

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ  
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»  
(САМАРСКИЙ УНИВЕРСИТЕТ)

На правах рукописи

**Жданова Александра Николаевна**

**ИНТЕРПРЕТИРУЕМЫЙ АНАЛИЗ МУЛЬТИМОДАЛЬНОЙ  
ИНФОРМАЦИИ НА ОСНОВЕ НЕЙРОСИМВОЛЬНОЙ МОДЕЛИ**

**2.3.8. Информатика и информационные процессы**

Диссертация

на соискание ученой степени доктора технических наук

Научный консультант:

**Куприянов Александр Викторович**

доктор технических наук, доцент

Самара – 2026

## ОГЛАВЛЕНИЕ

|                                                                                                                                  |    |
|----------------------------------------------------------------------------------------------------------------------------------|----|
| ВВЕДЕНИЕ                                                                                                                         | 12 |
| 1 Анализ методов обработки мультимодальных данных и постановка задачи исследования                                               | 21 |
| 1.1 Мультимодальные данные и задачи их обработки                                                                                 | 21 |
| 1.1.1 Понятие и характеристики мультимодальных данных                                                                            | 21 |
| 1.1.2 Характеристики отдельных модальностей                                                                                      | 22 |
| 1.1.3 Задачи анализа мультимодальных данных                                                                                      | 24 |
| 1.1.4. Мультимодальность восприятия как теоретическое основание                                                                  | 25 |
| 1.2 Методы мультимодального слияния                                                                                              | 26 |
| 1.2.1 Раннее слияние                                                                                                             | 26 |
| 1.2.2. Позднее слияние                                                                                                           | 26 |
| 1.2.3 Гибридное слияние                                                                                                          | 27 |
| 1.2.4 Сравнительный анализ методов слияния                                                                                       | 28 |
| 1.3 Графы знаний в задачах анализа данных                                                                                        | 29 |
| 1.4 Графовые нейронные сети                                                                                                      | 31 |
| 1.5 Проблема интерпретируемости нейросетевых моделей                                                                             | 33 |
| 1.5.1 Методы апостериорной интерпретации                                                                                         | 34 |
| 1.5.2 Встроенная интерпретируемость                                                                                              | 35 |
| 1.6 Постановка научной проблемы и гипотеза исследования                                                                          | 37 |
| 1.7 Нейросимвольная интеграция графов знаний и графовых нейронных сетей                                                          | 38 |
| 1.8 Математическая постановка задачи                                                                                             | 40 |
| 1.9 Выводы по главе 1                                                                                                            | 42 |
| 2 Метод нейросимвольной интеграции графов знаний и графовых нейронных сетей для интерпретируемого анализа мультимодальных данных | 44 |
| 2.1 Концептуальная модель нейросимвольной интеграции мультимодальных данных                                                      | 45 |
| 2.2 Принципы нейросимвольной интеграции                                                                                          | 48 |
| 2.3 Обоснование выбора архитектуры                                                                                               | 50 |
| 2.4 Представление мультимодальных данных                                                                                         | 53 |
| 2.4.1 Функция мультимодальной интеграции                                                                                         | 56 |
| 2.4.2 Формальная модель графа знаний                                                                                             | 57 |
| 2.5 Типизированное представление входных данных                                                                                  | 58 |
| 2.6 Гибридная нейросимвольная мультимодальная модель                                                                             | 59 |

|       |                                                                      |    |
|-------|----------------------------------------------------------------------|----|
| 2.6.1 | Уровень 1 – Формирование первичного описания                         | 60 |
| 2.6.2 | Уровень 2 – Формирование вторичного описания                         | 60 |
| 2.6.3 | Адаптивное слияние вторичных описаний                                | 60 |
| 2.6.4 | Граф знаний и вывод графовой сети внимания (GAT)                     | 61 |
| 2.6.5 | Информирование нейронной сети предметной онтологией                  | 61 |
| 2.7   | Онтология предметной области                                         | 63 |
| 2.7.1 | Теоретические основы онтологического моделирования                   | 63 |
| 2.7.2 | Концепты онтологического паттерна                                    | 64 |
| 2.7.3 | Типы отношений в онтологии                                           | 64 |
| 2.7.4 | Формирование триплетов и семантических связей                        | 65 |
| 2.7.5 | Формализация онтологии в OWL и RDF                                   | 66 |
| 2.7.6 | Визуализация онтологического паттерна                                | 66 |
| 2.8   | Граф знаний                                                          | 67 |
| 2.9   | Нейросимвольная интеграция и механизм взаимодействия компонентов     | 69 |
| 2.9.1 | Двунаправленное взаимодействие данных и знаний                       | 70 |
| 2.9.2 | Символьный компонент                                                 | 71 |
| 2.9.3 | Организация обучения нейросимвольной модели                          | 73 |
| 2.9.4 | Сравнение способов представления знаний в интеллектуальных системах  | 75 |
| 2.10  | Роль метода в системах поддержки принятия решений                    | 76 |
| 2.11  | Выводы по главе 2                                                    | 77 |
| 3     | Применение метода к задаче интерпретируемого анализа цифрового следа | 79 |
| 3.1   | Постановка прикладной задачи                                         | 79 |
| 3.2   | Описание данных                                                      | 81 |
| 3.2.1 | Сбор и описание выборки                                              | 82 |
| 3.2.2 | Этические аспекты обработки данных                                   | 87 |
| 3.2.3 | Ограничения датасета                                                 | 88 |
| 3.3   | Текстовая модальность                                                | 89 |
| 3.3.1 | Функциональная типология никнеймов                                   | 89 |
| 3.3.2 | Типология и тематика пользовательских статусов                       | 90 |
| 3.4   | Визуальная модальность                                               | 91 |
| 3.5   | Табличная (числовая) модальность                                     | 92 |

|       |                                                                         |     |
|-------|-------------------------------------------------------------------------|-----|
| 3.5.1 | Признаковое описание цифровой активности                                | 93  |
| 3.5.2 | Темпоральный анализ активности                                          | 93  |
| 3.5.3 | Типологизация активности пользователя                                   | 94  |
| 3.6   | Предметная онтология и граф знаний                                      | 95  |
| 3.6.1 | Объект анализа                                                          | 95  |
| 3.6.2 | Целевые характеристики                                                  | 96  |
| 3.6.3 | Типы отношений и примеры триплетов                                      | 97  |
| 3.6.4 | Пример формирования графа знаний для пользователя                       | 98  |
| 3.7   | Состав и источники знаний графа                                         | 101 |
| 3.8   | Интерпретируемость результатов прикладного анализа                      | 101 |
| 3.8.1 | Механизм формирования объяснения                                        | 102 |
| 3.8.2 | Пример объясняющего подграфа                                            | 103 |
| 3.8.3 | Вклад модальностей и ограничения интерпретации                          | 105 |
| 3.9   | Выводы по главе 3                                                       | 105 |
| 4     | Архитектура системы                                                     | 107 |
| 4.1   | Общая архитектура клиент-серверной системы                              | 107 |
| 4.1.1 | Подсистемы и их функции                                                 | 109 |
| 4.1.2 | Интеграция компонентов и последовательность обработки запроса           | 110 |
| 4.2   | Технологическая реализация графа знаний                                 | 111 |
| 4.2.1 | Схема узлов и рёбер                                                     | 112 |
| 4.2.2 | Наполнение графа из данных                                              | 113 |
| 4.2.3 | Количественные параметры наполненного графа                             | 114 |
| 4.3   | Алгоритмическое обеспечение конвейера обработки профиля                 | 114 |
| 4.3.1 | Конвейер обработки профиля пользователя                                 | 114 |
| 4.3.2 | Оптимизация производительности                                          | 116 |
| 4.4   | Пользовательский интерфейс и REST API                                   | 117 |
| 4.4.1 | Веб-интерфейс для HR-специалистов                                       | 117 |
| 4.4.2 | Программный интерфейс системы                                           | 120 |
| 4.5   | Ограничения предложенного подхода и направления дальнейших исследований | 121 |
| 4.5.1 | Текущие ограничения                                                     | 121 |
| 4.5.2 | Направления дальнейших исследований                                     | 121 |
| 4.6   | Выводы по главе 4                                                       | 122 |



|                                                                                                   |     |
|---------------------------------------------------------------------------------------------------|-----|
| 5 Экспериментальная проверка разработанных методов и моделей                                      | 123 |
| 5.1 Программа экспериментальных исследований                                                      | 123 |
| 5.2 Метрики оценки и методика валидации                                                           | 124 |
| 5.2.1 Метрики точности предсказания                                                               | 125 |
| 5.2.2 Метрики валидности корреляций                                                               | 125 |
| 5.2.3 Метрики интерпретируемости                                                                  | 126 |
| 5.2.4 Базовые модели                                                                              | 128 |
| 5.2.5 Статистическая значимость и кросс-валидация                                                 | 129 |
| 5.3 Психометрическая валидация признаков                                                          | 130 |
| 5.4 Мульти模альная интеграция признаков                                                            | 131 |
| 5.4.1 Интерпретируемость мульти模альной модели                                                     | 135 |
| 5.5 Экспериментальное сравнение GAT с базовыми моделями                                           | 137 |
| 5.6 Валидация метрики GPI и примеры интерпретации                                                 | 143 |
| 5.6.1 Примеры интерпретации предсказаний                                                          | 143 |
| 5.6.2 Валидация метрики интерпретируемости GPI                                                    | 148 |
| 5.6.3 Визуализация весов внимания                                                                 | 149 |
| 5.7 Апробация системы на прикладной задаче                                                        | 151 |
| 5.7.1 Экспериментальный сценарий: задача предварительного отбора профилей                         | 151 |
| 5.7.2 Анализ ошибок                                                                               | 152 |
| 5.7.3. Сравнение с альтернативными подходами                                                      | 154 |
| 5.8 Абляция онтологических ограничений                                                            | 155 |
| 5.9 Кросс-доменная проверка переносимости метода                                                  | 157 |
| 5.10 Выводы по главе 5                                                                            | 158 |
| ЗАКЛЮЧЕНИЕ                                                                                        | 160 |
| СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ                                                          | 164 |
| СПИСОК ЛИТЕРАТУРЫ                                                                                 | 166 |
| ПРИЛОЖЕНИЕ А Описание программной реализации нейросимвольной системы анализа мульти模альных данных | 188 |
| А.1 Модуль сбора данных из API социальной сети VK                                                 | 189 |
| А.2 Модуль предобработки мульти模альных данных                                                     | 190 |
| А.3 Модуль извлечения текстовых признаков                                                         | 192 |
| А.4 Модуль извлечения визуальных признаков                                                        | 193 |

|                                                                           |     |
|---------------------------------------------------------------------------|-----|
| А.5 Модуль извлечения поведенческих признаков                             | 194 |
| А.6 Модуль формирования графа знаний                                      | 195 |
| А.7 Модуль графового вывода на основе GAT                                 | 197 |
| А.8 Модуль генерации интерпретируемых объяснений                          | 198 |
| А.9 Модуль кэширования эмбедингов                                         | 198 |
| А.10 Модуль управления моделями(предзагрузка)                             | 199 |
| А.11 Модуль API взаимодействия                                            | 199 |
| ПРИЛОЖЕНИЕ Б Технические детали обучения моделей мультимодального анализа | 201 |
| Б.1 Классификация никнеймов: модель RoBERTa                               | 201 |
| Б.1.1 Архитектура модели                                                  | 201 |
| Б.1.2 Двухуровневая стратегия обучения                                    | 202 |
| Б.1.3. Предобработка данных                                               | 207 |
| Б.1.4 Процедура обучения и валидации                                      | 209 |
| Б.1.5 Технические детали реализации                                       | 210 |
| Б.2 Классификация статусов: коммерческие и ситуативные (RuBERT)           | 211 |
| Б.2.1 Архитектура модели                                                  | 211 |
| Б.2.2 Датасет и разметка                                                  | 211 |
| Б.2.3 Процедура обучения                                                  | 213 |
| Б.2.4 Предобработка данных                                                | 214 |
| Б.2.5 Технические детали реализации                                       | 216 |
| Б.3 Тематическая классификация статусов                                   | 216 |
| Б.3.1 Архитектура модели                                                  | 217 |
| Б.3.2 Датасет и разметка                                                  | 217 |
| Б.3.3 Процедура обучения                                                  | 218 |
| Б.3.4 Технические детали реализации                                       | 220 |
| Б.4 Классификация поведенческих паттернов                                 | 220 |
| Б.4.1 Архитектура модели                                                  | 221 |
| Б.4.2 Датасет и разметка                                                  | 222 |
| Б.4.3 Процедура обучения                                                  | 223 |
| Б.5 Модели визуального анализа                                            | 224 |
| Б.5.1 Распознавание возраста: ResNet50                                    | 225 |
| Б.5.2 Распознавание пола: VGG16                                           | 226 |

|                                                                                                      |     |
|------------------------------------------------------------------------------------------------------|-----|
| Б.5.3 Классификация типов визуального контента: ResNet50                                             | 228 |
| Б.5.4 Классификация зоонимов: ResNet50                                                               | 231 |
| Б.5.5 Распознавание эмоций: CNN                                                                      | 233 |
| Б.5.6 Технические детали реализации                                                                  | 236 |
| Б.6 Мультимодальная модель интеграции (Hybrid Fusion)                                                | 236 |
| Б.6.1 Процедура обучения                                                                             | 239 |
| Б.6.2 Анализ вклада компонентов                                                                      | 240 |
| Б.6.3 Сравнение с базовыми моделями                                                                  | 241 |
| Б.6.4 Интерпретируемость SHAP                                                                        | 243 |
| Б.6.5 Технические детали реализации                                                                  | 247 |
| Б.7 Общие примечания                                                                                 | 248 |
| ПРИЛОЖЕНИЕ В Детальные таблицы метрик моделей лингвистического и<br>визуального анализа              | 249 |
| В.1 Метрики классификации никнеймов (RoBERTa)                                                        | 249 |
| В.1.1 Этап 1: Синтетические данные                                                                   | 249 |
| В.1.2. Этап 2: Реальные данные ВКонтакте                                                             | 250 |
| В.1.3 Матрица ошибок для реальных данных ВК                                                          | 251 |
| В.2 Метрики классификации статусов                                                                   | 252 |
| В.2.1 Коммерческие и ситуативные                                                                     | 252 |
| В.2.2. Тематическая классификация                                                                    | 253 |
| В.2.3 Мировоззренческие установки                                                                    | 254 |
| В.2.4 Типы поведения                                                                                 | 255 |
| В.3 Метрики визуального анализа                                                                      | 257 |
| В.3.1 Классификация типа аватара (6 категорий, ResNet50)                                             | 257 |
| В.3.2 Классификация зоонимов в аватарах (10 категорий, ResNet50)                                     | 258 |
| В.3.3 Распознавание эмоций                                                                           | 258 |
| В.4 Метрики определения демографических характеристик                                                | 260 |
| В.4.1 Определение возраста (регрессия, ResNet50)                                                     | 260 |
| В.4.2 Определение пола (бинарная классификация, VGG16)                                               | 260 |
| В.5 Сводная таблица метрик всех моделей                                                              | 261 |
| ПРИЛОЖЕНИЕ Г Статистическое обоснование признакового пространства<br>системы анализа цифрового следа | 263 |
| Г.1 Методология отбора признаков                                                                     | 263 |

|                                                                      |     |
|----------------------------------------------------------------------|-----|
| Г.1.1 Датасет                                                        | 263 |
| Г.1.2 Разметка и контроль качества данных                            | 263 |
| Г.1.3 Процедура отбора признаков                                     | 264 |
| Г.2 Текстовая модальность – никнеймы                                 | 265 |
| Г.2.1 Регрессионная валидация независимости признаков                | 267 |
| Г.3 Текстовая модальность – тематика статусов                        | 268 |
| Г.4 Поведенческие типы                                               | 269 |
| Г.5. Визуальная модальность                                          | 270 |
| Г.6 Сводная таблица признакового пространства                        | 271 |
| Г.7 Формирование графа знаний                                        | 272 |
| Г.8 Ограничения признакового пространства                            | 273 |
| ПРИЛОЖЕНИЕ Д Технические детали визуальных и мультимодальных моделей | 275 |
| Д.1 Определение возраста по аватару                                  | 275 |
| Д.1.1 Постановка задачи                                              | 275 |
| Д.1.2 Конвейер обработки данных                                      | 276 |
| Д.1.3 Сравнение архитектур                                           | 277 |
| Д.1.4 Детальная архитектура ResNet50 для регрессии возраста          | 278 |
| Д.1.5 Обучение модели ResNet50                                       | 279 |
| Д.2 Определение пола по аватару                                      | 281 |
| Д.2.1 Архитектура модели                                             | 281 |
| Д.2.2 Обучение модели                                                | 282 |
| Д.3. Классификация типа аватара (6 категорий)                        | 283 |
| Д.3.1 Архитектура модели                                             | 283 |
| Д.3.2 Обучение модели                                                | 284 |
| Д.4 Распознавание эмоций                                             | 285 |
| Д.4.1 Архитектура модели                                             | 285 |
| Д.4.2 Обучение и адаптация к ВК – аватарам                           | 287 |
| Д.5 Мультимодальная интеграция: архитектура Hybrid Fusion            | 287 |
| Д.5.1 Общая схема архитектуры                                        | 288 |
| Д.5.2 Компонент 1: Модальные энкодеры                                | 289 |
| Д.5.2.1 Текстовый энкодер                                            | 289 |
| Д.5.2.2 Визуальный энкодер (ResNet50)                                | 289 |

|                                                                                                             |     |
|-------------------------------------------------------------------------------------------------------------|-----|
| Д.5.3 Компонент 2: Проекционные слои                                                                        | 289 |
| Д.5.4 Компонент 3: Двусторонний механизм внимания                                                           | 290 |
| Д.5.5 Компонент 4: Адаптивное слияние модальностей                                                          | 291 |
| Д.5.6 Компонент 5: Классификационная голова                                                                 | 291 |
| Д.5.7 Функция потерь                                                                                        | 292 |
| Д.5.8 Обучение модели                                                                                       | 292 |
| Д.6 Метрики качества                                                                                        | 293 |
| Д.6.1 Средняя абсолютная ошибка (MAE)                                                                       | 293 |
| Д.6.2 Коэффициент корреляции Пирсона (r)                                                                    | 294 |
| Д.6.3 F1-score (для дискретизированных факторов)                                                            | 294 |
| ПРИЛОЖЕНИЕ Е Результаты исследования вклада компонентов модели и таблицы корреляций по 26 факторам личности | 295 |
| Е.1 Оценка вклада архитектурных компонентов модели                                                          | 295 |
| Е.1.1 Результаты исследования вклада компонентов модели                                                     | 296 |
| Е.1.2 Интерпретация результатов                                                                             | 297 |
| Е.2 Таблица корреляций мультимодальной модели по 26 факторам                                                | 299 |
| Е.2.1 Анализ результатов                                                                                    | 301 |
| Е.3 SHAP для вклада модальностей по 26 факторам                                                             | 302 |
| ПРИЛОЖЕНИЕ Ж Математические детали SHAP для интерпретации мультимодальной модели                            | 306 |
| Ж.1 Теоретические основы из теории игр Shapley Values                                                       | 306 |
| Ж.2 Адаптация Shapley Values для машинного обучения                                                         | 308 |
| Ж.3 SHAP для мультимодальной модели                                                                         | 309 |
| Ж.4 Пример вычисления SHAP для конкретного пользователя                                                     | 312 |
| Ж.5 Усреднение SHAP по тестовой выборке                                                                     | 314 |
| Ж.6 Вычислительная сложность и оптимизация                                                                  | 314 |
| Ж.7 Ограничения метода SHAP для модели                                                                      | 315 |
| ПРИЛОЖЕНИЕ И Визуализации системы анализа цифрового следа                                                   | 316 |
| И.1 Трёхуровневая архитектура системы                                                                       | 316 |
| И.2 Алгоритмический конвейер обработки                                                                      | 318 |
| И.3 Граф знаний: структура узлов и связей                                                                   | 320 |
| И.4 Скриншоты веб-интерфейса                                                                                | 321 |
| И.4.1 Главная страница                                                                                      | 321 |

|                                                                                                  |     |
|--------------------------------------------------------------------------------------------------|-----|
| И.4.2 Отчёт «Профиль личности»                                                                   | 322 |
| И.4.3 Интерпретация результатов                                                                  | 323 |
| И.4.4 Визуализация графа знаний                                                                  | 324 |
| И.5 Диаграммы компонентов и взаимодействий                                                       | 325 |
| И.5.1 Диаграмма последовательности: Анализ профиля                                               | 325 |
| И.5.2 Компонентная диаграмма: ML–конвейер                                                        | 328 |
| И.6 Результаты валидации и апробации                                                             | 330 |
| И.6.1 Апробация системы на задаче анализа цифровых профилей пользователей                        | 330 |
| И.7 Примечания к визуализациям                                                                   | 332 |
| ПРИЛОЖЕНИЕ К. Кросс-доменный перенос метода: дерматоскопическая диагностика новообразований кожи | 333 |
| К.1 Постановка задачи переноса                                                                   | 333 |
| К.2 Набор данных и опорная сеть классификации                                                    | 333 |
| К.3 Реализация онтологического паттерна: дерматоскопические признаки                             | 334 |
| К.4 Граф знаний «признак – диагноз»                                                              | 335 |
| К.5 Рассуждающие модули и применение метрики GPI                                                 | 336 |
| К.6 Результаты                                                                                   | 337 |
| К.7 Выводы по приложению                                                                         | 338 |
| ПРИЛОЖЕНИЕ Л Кросс-доменная проверка метода на задачах интерпретируемого анализа текста          | 340 |
| Л.1 Постановка задачи переноса                                                                   | 340 |
| Л.2 Набор данных                                                                                 | 341 |
| Л.3. Реализация паттерна                                                                         | 342 |
| Л.4 Формирование графа знаний                                                                    | 343 |
| Л.5 Проверка метода на различных моделях вывода                                                  | 344 |
| Л.6 Результаты                                                                                   | 345 |
| Л.7 Согласованность объяснений с человеческими обоснованиями                                     | 346 |
| Л.8 Анализ смещения                                                                              | 347 |
| Л.9. Языково-согласованная проверка на русскоязычном наборе                                      | 348 |
| Л.10 Реализация полного паттерна на задаче эмоциональной классификации                           | 350 |
| ПРИЛОЖЕНИЕ М Формализация онтологии и графа знаний предметной области                            | 353 |

|                                                                        |     |
|------------------------------------------------------------------------|-----|
| М.1 Концепты и предикаты онтологии                                     | 353 |
| М.2 Спецификация онтологии в OWL/RDF                                   | 354 |
| М.3 Схема графа знаний в Neo4j                                         | 355 |
| М.4 Примеры триплетов и логического вывода                             | 356 |
| ПРИЛОЖЕНИЕ Н Онтологический паттерн проектирования                     |     |
| «объект анализа – признак – целевая характеристика»                    | 357 |
| Н.1 Классы концептов                                                   | 357 |
| Н.2 Предикаты                                                          | 357 |
| Н.3 Формализация в OWL и RDF                                           | 358 |
| Н.4. Аксиомы и ограничения                                             | 359 |
| Н.5. Порядок проектирования системы на основе паттерна                 | 360 |
| Н.6 Объяснимость как следствие структуры и переносимость               | 360 |
| ПРИЛОЖЕНИЕ П Метрика интерпретируемости путей в графе GPI              | 363 |
| П.1 Идея и определение объясняющего пути                               | 363 |
| П.2 Основная метрика – покрытие путями (GPI_cov)                       | 364 |
| П.3 Вспомогательная метрика – взвешенная согласованность путей (GPI_w) | 364 |
| П.4 Свойства метрики                                                   | 364 |
| П.5 Проверка верности методом удаления концептов                       | 365 |
| П.6 Валидация и применение                                             | 365 |
| ПРИЛОЖЕНИЕ Р Акты внедрения                                            | 366 |

## ВВЕДЕНИЕ

Диссертационная работа посвящена разработке нейросимвольных методов и моделей интерпретируемого анализа мультимодальных данных на основе интеграции графов знаний и графовых нейронных сетей. В качестве прикладной области рассматривается задача формирования психологических профилей личности по данным цифровой активности пользователей социальных сетей. Основной акцент сделан на обеспечении встроенной интерпретируемости результатов анализа, отличающей предлагаемый подход от существующих моделей типа «чёрный ящик».

### Актуальность темы

Современные информационные системы генерируют большие объёмы мультимодальных данных, объединяющих текстовую, визуальную и поведенческую информацию. По данным International Data Corporation, глобальный объём данных превысил 181 зеттабайт в 2025 году. Значительная часть этой информации представлена в слабоструктурированной форме, что существенно усложняет её анализ традиционными методами [1].

Совместная обработка данных различных модальностей не сводится к независимому анализу каждой из них с последующим объединением результатов. Семантические зависимости между модальностями: согласованность, дополняемость или противоречивость информации – требуют методов, способных моделировать кросс-модальные связи в едином пространстве представлений [2, 3]. Как показано в работах Балтрушайтиса Т., Ахуджи Ч. и Моренси Л.-Ф. [4], эффективная мультимодальная интеграция позволяет повысить точность анализа на 15–30 % по сравнению с одномодальными подходами.

Современные методы машинного обучения, основанные на трансформерных архитектурах [5, 6] и глубоких свёрточных сетях [7, 8], достигли значительных успехов в извлечении признаков из текстовой и визуальной информации. Однако результат формируется без возможности проследить логику вывода, ограничивая применение таких моделей в задачах, требующих обоснования принимаемых решений [9]. Проблема интерпретируемости особенно остро стоит в медицинской диагностике, финансовом анализе и системах поддержки принятия решений, где ошибочные или необъяснимые решения могут иметь серьёзные последствия [10].

Одним из перспективных направлений преодоления этого ограничения является нейросимвольная интеграция, объединяющая обучаемость нейросетевых моделей с



формализованным представлением знаний в виде онтологий и графов знаний [11, 12]. Благодаря использованию графовых моделей знаний стало возможным явно задавать семантические связи между элементами данных и целевыми показателями, а применению графовых нейронных сетей [13, 14] – учитывать эти связи при формировании результатов.

Вместе с тем в настоящее время недостаточно разработаны методы и архитектуры, в которых интерпретируемость является встроенным свойством процесса обработки данных, а не результатом последующего постфактум анализа. Существующие апостериорные методы интерпретации [15, 16] формируют локальные приближения уже полученных решений, при этом не гарантируется соответствие объяснений реальному механизму вывода, а результаты не верифицируются на основе предметных знаний. Кроме того, отсутствуют унифицированные подходы к количественной оценке интерпретируемости нейросимвольных моделей при работе с мультимодальными данными.

Таким образом, разработка методов и архитектур интерпретируемой обработки мультимодальных данных на основе нейросимвольной интеграции является актуальной научной задачей, имеющей существенное значение для развития интеллектуальных информационных систем и теории обработки информации.

#### Степень разработанности проблемы

Фундаментальные принципы мультимодального слияния систематизированы в работах Балтрушайтиса Т. и соавторов [4], где выделены подходы раннего, позднего и гибридного объединения признаков. Нгиама Дж. и соавторы [17] исследовали глубокое мультимодальное обучение для задач распознавания речи и видео. Заде А. и соавторы [18] предложили механизмы тензорного слияния.

Развитие трансформерных архитектур [5] стало основой современных языковых моделей. Девлин Дж. и соавторы [19] разработали BERT, Лю И. и соавторы [6] – RoBERTa. Для русскоязычных текстов адаптирована модель RuBERT [20]. В области компьютерного зрения Хе К. и соавторы [7] предложили ResNet, Тана М. и Ле К. [21] – EfficientNet, Досовицкий А. и соавторы [8] – Vision Transformer.

Концептуальные основы графов знаний заложены в работах Совы Дж. [22]. Методы векторного представления графов знаний развиты в работах Бордеса А. [23], Янга Б. [24], Труйона Т. [25]. Обзор существующих подходов приведён Вангом Ц. и соавторами [26].

Базовые подходы к обработке графовых структур представлены в работах Кипфа Т. и Веллинга М. [13] (GCN), Величковича П. и соавторов [14] (GAT), Хэмилтона У. и соавторов [27] (GraphSAGE), Шлихткрулла М. и соавторов [28] (R-GCN).

Основные подходы к интерпретации моделей машинного обучения представлены в работах Рибейру М. [15] (LIME), Лундберга С. и Ли С. [16] (SHAP), Селвараджу Р. и соавторов [29] (Grad-CAM). Систематизация методов дана в работах Арриеты А. и соавторов [30] и Мольнара К. [31].

Принципы нейросимвольного подхода изложены в работах Гарсеса А. и Лэмба Л. [11], Маркуса Г. [32], Де Рэдта Л. и соавторов [12].

В более ранних работах автора [33\*–36\*] рассмотрены вопросы интеллектуальной поддержки принятия решений, психодиагностики, анализа личностных характеристик и применения нейросетевых методов обработки данных. Эти исследования сформировали основу для последующей разработки нейросимвольного подхода, представленного в настоящей диссертационной работе. Переносимость разработанного в диссертации подхода дополнительно подтверждена на задачах дерматоскопической диагностики, классификации токсичных комментариев и распознавания эмоций (приложения К, Л)

Несмотря на развитие методов мультимодального анализа, графовых нейронных сетей и объяснимого искусственного интеллекта, остаются нерешёнными следующие задачи:

- отсутствуют унифицированные архитектуры, объединяющие мультимодальные признаки и семантические связи в рамках единого интерпретируемого представления;
- не разработаны методы количественной оценки интерпретируемости нейросимвольных моделей;
- недостаточно исследованы механизмы взаимодействия нейронной и символьной компонент при обработке мультимодальных данных.

Указанные ограничения определяют научную задачу разработки методов, моделей и архитектур интерпретируемого анализа мультимодальных данных на основе нейросимвольной интеграции графов знаний и графовых нейронных сетей.

#### Объект исследования

Объектом исследования являются процессы интерпретируемого анализа типизированной мультимодальной информации.

### Предмет исследования

Предметом исследования являются методы и модели интерпретируемого анализа типизированной мультимодальной информации, основанные на гибридной нейросимвольной интеграции графов знаний и графовых нейронных сетей.

### Цель и задачи исследования

Целью диссертационной работы является разработка метода интерпретируемого анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, интегрирующей графы знаний и графовые нейронные сети для информирования нейросетевой модели предметными знаниями и формирования интерпретируемого результата анализа.

Для достижения поставленной цели решены следующие задачи:

1. Анализ существующих методов представления, интеграции и интерпретации мультимодальных данных, а также методов нейросимвольного моделирования.
2. Разработка графовой модели представления мультимодальных данных, позволяющей совместно описывать текстовые, визуальные и табличные признаки в единой семантической структуре на основе онтологического паттерна «объект анализа – признак – целевая характеристика».
3. Разработка нейросимвольного метода интерпретируемого анализа мультимодальных данных, использующего граф знаний для структурирования предметных знаний и графовую нейронную сеть для формирования прогнозов и объяснений.
4. Разработка доменно-инвариантного онтологического паттерна проектирования интерпретируемых нейросимвольных систем «объект анализа – признак – целевая характеристика», выделяющего переносимое между предметными областями ядро.
5. Разработка механизма встроенной интерпретируемости результатов анализа, при котором объяснения формируются непосредственно в процессе нейросимвольного вывода в виде семантических путей графа знаний, допускающих экспертную верификацию.
6. Разработка модели количественной оценки интерпретируемости результатов анализа на основе семантических путей графа знаний.
7. Разработка технологии и программного комплекса, реализующих предложенные методы и модели, и проведение апробации на задаче анализа цифрового

следа пользователей социальной сети с оценкой точности прогнозирования и качества интерпретации результатов.

### Методы исследования

Методологическую основу исследования составляют теория графов [37, 38], теория нейронных сетей [39, 40], теория информации [41], онтологическое моделирование [42, 43] и методы математической статистики [44]. Для решения поставленных задач применялись методы мультимодальной обработки данных, построения графов знаний, нейросимвольной интеграции, графовые нейронные сети, а также методы количественной оценки интерпретируемости.

### Научная новизна

Научная новизна диссертационной работы заключается в разработке нейросимвольного подхода к интерпретируемому анализу типизированной мультимодальной информации, основанного на интеграции мультимодальных представлений, графов знаний и графовых нейронных сетей.

В работе впервые:

1. Разработана гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации, интегрирующая нейросетевую модель с интерпретируемым графом знаний, конструируемым на основе доменно-инвариантного онтологического паттерна «объект анализа – признак – целевая характеристика».
2. Разработан нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, включающий встроенный механизм интерпретируемости результатов.
3. Предложена метрика интерпретируемости путей GPI (Graph Path Interpretability) для количественной оценки результатов нейросимвольного анализа по семантическим путям вывода в графе знаний.
4. Разработана технология обработки и анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, включающая извлечение признаков из модальностей, построение предметной онтологии и графа знаний, нейросимвольный вывод и формирование интерпретируемых объяснений.

5. Разработана информационная система обработки и анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, реализующая полный цикл обработки данных от извлечения признаков до формирования интерпретируемых объяснений.

#### Теоретическая значимость

Теоретическая значимость работы заключается в развитии методов интерпретируемого анализа типизированной мультимодальной информации на основе нейросимвольного подхода. Разработанная гибридная нейросимвольная модель формализует совместное использование вычислительного и семантического представлений объекта анализа и информирование нейросетевой модели знаниями предметной области. Предложенный метод развивает подходы к встроенной интерпретации результатов посредством семантических путей графа знаний. Метрика GPI дополняет аппарат количественной оценки интерпретируемости нейросимвольных моделей.

#### Практическая значимость

Практическая значимость работы заключается в разработке технологии и информационной системы интерпретируемого анализа типизированной мультимодальной информации, реализующих полный цикл обработки данных: извлечение признаков, построение графа знаний, нейросимвольный анализ и формирование интерпретируемых объяснений. Использование предобученных энкодеров с фиксированными параметрами сокращает число обучаемых параметров по сравнению со сквозными мультимодальными архитектурами и снижает вычислительные затраты на этапе обучения. Разработанные методы и программные средства могут применяться в интеллектуальных информационных системах, требующих совместного анализа разнородных данных и содержательной интерпретации получаемых результатов.

Результаты диссертационного исследования внедрены в:

- ГБУ ДПО «Региональный социопсихологический центр» (г. Самара);
- ФГБУ «ЦЭНКИ» – Космодром «Восточный» (г. Циолковский);
- ПАО «Сбербанк России» (г. Самара);
- ООО «Профиль» (г. Тольятти);
- ООО «ФОТОН-БИО» (г. Черноголовка);
- АНО «Иннокуб» (г. Краснодар);

- ООО НПП «Интеллект Софт» (г. Самара);
- ООО «Международный образовательный центр «ОПТИМА СТАДИ»» (г. Самара);
- филиал РТРС Саратовский ОРТПЦ (г. Саратов);
- ООО «Самара-Сигнал» (г. Самара);
- учебный процесс Самарского университета (курсы «Системы искусственного интеллекта», «Нейронные сети», «ИТ-практикум») (г. Самара).

Результаты диссертационного исследования использованы при выполнении государственных заданий Министерства науки и высшего образования Российской Федерации (FSSS-2020-0017, FSSS-2026-0003), проекта Благотворительного фонда Владимира Потанина «Трансформация курса «Нейронные сети» в исследовательскую AI-лабораторию», мероприятий программы развития Научно-образовательного математического центра Приволжского федерального округа (2023–2025 гг.), а также комплексного проекта «Разработка мультиагентной платформы адаптивного планирования и организация на ее основе высокотехнологичного производства по созданию промышленных интеллектуальных систем управления ресурсами предприятий в реальном времени», выполнявшегося в рамках реализации Постановления Правительства Российской Федерации от 09.04.2010 № 218.

#### Достоверность результатов

Достоверность полученных результатов обеспечивается:

- корректным применением математического аппарата теории графов, теории нейронных сетей и методов статистического анализа;
- использованием широко применяемых архитектур глубокого обучения [19, 7, 14] и современных программных средств их реализации (PyTorch, PyTorch Geometric, Hugging Face Transformers);
- экспериментальной проверкой на наборах данных, при формировании которых было собрано более 31 тыс. пользовательских профилей;
- статистической валидацией результатов с использованием коррекции Бонферрони для множественных сравнений;
- сопоставлением результатов с данными психометрического тестирования ( $N = 1722$ ) и экспертной оценкой качества формируемых объяснений (87 % экспертных оценок не ниже 4 баллов по пятибалльной шкале);

– воспроизводимостью вычислительных экспериментов при фиксированных параметрах инициализации, что обеспечивает повторяемость полученных результатов.

#### Апробация результатов

Результаты исследований, полученные на различных этапах выполнения диссертационной работы, докладывались и обсуждались на международных и всероссийских научных конференциях, в том числе: Международной научной конференции «Конвергентные когнитивно-информационные технологии» (Москва, МГУ, 2025 г.); Международной конференции и молодёжной школе «Информационные технологии и нанотехнологии» (ИТНТ, Самара, 2015–2024 гг.); Международной конференции Propulsion and Energy 2015 (Орландо, США, 2015 г.); Международном симпозиуме Environmental and Engineering Aspects for Sustainable Living (Euro-Eco, Ганновер, Германия, 2014 г.); Международной конференции «Образование через всю жизнь для устойчивого развития» (Саранск, 2014 г.); Международных общественно-научных чтениях, посвящённых памяти Ю.А. Гагарина (Гагарин, 2015 г.); Международной молодёжной научной конференции «Королёвские чтения» (Самара, 2017 г.); молодёжных научных конференциях Самарского университета (Самара, 2017–2018 гг.).

Отдельные результаты диссертационного исследования представлялись и обсуждались в рамках Летней цифровой школы Сбера (2021, 2024, 2025 гг.) и программы дополнительного профессионального образования «Технологии искусственного интеллекта и их применение», реализованной Университетом ИТМО совместно с Альфа-Банком (2024 г.).

#### Публикации

Соискатель имеет 75 опубликованных работ, в том числе 42 работы по теме диссертации, из которых 3 монографии, 5 статей в изданиях, индексируемых в международных базах данных Scopus и Web of Science, 11 статей, опубликованных в ведущих рецензируемых научных журналах и изданиях, рекомендованных Высшей аттестационной комиссией, а также 4 свидетельства о государственной регистрации программ для ЭВМ.

#### Личный вклад автора

Все основные результаты диссертационной работы, включая постановку задачи, разработку моделей, методов, алгоритмов и программного комплекса, проведение экспериментов и анализ полученных результатов, получены автором лично.

### Структура и объём диссертации

Диссертация состоит из введения, пяти глав, заключения, списка условных обозначений и сокращений, списка литературы и приложений. Общий объём работы составляет 377 страниц машинописного текста, включая 59 рисунков и 122 таблицы. Список литературы включает 256 наименований.

### Положения, выносимые на защиту

На защиту выносятся следующие положения, содержащие элементы научной новизны:

1. Разработана гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации, в которой структура графа знаний задаётся семантическими предикатами предметной онтологии, а веса связей инициализируются статистически значимыми зависимостями и уточняются в процессе обучения; интерпретируемость вывода достигается без статистически значимого ухудшения точности предсказания.

2. Разработан нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации, основанный на двухуровневом формировании первичных и вторичных описаний и построении вычислительного и семантического представлений объекта анализа; при фиксированных параметрах модальных энкодеров обучению подлежит только графовый модуль, что сокращает число обучаемых параметров.

3. Предложена метрика GPI для количественной оценки интерпретируемости нейросимвольного анализа по статистически значимым семантическим путям графа знаний и их весам; значения метрики согласуются с экспертной оценкой, а верность объяснений проверяется методом удаления концептов.

4. Разработана технология обработки и анализа типизированной мультимодальной информации с доменно-инвариантным ядром, включающим онтологический паттерн, структуру графа знаний, механизм нейросимвольного вывода и метрику GPI; при переносе между предметными областями сохраняется инвариантное ядро метода, а изменяются предметное наполнение онтологии, состав признаков и целевых характеристик.

5. Разработана информационная система интерпретируемого анализа типизированной мультимодальной информации, объединяющая извлечение признаков, построение графа знаний, нейросимвольный вывод и формирование объяснений в едином цикле обработки и поддерживающая пакетную обработку данных.



# 1 Анализ методов обработки мультимодальных данных и постановка задачи исследования

## 1.1 Мультимодальные данные и задачи их обработки

### 1.1.1 Понятие и характеристики мультимодальных данных

Термин «мультимодальные данные» обозначает информацию, представленную одновременно в нескольких формах, каждая из которых характеризуется собственным способом кодирования, статистическими свойствами и семантической структурой [4]. В задачах машинного обучения под модальностью понимается определённый тип входных данных: текст, изображения, аудио-, видеоданные, временные ряды, табличные и графовые структуры [3].

Характерной особенностью мультимодальных данных является гетерогенность, проявляющаяся в различии форматов представления, уровня абстракции и семантической насыщенности отдельных модальностей [4]. Анализ мультимодальных данных не сводится к независимой обработке каждой модальности с последующим объединением результатов – существенную роль играют взаимосвязи между модальностями, отражающие согласованность, взаимное дополнение либо противоречивость представленной информации [45].

С теоретической точки зрения мультимодальные данные могут рассматриваться как совокупность согласованных описаний одного объекта, заданных в различных пространствах признаков. Текстовая модальность представляется последовательностью языковых единиц, визуальная – пространственной структурой изображения, табличная (числовая) – совокупностью числовых и категориальных признаков, включая временные ряды и агрегированные показатели. Каждая модальность характеризуется собственными источниками шума, масштабами представления и особенностями формирования признаков, что требует применения специализированных методов их извлечения и обработки [2]. Графовые структуры в данном контексте рассматриваются как способ представления взаимосвязей между сущностями, а не как самостоятельная модальность.

Мультимодальность является естественным свойством многих прикладных областей: медицинской диагностики, промышленного мониторинга, финансовой аналитики и других, где использование информации только одной модальности приводит к потере части сведений, содержащихся в данных. Обработка мультимодальных данных осуществляется в условиях высокой размерности,

неполноты информации и гетерогенности источников [46], что усложняет их анализ традиционными методами [47].

Вопросам интеграции мультимодальных данных посвящён ряд исследований. Лахат Д. и соавторы [48] рассмотрели взаимосвязи между модальностями на основе концепции общих и специфичных компонент представления данных. Рамачандрам Д. и Тейлор Г. [49] выполнили систематизацию подходов к глубокому мультимодальному обучению. В работе Гао Дж. и соавторов [50] представлен обзор современных методов мультимодального обучения применительно к задачам обработки естественного языка и компьютерного зрения.

Таким образом, несмотря на значительное количество исследований, вопросы совместного использования мультимодальных представлений и графовых моделей знаний, а также обеспечения интерпретируемости результатов анализа остаются недостаточно разработанными.

### 1.1.2 Характеристики отдельных модальностей

Текстовая модальность характеризуется высокой семантической насыщенностью и зависимостью интерпретации от контекста. Естественный язык обладает полисемией, синонимией и сложной синтаксической структурой, что требует применения контекстно-зависимых методов обработки [51]. Современные подходы основаны на трансформерных архитектурах [5], формирующих контекстуальные представления слов и фрагментов текста. Для обработки русскоязычных данных разработаны специализированные модели, включая RuBERT [20].

Развитие методов векторного представления текста прошло путь от статистических подходов к контекстуальным нейросетевым моделям. Распределённые векторные представления слов предложены Миколовым Т. и соавторами [52] (Word2Vec) и Пеннингтоном Дж., Сочером Р. и Мэннингом К. Д. [53] (GloVe). Дальнейшее развитие получили трансформерные языковые модели BERT [19] и RoBERTa [6]. Теоретические основы переноса обучения рассмотрены в работе Пана С. Дж. и Яна Ц. [54].

Отдельное направление связано с анализом тональности и эмоциональной окраски высказываний, для чего разработаны специализированные модели и корпуса [56, 57]. Сравнительный анализ эффективности различных архитектур рекуррентных нейронных сетей для анализа тональности русскоязычных комментариев выполнен автором в [36\*].

Визуальная модальность характеризуется пространственной организацией данных и высокой вариативностью условий формирования изображений. Информация

в изображениях представлена на различных уровнях абстракции: от локальных признаков (границ, текстур) до семантических представлений объектов и сцен [58]. Для извлечения признаков используются глубокие нейронные сети: свёрточные архитектуры ResNet [7] и EfficientNet [21], а также трансформерные модели Vision Transformer [8].

Для задач анализа визуальной информации о человеке разработаны специализированные модели. Система DeepFace [59] обеспечивает распознавание лиц с точностью, сопоставимой с человеческой, модель FaceNet [60] формирует унифицированные эмбединги лиц, а функция потерь ArcFace [61] повышает разделимость признакового пространства. Для локализации и предобработки лиц применяются каскадные свёрточные сети MTCNN [62] и однофазный детектор RetinaFace [63].

Отдельное направление связано с анализом визуальной самопрезентации пользователей в социальных сетях, где категория изображения определяется как объектным составом, так и функциональной ролью сцены. В работах автора [64\*] рассмотрен семантический разрыв при классификации изображений по функциональному назначению и показано, что объектный состав изображения не всегда достаточен для определения его функциональной категории. На реальных изображениях профилей пользователей социальной сети продемонстрировано, что наибольшие ошибки классификации связаны с категориями, различающимися отношением объектов в кадре, а не их набором. Полученные результаты подтверждают наличие семантического разрыва между объектными признаками и признаками, необходимыми для функциональной классификации изображений цифрового следа пользователя.

Табличная (числовая) модальность объединяет структурированные числовые и категориальные данные, временные ряды и агрегированные показатели. Для неё характерны неоднородность признаков, наличие временной динамики и различная частота поступления данных [65]. Анализ табличных данных осуществляется с использованием методов анализа временных рядов, рекуррентных нейронных сетей [66] и алгоритмов градиентного бустинга XGBoost [67] и LightGBM [68].

Обобщённые характеристики модальностей данных приведены в таблице 1. Модальности существенно различаются по способам представления информации и методам формирования признаков, что затрудняет их прямую интеграцию и требует разработки специализированных методов мультимодального анализа.

Таблица 1 – Характеристики модальностей данных

| Модальность          | Тип представления                                  | Типичные методы извлечения признаков | Особенности                                         |
|----------------------|----------------------------------------------------|--------------------------------------|-----------------------------------------------------|
| Текстовая            | Последовательность языковых единиц                 | Трансформеры (BERT, RoBERTa, RuBERT) | Контекстуальность, полисемия                        |
| Визуальная           | Пространственное представление изображения         | CNN (ResNet, EfficientNet), ViT      | Пространственная структура, многоуровневые признаки |
| Табличная (числовая) | Числовые и категориальные признаки, временные ряды | Градиентный бустинг, LSTM, GRU       | Неоднородность признаков, временная динамика        |

### 1.1.3 Задачи анализа мультимодальных данных

Согласно таксономии Балтрушайтиса Т. и соавторов [4], исследования в области мультимодального машинного обучения включают несколько направлений: мультимодальное представление, трансляцию между модальностями, выравнивание модальностей, мультимодальное слияние и совместное обучение. Наиболее близким к тематике настоящего исследования является мультимодальное слияние – объединение информации из нескольких модальностей для решения прикладных задач классификации, регрессии и прогнозирования.

Развитие трансформерных архитектур привело к появлению унифицированных подходов к обработке данных различных модальностей на основе механизма внимания [3]. В последнее время появились обобщающие обзоры, систематизирующие современные методы выравнивания и интеграции разнородных модальностей [69], а также рассматривающие развитие универсальных мультимодальных архитектур [69]. Вместе с тем большинство существующих исследований ориентировано на повышение точности прогнозирования, тогда как вопросы интеграции мультимодальных представлений с графовыми моделями знаний и формирования встроенной интерпретируемости результатов остаются недостаточно разработанными.

Прикладной областью, в которой методы мультимодального анализа цифрового следа получили широкое распространение, является вычислительная оценка психологических характеристик пользователей. В работах Косински М., Стиллвелла Д. и Грэпеля Т. [70] показано, что личностные характеристики могут быть предсказаны на основе цифровых следов поведения в социальных сетях, а компьютерные оценки личности в ряде случаев сопоставимы или превосходят оценки, формируемые людьми. Методологические основы вычислительной психометрики систематизированы в работах [71, 72].

Значительный вклад в развитие данного направления внесли методы анализа текстового контента социальных сетей. Подход открытого словаря предложен

Шварцем Х. А. и соавторами [74], а его развитие применительно к автоматической оценке личности представлено в работе Парка Г. и соавторов [75]. Онтологический подход к предсказанию психологических характеристик пользователей социальных сетей на основе графов знаний предложен автором в [211\*].

Отдельное направление связано с анализом визуального самопредставления пользователей. В работах [78, 79] исследована связь между характеристиками личности и визуальным контентом социальных сетей. Практические возможности и этические аспекты использования цифрового следа для анализа личности рассмотрены в работах [78, 80].

#### 1.1.4 Мультимодальность восприятия как теоретическое основание

Теоретическим основанием совместной обработки различных модальностей служит концепция мультимодального восприятия, согласно которой информация об объектах и явлениях формируется на основе интеграции данных из нескольких источников [81]. В когнитивных исследованиях показано, что человек воспринимает окружающий мир не через независимые сенсорные каналы, а посредством формирования единого представления, объединяющего зрительную, слуховую и тактильную информацию [82, 83].

Психологические основания концепции мультимодального восприятия заложены в работах отечественных исследователей. Рубинштейн С. Л. [84], Ананьев Б. Г. [85], Леонтьев А. Н. [86] и Выготский Л. С. [87] рассматривали восприятие, сознание и личность как системные образования, формирующиеся в процессе деятельности и социального взаимодействия. Психофизиологические механизмы интеграции информации различных сенсорных каналов исследованы Соколовым Е. Н. [88] и Александровым Ю. И. [89], когнитивные аспекты – Величковским Б. М. [90] и Ушаковым Д. В. [91].

Теоретико-методологические основы психологии личности представлены в работах Петровского А. В. и Ярошевского М. Г. [92], Бодалёва А. А. [93] и Леонтьева Д. А. [94]. Для решения задачи оценки психологических характеристик пользователей особое значение имеет факторный подход к описанию личности. Кэттелл Р. Б. [95, 96] разработал многофакторную модель личности и опросник 16PF, адаптированный для русскоязычной аудитории Капустиной А. Н. [97]. Методологические аспекты психодиагностики систематизированы Шмелёвым А. Г. [98]. Дополнительное значение для анализа цифрового следа имеют методы психолингвистического исследования текста [99, 100], а также учёт культурного контекста [101].

С точки зрения обработки информации различные модальности отражают различные аспекты одного и того же объекта и обладают различной степенью информативности в зависимости от контекста. Использование нескольких модальностей позволяет компенсировать неполноту отдельных источников данных, повысить устойчивость анализа и учитывать дополнительные закономерности, возникающие при совместном рассмотрении различных форм представления информации. Этот принцип лежит в основе современных методов мультимодального машинного обучения, в которых объект описывается совокупностью взаимодополняющих представлений [100, 101]. Значимую роль играют взаимосвязи между модальностями, отражающие согласованность, взаимное дополнение либо противоречивость представленной информации.

## 1.2 Методы мультимодального слияния

Интеграция признаков из различных модальностей является основной задачей мультимодального анализа [4, 49]. Слияние принято разделять на раннее, позднее и гибридное [3].

### 1.2.1 Раннее слияние

Раннее слияние (Early Fusion) основано на объединении признаков различных модальностей до этапа построения модели – чаще всего путём конкатенации признаков векторов в единое представление.

Достоинства: возможность совместного формирования признакового пространства и выявления зависимостей между модальностями на ранних этапах обработки [102]. В работе [17] показано, что совместное обучение на аудио- и видеоданных формирует более устойчивые представления по сравнению с независимым обучением каждой модальности.

Ограничения: объединение признаков увеличивает размерность пространства, усложняет обучение и повышает риск переобучения [103]. Различия в статистических свойствах модальностей могут вызывать дисбаланс их вклада: при наличии высокоинформативной модальности модель может игнорировать остальные источники данных [104]. Кроме того, раннее объединение ограничивает возможность учёта особенностей отдельных модальностей [4].

### 1.2.2 Позднее слияние

При позднем слиянии (Late Fusion) каждая модальность обрабатывается отдельной моделью, а объединение информации выполняется на уровне выходных

представлений или итоговых решений. Для интеграции результатов используются усреднение, голосование либо обучаемые механизмы комбинирования.

Достоинства: независимая разработка и настройка моделей для отдельных модальностей, использование предобученных моделей без совместного дообучения [105]. Система может функционировать при отсутствии одной из модальностей за счёт использования оставшихся источников информации.

Ограничения: позднее слияние ограничивает взаимодействие модальностей на уровне признаков, вследствие чего модель не учитывает часть семантических зависимостей между различными представлениями данных [106]. Как отмечают Рамачандрам Д. и Тейлор Г. [49], такой подход затрудняет моделирование комплементарности модальностей, при которой информация одной модальности уточняет или дополняет информацию другой.

Гибридное слияние, рассматриваемое далее, сочетает элементы раннего и позднего подходов, что позволяет частично преодолеть указанные ограничения.

### 1.2.3 Гибридное слияние

Гибридное слияние (Hybrid Fusion) сочетает элементы раннего и позднего подходов, интегрируя информацию на промежуточных этапах обработки. В отличие от раннего слияния, данный подход сохраняет возможность независимого формирования признаков отдельных модальностей; в отличие от позднего – учитывает их взаимодействие на уровне внутренних представлений.

Наиболее распространённым механизмом реализации гибридного слияния являются механизмы внимания, адаптивно распределяющие вклад различных модальностей в зависимости от контекста задачи [107]. Механизм кросс-модального внимания (Cross-Modal Attention) организует взаимодействие между модальностями посредством обмена информацией между их представлениями [5]. Развитие данного направления применительно к совместной обработке текстовой и визуальной информации представлено в работах [108, 109].

Ранее автором рассматривались гибридные интерпретируемые методы анализа данных, в том числе извлечение нечётких правил из числовых данных [110\*], нейронечёткая классификация [111\*] и кластеризация данных на самоорганизующихся картах Кохонена [112\*].

Помимо механизмов внимания, используются и другие подходы. Тензорное слияние (Tensor Fusion) моделирует взаимодействия между модальностями через операции над совместными представлениями [18], однако его применение ограничивается быстрым ростом размерности. Адаптивное слияние основано на

использовании обучаемых весов, определяющих вклад каждой модальности в итоговое представление [113].

В настоящей работе используется гибридное слияние, объединяющее кросс-модальное внимание и адаптивное объединение признаков. Таким образом учитываются взаимосвязи между текстовой, визуальной и табличной (числовой) модальностями, а также динамически определяется вклад каждой из них в зависимости от информативности для конкретного объекта анализа. Как отмечается в современных обзорах [69], гибридные подходы рассматриваются как наиболее перспективные для задач, требующих одновременно высокой точности и интерпретируемости результатов.

#### 1.2.4 Сравнительный анализ методов слияния

В контексте мультимодального слияния активно развиваются графовые методы интеграции разнородных модальностей [114, 115]. В отличие от традиционных подходов, графовые методы позволяют явно моделировать связи между элементами различных модальностей и использовать структуру этих связей при формировании итогового представления. Однако большинство существующих графовых мультимодальных методов нацелены на повышение точности и остаются моделями типа «чёрный ящик». Предлагаемый в работе подход дополняет это направление обеспечением верифицируемой интерпретируемости мультимодального вывода на основе графа знаний.

Сравнительные характеристики основных методов мультимодального слияния представлены в таблице 2.

Таблица 2 – Сравнительный анализ методов мультимодального слияния

| Метод                   | Кросс-модальное взаимодействие | Модульность | Интерпретируемость | Основные ограничения                                       |
|-------------------------|--------------------------------|-------------|--------------------|------------------------------------------------------------|
| Раннее слияние          | Неявное                        | Низкая      | Низкая             | Рост размерности признакового пространства                 |
| Позднее слияние         | Отсутствует                    | Высокая     | Средняя            | Отсутствие взаимодействия модальностей на уровне признаков |
| Гибридное слияние       | Явное                          | Средняя     | Средняя            | Высокая вычислительная сложность                           |
| Тензорное слияние       | Явное                          | Низкая      | Низкая             | Экспоненциальный рост размерности                          |
| Нейросимвольное слияние | Явное                          | Высокая     | Высокая            | Необходимость построения базы знаний                       |



Сравнение подходов показывает, что наиболее высокое качество анализа достигается при использовании методов, учитывающих взаимодействие между модальностями на уровне внутренних представлений данных [46]. Особый интерес представляют гибридные и нейросимвольные подходы, поскольку они позволяют одновременно учитывать межмодальные зависимости и сохранять структуру предметных знаний. Использование графов знаний в составе мультимодальных моделей создаёт дополнительные возможности для интерпретации результатов анализа и формирования объяснений [116].

Развитие гибридных подходов во многом обусловлено появлением кросс-модальных архитектур внимания [117] и предобученных моделей совместного представления изображений и текста, таких как CLIP [118]. Дальнейшее развитие привело к появлению мультимодальных фундаментальных моделей (GPT-4 [119], Gemini [120]) и универсальных архитектур, связывающих несколько модальностей в едином пространстве эмбедингов (ImageBind [121]). Вопросы эффективности механизмов внимания исследуются в работах [122], показывающих, что значительная часть голов внимания может быть удалена без существенной потери качества.

### 1.3 Графы знаний в задачах анализа данных

Граф знаний (knowledge graph) представляет собой формализованное представление предметной области в виде ориентированного графа, узлы которого соответствуют сущностям (объектам, понятиям, характеристикам), а рёбра – семантическим отношениям между ними [123]. В отличие от реляционных баз данных, где связи между таблицами не имеют явной семантики, в графе знаний каждое ребро типизировано и несёт смысловую нагрузку: «является», «часть», «связан с», «влияет на» и т. д. Благодаря этому граф знаний позволяет хранить факты и строить цепочки рассуждений, прослеживая пути от одного понятия к другому через систему связей. Например, в медицинской диагностике граф знаний может связывать симптомы, заболевания и методы лечения, позволяя проследить путь «симптом – возможный диагноз – рекомендуемое обследование». В контексте настоящего исследования граф знаний используется для явного представления связей между признаками цифрового следа и прогнозируемыми характеристиками личности, что обеспечивает интерпретируемость результатов анализа.

Онтологические паттерны проектирования (Ontology Design Patterns) представляют собой переиспользуемые решения для моделирования типовых ситуаций предметной области, позволяющие отделить инвариантную часть модели от доменно-специфичных компонентов. В развитие данного направления автором предложен

доменно-инвариантный онтологический паттерн «объект анализа – признак – целевая характеристика» для интерпретируемых нейросимвольных систем [204\*]. Паттерн обобщает ранее разработанную онтологию психопрофилирования пользователей социальных сетей [211\*] до уровня переиспользуемого проектного решения. Возможность его применения в иной предметной области продемонстрирована на задаче дерматоскопической диагностики новообразований кожи.

Истоки представления знаний в виде графовых структур начинаются с моделей семантической памяти Квиллиана М. Р. [124] и фреймов Минского М. [125]. Современное понятие графа знаний систематизировано в обзоре Хогана А. и соавторов [123], а его практическое внедрение обозначено публикацией Сингхала А. [126], сформулировавшего принцип «вещи, а не строки». Графы знаний находят применение в задачах семантического поиска, рекомендательных систем и интеллектуального анализа данных. Крупные базы знаний, такие как DBpedia, ConceptNet и BabelNet, служат основой для многих прикладных решений [127, 128, 129].

Для интеграции графов знаний с методами машинного обучения разработаны методы векторного представления графов знаний (Knowledge Graph Embeddings, KGE), отображающие сущности и отношения в непрерывное векторное пространство [26, 130]. Различные семейства методов отличаются способом моделирования отношений: трансляционные модели (TransE [23], TransH [131]), модели на основе скалярных произведений (DistMult [24]) и модели в комплексном пространстве (ComplEx [25], RotatE [132]). Детальный анализ существующих подходов и обоснование выбора используемых методов приведены в последующих параграфах работы.

Несмотря на существенное развитие методов представления и обработки графов знаний, ряд задач остаётся открытым [133]. Большинство методов KGE ориентировано на работу со статическими графами и не учитывает динамику изменения знаний во времени. Формируемые векторные представления преимущественно используются для решения задач представления и дополнения знаний, тогда как вопросы их интеграции с нейросетевыми моделями обработки мультимодальных данных исследованы в меньшей степени. Кроме того, переход от символьного представления знаний к векторному пространству затрудняет интерпретацию результатов в терминах исходных сущностей и отношений графа. Указанные ограничения определяют актуальность разработки методов нейросимвольной интеграции, объединяющих нейросетевое обучение и формализованное представление знаний.

#### 1.4 Графовые нейронные сети

Графовые нейронные сети (Graph Neural Networks, GNN) являются одним из основных подходов к обработке данных, представленных в виде графов [134]. В отличие от традиционных нейронных сетей, ориентированных на регулярные структуры (последовательности, изображения), GNN работают с графами произвольной топологии и учитывают структуру связей между объектами [135]. Основным принцип их функционирования – механизм передачи сообщений (message passing), при котором представления узлов формируются путём последовательной агрегации информации от соседних вершин [136].

Развитие GNN связано с появлением ряда архитектур, различающихся способом агрегации информации.

Графовые свёрточные сети (Graph Convolutional Networks, GCN) [13] – представление узла формируется на основе агрегированной информации о его соседях, веса связей определяются структурой графа.

Графовые сети внимания (Graph Attention Networks, GAT) [14] – вклад каждого соседнего узла определяется обучаемым коэффициентом внимания, вычисляемым на основе признаков взаимодействующих вершин. GAT учитывает неодинаковую значимость соседних узлов; коэффициенты внимания интерпретируемы по построению и отражают вклад каждого соседа в формирование представления узла.

Метод GraphSAGE [27] – предназначен для индуктивного обучения на графах, допускает обработку ранее не встречавшихся узлов.

Реляционные графовые сети (R-GCN) [28] – ориентированы на работу с графами, содержащими несколько типов отношений между сущностями, что актуально для графов знаний.

Применение GNN к графам знаний представляет особый интерес для задач интеллектуального анализа данных [137]. В этом случае структура графа отражает предметные знания, а механизм обучения позволяет учитывать статистические закономерности, содержащиеся в данных. Сочетание структурного и статистического подходов рассматривается как одно из направлений интеграции символьных и нейросетевых методов [11].

Развитие графовых архитектур для работы с мультимодальными данными привело к появлению реляционных и гетерогенных сетей: R-GCN [138], R-GAT [139], HAN [140], гиперболических графовых трансформеров [141] и масштабируемых графовых сетей [142]. Систематизация гетерогенных графовых

сетей приведена в обзоре [143]. В системах поддержки принятия решений мультимодальный анализ применяется для интеграции разнородных источников при выработке решений [144, 145]. Предложены мультимодальные графовые архитектуры интеграции и кросс-модального поиска [146, 147], а также реляционные графовые модели для мультимодальных задач [148]. Существующие подходы к работе с мультимодальными данными на графах включают иерархические механизмы внимания (MM-HGNN [149]), сквозное обучение структуре графа (HGN [150]), явное обучение кросс-модальных связей (HGCN [151]), гиперграфовые модели (HHN [152]) и взвешенное обучение векторных представлений (WeightedHGE [153]).

Принципиально, что в методах данной группы граф знаний выступает предметом предсказания: задача состоит в восстановлении недостающих рёбер и сущностей самого графа, а мультимодальные описания служат вспомогательной информацией. В настоящей работе постановка обратная: граф знаний конструируется на основе заданной экспертом предметной онтологии и выступает априорной структурой, информирующей нейронную сеть при решении внешней задачи анализа объекта; интерпретируемость решения (объясняющие пути, метрика GPI) при этом является целевым свойством, а не побочным. Кроме того, табличная (числовая) модальность в моделях МКГ сводится к атрибутам сущностей, тогда как в предлагаемом методе она обрабатывается как равноправная модальность с собственными доменами и типами.

В настоящей работе в качестве основной архитектуры выбрана графовая сеть внимания GAT. Этот выбор обусловлен следующими факторами:

- механизм внимания позволяет учитывать различную значимость связей между признаками и целевыми характеристиками;
- коэффициенты внимания могут быть использованы для построения интерпретируемых объяснений в виде семантических путей графа знаний;
- GAT демонстрирует высокую эффективность на гетерогенных графах, к которым относятся графы знаний.

Использование GNN в составе графов знаний создаёт основу для построения нейросимвольных систем, объединяющих формализованное представление знаний и методы машинного обучения. Данное направление рассматривается как одно из наиболее перспективных для создания интерпретируемых интеллектуальных систем.

Сопоставление существующих графовых архитектур и обоснование выбора модели приведены в последующих главах диссертации.

## 1.5 Проблема интерпретируемости нейросетевых моделей

Интерпретируемость машинного обучения (Interpretable Machine Learning, IML) рассматривается как способность человека понимать причины и логику формирования результатов работы модели [31]. В литературе выделяются несколько взаимосвязанных понятий: интерпретируемость, объяснимость и прозрачность [30]. Под интерпретируемостью обычно понимается возможность анализа механизма формирования решения, под объяснимостью – представление результатов в форме, доступной для восприятия человеком, а под прозрачностью – доступность внутренней структуры модели для исследования и анализа. Актуальность проблемы интерпретируемости подтверждается в современных обзорах [155], систематизирующих вызовы и перспективы развития данного направления.

Существуют различные подходы к классификации интерпретируемости. В работе Липтона З. [156] выделяются такие свойства интерпретируемых моделей, как симулируемость, предполагающая возможность воспроизведения логики работы модели человеком, и декомпозируемость, связанная с интерпретацией отдельных компонентов модели и их вклада в итоговый результат.

Значимость интерпретируемости возрастает по мере расширения областей применения методов машинного обучения. Как отмечает Рудин С. [9], в социально значимых сферах недостаточно получить высокую точность прогнозирования; требуется понимание причин, по которым модель формирует то или иное решение. Интерпретируемость облегчает выявление ошибок, систематических смещений и некорректных зависимостей в данных, а также способствует повышению доверия пользователей к интеллектуальным системам. Дополнительное значение данному направлению придают нормативные требования в области регулирования искусственного интеллекта и обработки персональных данных [157].

Доши-Велес Ф. и Ким Б. [10] подчёркивают особую важность интерпретируемости в условиях ограниченности данных, высокой стоимости ошибки и возможного расхождения между обучающей и эксплуатационной средой модели. В таких ситуациях объяснение результатов становится дополнительным инструментом оценки корректности принимаемых решений.

В зависимости от способа формирования объяснений существующие подходы к интерпретации результатов машинного обучения принято разделять на встроенные и апостериорные. Встроенные методы предполагают, что интерпретируемость является свойством самой модели и учитывается при её проектировании. Апостериорные методы формируют объяснение после получения

результата путём анализа уже обученной модели. В последнее время активно развивается направление самоинтерпретируемых нейронных сетей (Self-Interpretable Neural Networks), в которых объяснение является встроенным свойством модели [155]. Данное различие имеет принципиальное значение для настоящего исследования, поскольку рассматриваемый нейросимвольный подход относится к классу моделей со встроенной интерпретируемостью.

#### 1.5.1 Методы апостериорной интерпретации

Методы апостериорной интерпретации формируют объяснение после получения результата работы модели, тогда как во встроенных подходах интерпретируемость учитывается на этапе проектирования [158].

К числу наиболее известных апостериорных методов относятся:

- LIME (Local Interpretable Model-agnostic Explanations) [15] – строит локальную интерпретируемую модель в окрестности исследуемого объекта, аппроксимируя поведение исходной модели.
- SHAP (SHapley Additive exPlanations) [16] – использует значения Шэпли из теории кооперативных игр для оценки вклада признаков в формирование результата.
- Grad-CAM [29] – формирует визуальное объяснение решений свёрточных нейронных сетей на основе анализа градиентов целевого класса.

Помимо этих методов, разработаны подходы на основе анализа градиентов и активаций [159, 160], концептуальные объяснения [161, 162], контрфактические объяснения [163], а также методы анализа внутренних механизмов трансформерных моделей [164, 165]. Обобщение подходов к оценке качества объяснений представлено в работе [166].

Несмотря на широкое распространение, апостериорные методы имеют ряд ограничений [167, 168]:

Формируемые объяснения представляют собой приближение поведения модели и необязательно отражают фактический механизм принятия решения. Различные методы могут формировать несовпадающие объяснения для одного и того же объекта [9]. Объяснения чувствительны к изменениям входных данных [169].

Вычислительная сложность некоторых методов, в частности SHAP, существенно возрастает с увеличением числа признаков.

Апостериорные методы работают с уже сформированным результатом и, как правило, не используют предметные знания при построении объяснений. Вследствие

этого затрудняется проверка согласованности объяснений с экспертными представлениями.

В работе [170] показано, что отдельные модели могут демонстрировать дискриминационное поведение, оставаясь при этом внешне корректными с точки зрения методов LIME и SHAP.

Указанные ограничения обусловили развитие подходов со встроенной интерпретируемостью, в которых объяснение формируется непосредственно в процессе получения результата и связано со структурой используемой модели. Эти подходы рассматриваются в следующем подпараграфе.

Помимо методов интерпретации, ориентированных на полносвязные и свёрточные архитектуры, разработаны специализированные подходы для графовых нейронных сетей. К числу апостериорных методов объяснения решений GNN относятся GNNExplainer [171], PGExplainer [172] и GraphMask [173], выделяющие значимые подграфы и рёбра, оказавшие наибольшее влияние на результат. Обобщение методов интерпретируемости графовых нейронных сетей представлено в обзорах [174, 175].

В отличие от перечисленных апостериорных методов, восстанавливающих объясняющий подграф для уже обученной модели «чёрного ящика», в настоящей работе объяснение является встроенным: значимые пути в графе знаний формируются непосредственно в процессе вывода, что делает применение внешних методов постфактум избыточным. Предлагаемая метрика GPI (глава 5, приложение П) не конкурирует с указанными методами, а решает иную задачу – количественную оценку верности и полноты интерпретации; её компонент проверки верности (удаление концептов) методологически согласован с критериями точности объяснений (fidelity), применяемыми для оценки апостериорных методов.

### 1.5.2 Встроенная интерпретируемость

Альтернативу апостериорным методам составляют подходы со встроенной интерпретируемостью (explainability-by-design), в которых возможность интерпретации учитывается на этапе проектирования модели [176]. К данному направлению относятся интерпретируемые архитектуры, основанные на деревьях решений, правилах и обобщённых аддитивных моделях [177], модели с механизмами внимания [178], подходы, использующие промежуточные концептуальные представления (Concept Bottleneck Models) [179], а также самоинтерпретируемые нейронные сети, активно развивающиеся в последнее время [155].

Отдельное направление образуют нейросимвольные модели, объединяющие нейронные сети и формализованные структуры знаний – онтологии и графы знаний [180]. В отличие от апостериорных методов, интерпретация результатов в таких системах связана со структурой модели и может выполняться на уровне используемых концептов, отношений и цепочек вывода. Современные обзоры [69] подчёркивают, что нейросимвольная интеграция является одним из наиболее перспективных направлений для создания интерпретируемых систем искусственного интеллекта.

Следует отметить, что интерпретационные возможности механизмов внимания имеют определённые ограничения. В работе [181] показано, что значения весов внимания не всегда совпадают с фактической значимостью признаков для принимаемого решения. Это обстоятельство стимулирует поиск подходов, в которых интерпретация опирается на статистические зависимости и формализованные предметные знания.

По мнению Гарсеса А. д'Авилы и Лэмба Л. К. [11], нейросимвольная интеграция относится к числу наиболее перспективных направлений развития интерпретируемого искусственного интеллекта, поскольку сочетает обучаемость нейронных сетей с возможностью использования формализованных структур знаний.

Значительный вклад в развитие интеллектуальных систем и методов машинного обучения внесли диссертационные исследования: технологии построения гибридных интеллектуальных систем [182, 183], обеспечения жизнеспособности интеллектуальных систем с базами знаний [184], векторных моделей графов в задачах машинного обучения [185], мультимодального машинного зрения [186] и методов машинного обучения для анализа данных [187].

В области интерпретируемых графовых нейронных сетей разрабатываются также методы со встроенной интерпретируемостью, формирующие объяснение непосредственно в процессе вывода: CI-GNN [188] на основе причинно-следственного анализа и FireGNN [189] с обучаемыми нечёткими правилами. Отдельное направление образуют нейросимвольные методы, представляющие поведение графовой сети в структуре графа знаний [190, 191], и метрики объяснимости на основе путей вывода – Power-Link [192] и PaGE-Link [193]. Предлагаемый в работе подход относится к методам со встроенной интерпретируемостью и отличается организацией объяснений как семантических путей в онтологически типизированном графе знаний и их количественной оценкой метрикой интерпретируемости путей (GPI). К близким относятся методы рассуждения по путям в графах знаний [194].



В настоящей работе развивается именно этот подход: предлагается нейросимвольная архитектура, в которой интерпретируемость обеспечивается за счёт использования графа знаний, структурирующего предметные знания, и графовой сети внимания, формирующей объяснения в виде семантических путей вывода. В отличие от чистых механизмов внимания, объяснение в предлагаемом подходе опирается на формализованные онтологические связи, что позволяет верифицировать его с использованием предметных знаний.

#### 1.6 Постановка научной проблемы и гипотеза исследования

Проведённый анализ показал наличие ряда ограничений существующих подходов к обработке мультимодальных данных.

Ограничения методов мультимодального слияния связаны с трудностью одновременного учёта межмодальных зависимостей и сохранения особенностей отдельных модальностей. Раннее слияние приводит к увеличению размерности представления и частичной потере модальностной специфики; позднее слияние ограничивает взаимодействие между модальностями; гибридные методы учитывают межмодальные связи, однако формируемые представления, как правило, не связаны с предметными знаниями.

Ограничения методов работы с графами знаний связаны с недостаточной интеграцией символьного представления знаний и нейросетевых моделей анализа данных. Большинство методов векторного представления графов знаний ориентировано на формирование эмбедингов сущностей и отношений и не рассматривает их совместное обучение с моделями извлечения признаков. Кроме того, структура графа знаний обычно остаётся фиксированной и не адаптируется в процессе обучения.

Ограничения существующих методов интерпретации связаны с преобладанием апостериорных подходов. Формируемые объяснения не всегда отражают фактический механизм принятия решения, не опираются на предметные знания и не предусматривают их формальной верификации. Для нейросимвольных моделей также отсутствуют общепринятые количественные методы оценки интерпретируемости.

Ограничения существующих подходов к нейросимвольной интеграции связаны с тем, что нейросетевые представления и графовые структуры знаний в большинстве случаев рассматриваются как отдельные компоненты системы. При этом недостаточно исследованы механизмы их совместного использования при формировании результата и построении объяснений в виде цепочек рассуждений по графу знаний.

Выявленные ограничения определяют научную задачу разработки методов интеллектуальной обработки мультимодальных данных, объединяющих нейросетевые модели и графовые структуры знаний в рамках единой интерпретируемой архитектуры. Решение данной задачи предполагает совместное использование мультимодальных представлений и предметных знаний, формирование объяснений непосредственно в процессе вывода и разработку методов количественной оценки интерпретируемости получаемых результатов.

Гипотеза исследования состоит в том, что нейросимвольная интеграция графа знаний, построенного по доменно-инвариантному онтологическому паттерну «объект анализа – признак – целевая характеристика», и графовой нейронной сети обеспечивает встроенную интерпретируемость результатов анализа мультимодальных данных. Объяснения формируются в виде семантических путей вывода, допускающих экспертную проверку и количественную оценку. При этом качество прогнозирования сохраняется на уровне, сопоставимом со сквозными мультимодальными моделями, а при переносе между предметными областями сохраняется инвариантное ядро метода, тогда как изменяются предметное наполнение онтологии, состав признаков и целевых характеристик.

В контексте настоящего исследования важной задачей является разграничение объяснимости и интерпретируемости. Под объяснимостью понимается способность построить объяснение поведения уже обученной модели постфактум (апостериорные методы), тогда как интерпретируемость трактуется как свойство модели, при котором процесс вывода понятен по построению. В предлагаемом подходе объяснение формируется не как отдельная аппроксимация, а как семантический путь рассуждения в графе знаний, что соответствует встроенной интерпретируемости. Как отмечается в современных обзорах [30], это разграничение имеет принципиальное значение для проектирования систем, ориентированных на человека. Настоящая работа направлена именно на встроенную интерпретируемость, а не на апостериорную объяснимость.

### 1.7 Нейросимвольная интеграция графов знаний и графовых нейронных сетей

Принципиальное отличие от классических нейронных сетей состоит в следующем. Полносвязные, свёрточные и трансформерные сети работают с входом фиксированной структуры и остаются «чёрным ящиком» – их внутренние веса не интерпретируемы. Графовая нейронная сеть учитывает отношения между сущностями, однако над произвольным графом она также остаётся непрозрачной. В предлагаемом подходе графовая сеть применяется над графом знаний, структура которого

семантически осмыслена, поэтому сам путь вывода является человеко-проверяемым объяснением. Тем самым осуществляется переход от модели «чёрный ящик» к встроенной интерпретируемости: классическая сеть даёт ответ, но не объяснение, а графовая сеть над графом знаний – ответ вместе с проверяемым путём рассуждения в терминах предметной области.

Вместе с тем перечисленные подходы развивают отдельные аспекты задачи: одни – точность слияния модальностей, другие – апостериорное объяснение уже обученной модели, третьи – интерпретируемость в рамках одной предметной области. Ни один из них не объединяет эти аспекты в единую доменно-инвариантную систему встроенной интерпретируемости. Научная новизна настоящей работы состоит не в отдельных компонентах (каждый из которых имеет известные аналоги и на которые работа опирается), а в их интеграции в доменно-инвариантную систему: онтологический паттерн описания данных, нейросимвольная архитектура «граф знаний и графовые нейронные сети» с обучением, учитывающим онтологию, метрика интерпретируемости путей GPI и экспериментально подтверждённая переносимость метода между предметными областями. Результатом является метод, модели и метрика, проверенные на нескольких независимых предметных областях.

Систематизация рассмотренных подходов приведена на рисунке 1.

|           | Без предметных знаний                                                                                                                                                                  | Последовательная                                                               | Гибридная (дополнение MKG)                                                           | Информирование НС                                                   |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| 3 и более | <ul style="list-style-type: none"> <li>● GPT-4, Gemini, ImageBind</li> </ul>                                                                                                           |                                                                                | <ul style="list-style-type: none"> <li>● MKGformer, MoSE (дополнение MKG)</li> </ul> | <div> ★ Данная работа: онтологии, GNN, объясняющие пути, GPI </div> |
| 2         | <ul style="list-style-type: none"> <li>● CLIP</li> </ul>                                                                                                                               | <ul style="list-style-type: none"> <li>● GNN-RAG (LLM, граф знаний)</li> </ul> |                                                                                      |                                                                     |
| 1         | <ul style="list-style-type: none"> <li>● Классические DNN (CNN, RNN, трансформеры)</li> <li>● R-GAT, HGT (ретегор. графы)</li> <li>● GNNE explainer, PGExplainer, GraphMask</li> </ul> | <ul style="list-style-type: none"> <li>● KGE: TransE, DistMult</li> </ul>      | <ul style="list-style-type: none"> <li>● pLogicNet (правила, KGE)</li> </ul>         | <ul style="list-style-type: none"> <li>● CI-GNN, FireGNN</li> </ul> |

● «чёрный ящик»   
● апостериорная интерпретация   
● встроенная интерпретируемость   
★ предлагаемый метод

Рисунок 1 – Систематизация методов анализа мультимодальных данных

По способу нейросимвольной интеграции выделяются три класса [11]: последовательные (символьный и нейросетевой компоненты работают поочерёдно), гибридные (компоненты взаимодействуют как отдельные подсистемы) и

интегративные (символьные знания встроены в структуру и процесс обучения нейронной сети). По числу одновременно обрабатываемых модальностей методы группируются от одномодальных до охватывающих три и более; по способу формирования объяснений: от моделей «чёрного ящика» через апостериорные методы к встроенной интерпретируемости.

Существующие методы развивают отдельные аспекты задачи: мультимодальные фундаментальные модели (CLIP, GPT-4, ImageBind) обрабатывают три и более модальностей, но не используют предметных знаний и не интерпретируемы; методы мультимодальных графов знаний (MKGformer, MoSE) объединяют модальности вокруг графа, но решают задачу дополнения самого графа; методы со встроенной интерпретируемостью (CI-GNN, FireGNN) ограничены одной модальностью. Предлагаемый метод занимает не занятую существующими подходами область: интегративная нейросимвольная модель (информирование нейронной сети предметной онтологией) с тремя модальностями и встроенной интерпретируемостью.

## 1.8 Математическая постановка задачи

Пусть объект анализа  $o$  представлен мультимодальными данными

$$M(o) = \langle T, V, D \rangle,$$

где  $T$  – текстовая,  $V$  – визуальная,  $D$  – табличная (числовая) модальности. После извлечения признаков каждой модальности формируется единое признаковое описание объекта

$$x = \Phi(M(o)) \in X,$$

где  $\Phi(\cdot)$  – совокупность модальных энкодеров.

В классической постановке задачи машинного обучения требуется построить отображение

$$f: X \rightarrow Y,$$

которое по признаковому описанию объекта определяет значения целевых характеристик

$$y = (y_1, \dots, y_K),$$

минимизируя функцию ошибки на обучающей выборке.

Однако в рассматриваемой задаче целевые характеристики не являются независимыми. Они описывают один и тот же объект и обладают известными содержательными взаимосвязями. Независимое прогнозирование каждой характеристики приводит к возникновению противоречивых сочетаний, когда

отдельные предсказания оказываются согласованными с данными по отдельности, но несовместимыми между собой с точки зрения предметной области.

Кроме того, признаки различных модальностей характеризуют объект не напрямую, а через различные группы промежуточных интерпретируемых признаков. Следовательно, требуется учитывать два типа зависимостей:

- связи между признаками и целевыми характеристиками;
- взаимосвязи между самими целевыми характеристиками.

Для описания этих зависимостей вводится предметная онтология

$$O = (C, R),$$

где  $C$  – множество концептов, а  $R$  – множество семантических отношений. Онтология задаёт структуру предметных знаний и определяет допустимые связи между признаками и характеристиками, а также взаимосвязи между характеристиками.

Совместное использование мультимодальных признаков и онтологии формирует гетерогенный граф знаний

$$G = (N, E),$$

в котором узлами являются объекты, признаки и целевые характеристики, а рёбра отражают отношения имеет\_признак, ассоциируется\_с и коррелирует\_с. Граф знаний служит структурой распространения информации для графовой нейронной сети и обеспечивает включение предметных знаний в процесс обучения.

Таким образом, задача сводится к минимизации ошибки прогнозирования и сохранению известных предметных взаимосвязей между прогнозируемыми характеристиками.

В основном эксперименте обучение осуществляется посредством минимизации функции потерь

$$\theta^* = \arg \min_{\theta} (L_{MSE}(\theta) + \lambda L_{corr}(\theta)),$$

где

$$L_{corr} = \sum_{k < l} w_{kl} (\rho_{kl}^{pred} - \rho_{kl}^{true})^2,$$

а веса

$$w_{kl} = |\rho_{kl}^{true}|$$

определяются силой известных взаимосвязей между характеристиками.

При этом предикат «коррелирует\_с» реализуется через корреляционный член функции потерь, обеспечивающий сохранение согласованной структуры прогнозируемых характеристик, а предикат «ассоциируется\_с» реализуется

структурой графа знаний, содержащего только статистически значимые связи между признаками и характеристиками. Качество интерпретации получаемых решений оценивается с использованием метрики GPI.

Таким образом, предложенная постановка представляет собой классическую задачу машинного обучения, дополненную онтологическими ограничениями, которые учитываются при оптимизации модели, а также при построении структуры графа знаний.

## 1.9 Выводы по главе 1

1. Проведён анализ современных методов обработки мультимодальных данных. Показано, что мультимодальные данные характеризуются гетерогенностью, различием пространств признаков и наличием семантических зависимостей между модальностями. Установлено, что существующие методы мультимодального слияния в различной степени учитывают межмодальные взаимосвязи, однако не используют явное представление предметных знаний и не ориентированы на формирование встроенных интерпретируемых объяснений.

2. Выполнен анализ методов представления и обработки знаний на основе графовых моделей. Рассмотрены подходы к построению графов знаний и методы их векторного представления (TransE, DistMult, ComplEx, RotatE). Показано, что существующие подходы преимущественно ориентированы на формирование эмбедингов сущностей и отношений и в ограниченной степени учитывают совместное использование графов знаний и нейросетевых моделей анализа данных, а также не используют структуру графа для формирования интерпретируемых цепочек рассуждений.

3. Проведён анализ архитектур графовых нейронных сетей (GCN, GAT, GraphSAGE, R-GCN). Установлено, что графовые сети внимания учитывают неодинаковую значимость связей между узлами при формировании представлений и могут использоваться при построении интерпретируемых моделей анализа данных на графовых структурах. Механизм внимания создаёт предпосылки для объяснения процесса агрегации информации в терминах значимости связей между сущностями.

4. Исследованы современные подходы к интерпретации результатов машинного обучения. Анализ апостериорных методов интерпретации показал, что формируемые объяснения не связаны непосредственно с процессом получения результата и не предусматривают их проверки относительно предметных знаний. Рассмотрены подходы со встроенной интерпретируемостью, включая нейросимвольные методы, объединяющие нейросетевые модели и формализованные

структуры знаний. Показано, что наиболее перспективным направлением является интеграция графов знаний и графовых нейронных сетей, обеспечивающая встроенную интерпретируемость результатов.

5. Проведённый анализ показал, что существующие методы мультимодального анализа, графового представления знаний и интерпретации результатов не обеспечивают совместного решения задач интеграции мультимодальных данных, использования предметных знаний и формирования встроенных интерпретируемых объяснений. Выявленные ограничения определяют направление дальнейшего исследования, посвящённого разработке методов нейросимвольной интеграции графов знаний и графовых нейронных сетей для интерпретируемого анализа мультимодальных данных. Математическая формализация данной задачи, учитывающая онтологические ограничения, представлена в параграфе 1.8, где также дано позиционирование предлагаемого подхода в контексте существующих решений.

## 2 Метод нейросимвольной интеграции графов знаний и графовых нейронных сетей для интерпретируемого анализа мультимодальных данных

Научная гипотеза исследования заключается в том, что совместное использование графов знаний и графовых нейронных сетей в рамках многоуровневой нейросимвольной архитектуры позволит учитывать как статистические закономерности, содержащиеся в мультимодальных данных, так и формализованные предметные знания, представленные в графе знаний. При этом интерпретация результатов анализа может осуществляться посредством извлечения семантических путей вывода в пространстве графа знаний.

Гипотеза основывается на следующих предпосылках:

1. Графы знаний задают формализованное представление семантических связей между объектами, признаками и целевыми характеристиками предметной области.
2. Графовые сети внимания учитывают структуру графа при формировании представлений узлов, а коэффициенты внимания отражают относительную значимость связей в процессе распространения информации.
3. Нейросимвольная интеграция объединяет возможности обучения на данных с использованием формализованных структур знаний.
4. Представление объяснений в виде семантических путей графа знаний допускает их анализ и сопоставление с предметными знаниями.

Целью исследования является разработка метода интерпретируемого анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, интегрирующей графы знаний и графовые нейронные сети для информирования нейросетевой модели предметными знаниями и формирования интерпретируемого результата анализа.

В рамках разработки предлагаемого метода в настоящей главе решаются следующие задачи:

1. Анализ существующих методов представления, интеграции и интерпретируемого анализа мультимодальных данных.
2. Разработка графовой модели описания мультимодальных данных для совместного представления текстовых, визуальных и табличных данных в единой структуре.
3. Разработка модели графа знаний для семантической интеграции мультимодальных данных и предметных знаний.



4. Разработка нейросимвольного метода интерпретируемого анализа мультимодальных данных на основе совместного использования графов знаний и графовых нейронных сетей.

5. Разработка модели формирования и количественной оценки интерпретируемых результатов анализа.

6. Разработка программного комплекса и проведение апробации предложенных методов и моделей на задаче анализа цифрового следа пользователей социальной сети. Предметная независимость предложенных методов и моделей дополнительно подтверждается на независимых наборах данных других предметных областей (дерматоскопическая диагностика, классификация токсичных комментариев, эмоциональная классификация – приложения К, Л).

## 2.1 Концептуальная модель нейросимвольной интеграции мультимодальных данных

Современные системы поддержки принятия решений, выполняющие анализ мультимодальных данных, сталкиваются с двумя взаимосвязанными ограничениями. С одной стороны, нейросетевые модели широко применяются для извлечения признаков из разнородных данных, включая текстовую, визуальную и табличную информацию. С другой стороны, процесс формирования результатов в таких моделях зачастую недостаточно прозрачен для пользователя. В то же время символьные методы, основанные на использовании онтологий, экспертных правил и формализованных представлений знаний, позволяют описывать причинно-следственные связи и формировать интерпретируемые выводы, однако обладают ограниченными возможностями адаптации к большим объёмам неструктурированных данных.

Указанные ограничения приобретают особую значимость в интеллектуальных системах поддержки принятия решений, где наряду с высокой точностью анализа требуется возможность интерпретации результатов, позволяющая установить признаки и взаимосвязи, оказавшие наибольшее влияние на формирование вывода.

В рамках настоящего исследования решение данной задачи рассматривается на уровне архитектуры интеллектуальной системы, реализующей нейросимвольную интеграцию и объединяющей преимущества нейросетевого обучения и формализованного представления знаний.

Предлагаемая концептуальная модель основана на последовательном преобразовании мультимодальных данных в интерпретируемое представление и включает четыре функциональных уровня:

- уровень обработки данных;
- уровень семантического связывания;
- уровень нейросимвольного вывода;
- уровень формирования объяснений.

На первом уровне осуществляется извлечение признаков из разнородных модальностей входных данных объекта анализа. Для каждой модальности используются специализированные методы обработки, учитывающие особенности соответствующего типа данных. Результатом данного уровня являются векторные представления признаков, отражающие содержательные характеристики анализируемого объекта.

Особенностью первого уровня является параллельная обработка различных модальностей. Параллельная обработка различных модальностей даёт возможность учитывать взаимодополняющие аспекты объекта анализа, которые не выявляются при использовании только одного источника информации. Полученные представления передаются на следующий уровень для дальнейшей семантической интеграции.

На втором уровне выполняется семантическое связывание извлечённых признаков посредством графа знаний. Граф знаний представляет собой формализованную структуру, описывающую взаимосвязи между объектом анализа, выявленными признаками и прогнозируемыми характеристиками предметной области.

Теоретическую основу предложенного представления составляют формальные основы онтологий [195], классические результаты теории графов [196, 197] и методы векторного представления графов знаний [198].

Структура графа знаний включает три основных типа сущностей:

- объект анализа;
- признаки объекта анализа;
- целевые характеристики.

Связи между сущностями отражают наличие признаков у объекта анализа, ассоциации между признаками и целевыми характеристиками, а также взаимосвязи между самими целевыми характеристиками. В отличие от традиционных признаковых моделей, граф знаний явно представляет взаимосвязи между признаками и целевыми характеристиками предметной области и включает их в процесс анализа.

Граф знаний формируется динамически на основе признаков, извлечённых из входных данных, в соответствии с онтологией предметной области. В результате создаётся семантическое представление объекта анализа, объединяющее статистические характеристики данных и предметные знания в единой структуре.

На третьем уровне реализуется нейросимвольный вывод. В качестве вычислительного механизма используются графовые нейронные сети, осуществляющие обработку сформированного графа знаний. В процессе работы графовой нейронной сети представления узлов уточняются посредством многократной агрегации информации от связанных сущностей.

В отличие от традиционных методов анализа признакового пространства, графовая нейронная сеть учитывает непосредственные и опосредованные взаимосвязи между объектом анализа, признаками и целевыми характеристиками. При формировании прогнозов учитываются как характеристики отдельных сущностей, так и сложные взаимосвязи между элементами предметной области, представленные в графе знаний.

Дополнительным источником информации о процессе вывода являются коэффициенты внимания графовой нейронной сети, отражающие относительную значимость связей между узлами при формировании их представлений. Данные коэффициенты используются на следующем уровне для построения интерпретации результатов.

Результатом работы третьего уровня являются прогнозируемые значения целевых характеристик, а также параметры, характеризующие вклад отдельных элементов графа знаний в процесс формирования предсказания.

На четвёртом уровне выполняется формирование объяснений полученных результатов. Для этого используются структура графа знаний и информация о значимости связей, полученная на предыдущем этапе.

Процесс формирования объяснений включает:

1. Выделение цепочек связей, соединяющих объект анализа с прогнозируемыми характеристиками через систему признаков.
2. Ранжирование найденных путей по степени их влияния на результат анализа.
3. Оценку вклада различных модальностей в формирование предсказания.

Результатом данного этапа является граф объяснений – подграф графа знаний, содержащий наиболее значимые узлы и связи, участвовавшие в формировании

результата. Полученное объяснение может быть представлено в графической форме либо преобразовано в текстовое описание на естественном языке.

Таким образом, предложенная концептуальная модель реализует последовательный переход от мультимодальных данных к интерпретируемым результатам анализа. В отличие от традиционных подходов граф знаний участвует непосредственно в процессе формирования предсказаний, а не используется исключительно для представления предметных знаний или последующего объяснения результатов. Благодаря этому нейросетевые методы обработки данных и формализованное представление знаний объединяются в рамках единой нейросимвольной архитектуры.

Несмотря на то, что дальнейшая апробация предложенного метода выполняется на задаче анализа цифрового следа пользователей социальной сети, разработанные модели и алгоритмы рассматриваются как доменно-инвариантные и могут быть адаптированы к другим предметным областям посредством изменения состава признаков, целевых характеристик и предметных знаний. Переносимость подтверждена экспериментально на дерматоскопической диагностике, классификации токсичных комментариев, классификации эмоций (приложения К, Л).

## 2.2 Принципы нейросимвольной интеграции

Предложенная архитектура основана на ряде методологических положений, определяющих особенности совместного использования нейронных моделей и графовых представлений знаний при анализе мультимодальных данных.

1. На разных уровнях архитектуры реализовано разделение статистического обучения и предметных знаний. Нейронные модели Уровня 1 и графовые нейронные сети Уровня 3 выполняют обучение на данных и выявление закономерностей, содержащихся в наблюдаемых признаках. Граф знаний Уровня 2 формируется на основе онтологии предметной области и отражает формализованные представления о взаимосвязях между объектами, признаками и целевыми характеристиками. В результате обучение на данных дополняется использованием формализованного описания предметной области.

2. В архитектуре реализовано двустороннее взаимодействие статистического и символьного уровней. В процессе анализа информация распространяется от входных данных к признакам и далее к целевым характеристикам. Одновременно в ходе обучения уточняются веса связей, используемые при распространении информации по графу знаний, вследствие чего семантическая структура адаптируется к наблюдаемым

данным. В результате граф знаний служит одновременно средством представления предметных знаний и компонентом процесса анализа.

3. В предлагаемой архитектуре используется многоуровневое представление информации. Система оперирует представлениями нескольких уровней абстракции:

- низкий уровень – исходные данные (тексты, изображения, табличные записи);
- средний уровень – признаки, извлечённые из различных модальностей;
- высокий уровень – целевые характеристики предметной области.

Связь между указанными уровнями задаётся графом знаний, отражающим семантические отношения между признаками и целевыми характеристиками. Связь между уровнями делает возможным последовательный переход от наблюдаемых данных к содержательной интерпретации результатов анализа.

4. Предлагаемая архитектура формирует объяснения непосредственно в процессе нейросимвольного вывода. Интерпретируемость рассматривается как свойство архитектуры системы, а не как отдельный этап постобработки результатов. Граф знаний задаёт структуру предметных зависимостей, а графовая нейронная сеть использует данную структуру при формировании предсказаний. В результате каждому предсказанию может быть сопоставлен подграф объяснения, содержащий наиболее значимые признаки и связи, участвовавшие в формировании результата.

Апостериорные методы интерпретации могут использоваться дополнительно для анализа вкладов отдельных модальностей, однако основой интерпретации в предлагаемом подходе выступают структура графа знаний и пути вывода, формируемые непосредственно в процессе работы модели.

В настоящей работе разграничиваются понятия интерпретируемости и объяснения. Под интерпретируемостью понимается возможность проследить процесс формирования предсказания через семантические связи графа знаний – от наблюдаемого признака к целевой характеристике. Под объяснением понимается представление данного процесса в форме, доступной для анализа пользователем. Объяснение может быть представлено в виде подграфа вывода, ранжированных путей рассуждений или их текстового описания.

Сформулированные положения отражают основные отличия нейросимвольного подхода от традиционных нейронных и символьных методов. Их сравнительная характеристика приведена в таблице 3.

Таблица 3 – Сравнение нейронных, символьных и нейросимвольных подходов

| Критерий                    | Нейронные методы                  | Символьные методы                                         | Предложенная нейросимвольная архитектура                                                         |
|-----------------------------|-----------------------------------|-----------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Источник знаний             | Данные (эмпирическое обучение)    | Экспертные правила, онтологии                             | Совмещение данных и онтологического знания                                                       |
| Гибкость и адаптивность     | Высокая адаптивность к данным     | Низкая, требует ручного обновления правил                 | Сочетает автоматическое обучение и структурную стабильность                                      |
| Интерпретируемость          | Низкая, «чёрный ящик»             | Высокая, логическая прозрачность                          | Объяснимость по дизайну через граф знаний и механизмы внимания                                   |
| Требования к объёму данных  | Очень высокие (миллионы примеров) | Минимальные                                               | Средние: структура компенсирует нехватку данных                                                  |
| Интеграция знаний           | Отсутствует                       | Полная, но негибкая                                       | Формализованное знание встроено в адаптивную модель                                              |
| Учёт структуры зависимостей | Неявный (в скрытых слоях сети)    | Явный, но фиксированный                                   | Явный и адаптируемый в процессе обучения                                                         |
| Обработка сырых данных      | Возможна напрямую                 | Невозможна без предварительного конструирования признаков | Нейронный уровень выполняет извлечение признаков из сырых данных                                 |
| Учёт неопределённости       | Частичный                         | Отсутствует                                               | Поддерживается механизмами графового обучения                                                    |
| Основной недостаток         | Ограниченная интерпретируемость   | Низкая адаптивность                                       | Более высокая вычислительная сложность                                                           |
| Область применимости        | Массовые задачи анализа данных    | Формализованные предметные области                        | Системы поддержки принятия решений, требующие одновременно высокой точности и интерпретируемости |

Таким образом, интерпретируемость в предложенном подходе является встроенным свойством нейросимвольной архитектуры, а не результатом отдельного этапа постобработки. Граф знаний представляет предметные знания в явной форме, графовая нейронная сеть выполняет обучение на данных, а их совместное использование делает процесс формирования предсказания прослеживаемым. Для количественной оценки качества формируемых объяснений в главе 5 предложена метрика GPI (Graph Path Interpretability), основанная на анализе графа вывода. Метрика характеризует свойства сформированного объяснения и не предназначена для оценки когнитивных процессов человека при его восприятии.

### 2.3 Обоснование выбора архитектуры

Четырёхуровневая нейросимвольная архитектура, представленная на рисунке 2, выбрана исходя из требований к интерпретируемому анализу мультимодальных данных и включает следующие компоненты:

1. Мультимодальная природа данных требует использования специализированных моделей обработки текста, изображений и табличной информации для формирования информативных признаков представлений.

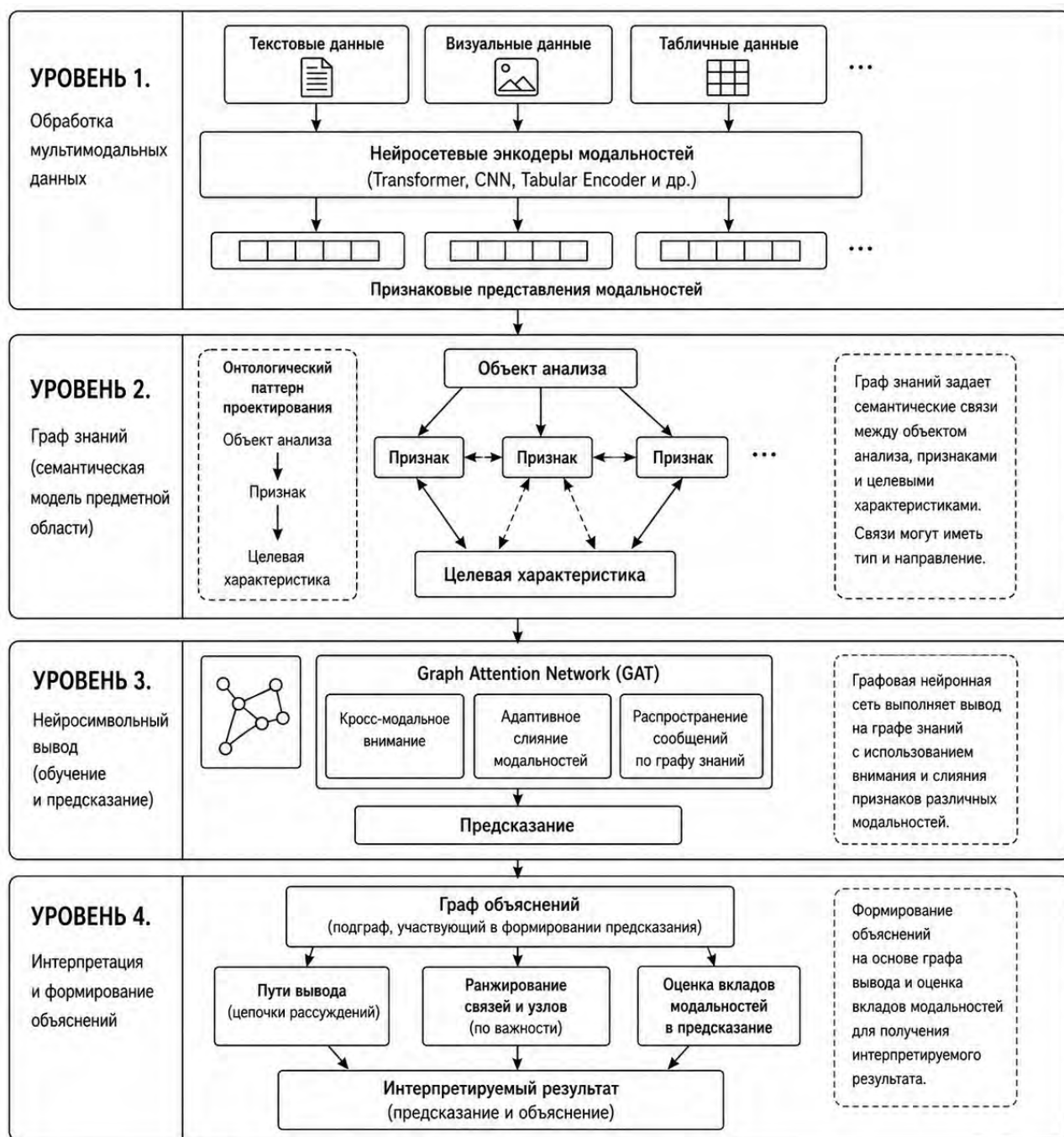


Рисунок 2 – Четырёхуровневая архитектура системы интерпретируемого анализа мультимодальных данных

Приведённая на рисунке 2 архитектура описывает последовательность обработки и не выявляет источник гибридности модели: уровни объединяют как процессы (обработку данных, вывод и формирование объяснений), так и используемые нейросетевые и графовые модели. Собственно гибридность определяется не числом уровней обработки, а наличием двух связанных слоёв представления: вычислительного и семантического – и пары отображений между ними. Двухуровневая структура модели приведена на рисунке 3.

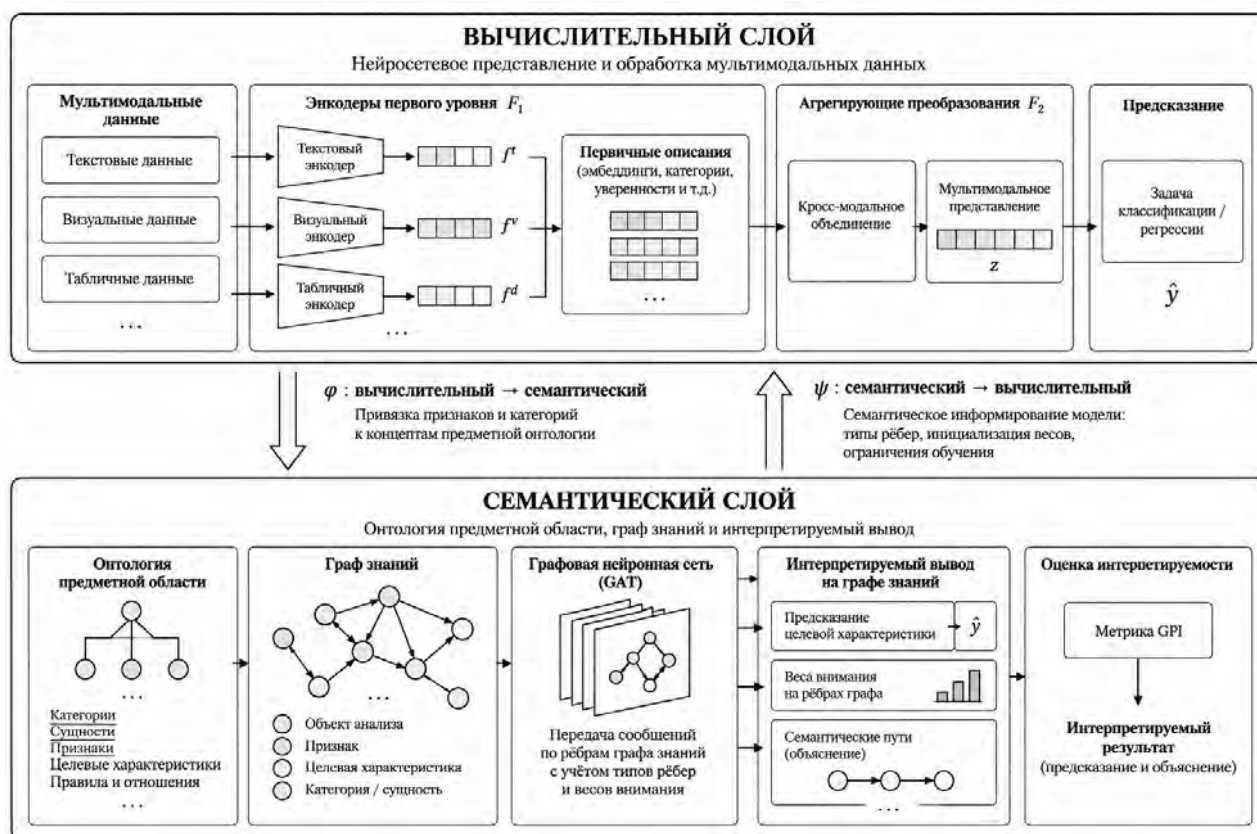


Рисунок 3 – Двухуровневая структура гибридной нейросимвольной модели

Вычислительный слой выполняет нейросетевое представление и обработку мультимодальных данных: энкодеры первого уровня формируют первичные описания, агрегирующие преобразования сводят их к мультимодальному представлению, по которому решается задача классификации или регрессии. Семантический слой содержит предметную онтологию, граф знаний и работающую по нему графовую нейронную сеть, формирующую предсказание вместе с весами внимания на рёбрах и семантическими путями вывода. Слои связаны двумя отображениями:  $\varphi$  переводит признаки и категории вычислительного слоя в концепты предметной онтологии, а  $\psi$  выполняет обратное семантическое информирование модели – задаёт типы рёбер графа знаний, инициализацию весов связей и ограничения обучения. Именно наличие  $\psi$ , то есть влияния символьного представления на обучаемую нейросетевую компоненту, и делает модель гибридной, а не последовательным соединением нейросетевого и графового блоков.

2. Интерпретируемость результатов достигается включением графа знаний, описывающего семантические связи между объектами анализа, признаками и целевыми характеристиками.



3. Предметные знания, представленные в онтологии и графе знаний, позволяют снизить зависимость качества анализа от объёма размеченных данных.

4. Адаптивность архитектуры достигается сочетанием обучаемых нейронных компонентов, учитывающих новые данные, и онтологии, сохраняющей устойчивую структуру предметной области.

5. Граф знаний включает предметные знания непосредственно в процесс анализа, а не только в параметры нейронной сети, что повышает обоснованность и объяснимость принимаемых решений.

Одной из особенностей предложенной архитектуры является перенос части предметных знаний из обучаемых параметров нейронной сети в граф знаний, используемый непосредственно в процессе нейросимвольного вывода. Благодаря этому семантические зависимости задаются в явной форме, а объём знаний, который необходимо кодировать исключительно параметрами нейронной сети, уменьшается.

Таким образом, предложенная архитектура:

- объединяет данные и предметные знания в рамках единой модели;
- поддерживает использование экспертных знаний на этапе формирования предсказаний;
- снижает зависимость качества анализа от объёма обучающей выборки;
- обеспечивает встроенную интерпретируемость результатов;
- допускает перенос на различные предметные области за счёт замены предметной онтологии при сохранении общей структуры метода.

## 2.4 Представление мультимодальных данных

Для строгого изложения предлагаемого метода введём формальные определения модели мультимодальных данных, графа знаний и объясняющего пути, а также сформулируем основные свойства метода и метрики интерпретируемости.

Определение 1. Объектом анализа  $o$  называется сущность, описываемая мультимодальными данными

$$M(o) = \langle T, V, D \rangle,$$

где  $T$  – текстовая модальность,  $V$  – визуальная модальность,  $D$  – табличная (числовая) модальность. Из модальностей извлекается интерпретируемое признаковое описание

$$c(o) = (c_1, \dots, c_d) \in \mathbb{R}^d,$$

где каждый признак  $c_i$  соответствует распознаваемой человеком характеристике одной из трёх модальностей.

Здесь  $c(o)$  обозначает интерпретируемое (именованное) признаковое описание объекта, в отличие от векторного мультимодального представления  $x = \Phi(M(o))$ , введённого в главе 1.

Определение 2. Графом знаний называется ориентированный типизированный граф  $G = (N, E)$  со множеством узлов

$$N = \{o\} \cup F \cup P, \quad F = \{f_1, \dots, f_d\}, \quad P = \{p_1, \dots, p_K\},$$

где  $o$  – узел объекта анализа,  $F$  – узлы признаков,  $P$  – узлы целевых характеристик, и множеством рёбер  $E$  трёх типов (предикатов):

- 1) имеет\_признак:  $o \rightarrow f_i$  с весом  $c_i$  (значение признака);
- 2) ассоциируется\_с:  $f_i \rightarrow p_k$  с весом  $r_{ik}$  (мера статистической связи признака  $f_i$  с целевой характеристикой  $p_k$ );
- 3) коррелирует\_с:  $p_k \leftrightarrow p_l$  с весом  $\rho_{kl}$  (корреляция целевых характеристик).

Определение 3. Объясняющим путём для предсказания целевой характеристики  $p_k$  объекта  $o$  называется путь  $\pi_{ik} = (o, f_i, p_k)$ , образованный ребром имеет\_признак ( $o \rightarrow f_i$ ) и ребром ассоциируется\_с ( $f_i \rightarrow p_k$ ).

Путь  $\pi_{ik}$  называется значимым, если одновременно:

- признак присутствует у объекта,  $c_i > \tau$  (порог присутствия);
- ребро ассоциируется\_с статистически значимо после поправки на множественность проверки гипотез (проверяется значимость меры статистической связи признака с классом целевой характеристики; формулы вычисления мер связи и порог значимости приведены в главе 3).

Определение 4. Для объекта  $o$  и целевой характеристики  $p_k$  метрика интерпретируемости путей GPI:

$$\text{GPI}_{cov}(o, p_k) = \mathbb{1}[\exists \text{ значимый путь } o \rightarrow f_i \rightarrow p_k],$$

$$\text{GPI}_w(o, p_k) = \frac{\sum_i \alpha_{ik} \cdot s_{ik}}{\sum_i \alpha_{ik}},$$

где  $\alpha_{ik}$  – коэффициент внимания ребра  $f_i \rightarrow p_k$  в модели для объекта  $o$ , а  $s_{ik}$  – индикатор значимости пути. Общая оценка по выборке вычисляется как среднее по объектам и характеристикам.

Метрика GPI вводится как система из основной и вспомогательной компонент. Основная компонента – покрытие путями:

$$\text{GPI}_{cov} = \frac{1}{N \cdot K} \sum_{n=1}^N \sum_{k=1}^K \mathbb{1}[\exists \pi_{ik} - \text{значимый объясняющий путь для } p_k(o_n)],$$

доля решений, для которых найден хотя бы один значимый объясняющий путь.

Вспомогательная компонента – взвешенная согласованность

$$GPI_w = \frac{1}{N \cdot K} \sum_{n=1}^N \sum_{k=1}^K \frac{\sum_i w_{ik}^{(n)} s_{ik}^{(n)}}{\sum_i w_{ik}^{(n)}},$$

где  $w_{ik}^{(n)} \geq 0$  – вес (внимание) ребра  $f_i \rightarrow p_k$  в решении модели для объекта  $o_n$ , а  $s_{ik}^{(n)} \in \{0,1\}$  – индикатор значимости соответствующего пути.

Утверждение 1. Пусть при построении графа знаний проверяется  $m$  гипотез о наличии связи  $f_i \rightarrow p_k$ , и ребро признаётся значимым по критерию Бонферрони при уровне значимости  $\alpha/m$ . Тогда групповая вероятность хотя бы одного ложноположительного ребра не превышает  $\alpha$ :

$$FWER = \Pr\left(\bigcup_{j=1}^m A_j\right) \leq \alpha,$$

где  $A_j$  – событие « $j$ -е ребро ложно признано значимым».

Обоснование – по неравенству Буля (объединительному неравенству)

$$\Pr\left(\bigcup_{j=1}^m A_j\right) \leq \sum_{j=1}^m \Pr(A_j) \leq m \cdot \frac{\alpha}{m} = \alpha.$$

В кросс-доменных проверках, приложения К и Л, применяется менее консервативная процедура Бенджамини–Хохберга, контролирующая ожидаемую долю ложных рёбер FDR на уровне  $q$ .

Утверждение 2. Метрики  $GPI_{cov}$  и  $GPI_w$  принимают значения в отрезке  $[0,1]$ . При этом  $GPI_w = 0$  для модели, не относящей вес решения на значимые пути (модель «чёрный ящик»), и  $GPI_w \rightarrow 1$  при сосредоточении всего веса решения на значимых путях.

В дальнейшем при использовании обозначения  $GPI$  без индекса используется для основной компоненты метрики  $GPI_{cov}$ ; вспомогательная компонента  $GPI_w$  указывается явно.

Обоснование.  $GPI_{cov}$  есть среднее индикаторов из  $\{0,1\}$ , следовательно,  $GPI_{cov} \in [0,1]$ . Каждое слагаемое в  $GPI_w$  есть взвешенное среднее величин  $s_{ik}^{(n)} \in \{0,1\}$  с неотрицательными весами  $w_{ik}^{(n)}$ , поэтому оно лежит в  $[0,1]$ , как и их среднее. Если веса сосредоточены на путях с  $s = 0$  (объяснение не опирается на значимые пути), числитель равен нулю и  $GPI_w = 0$ ; если весь вес сосредоточен на путях с  $s = 1$ , каждое слагаемое равно 1 и  $GPI_w = 1$ .

Утверждение 3. Если предсказание модели не зависит от признака  $i$ , то удаление этого признака ( $c_i = 0$ ) не изменяет предсказания. Как следствие, монотонное падение уверенности модели при последовательном удалении признаков в порядке убывания приписанного им веса является количественной мерой верности объяснений: чем раньше и резче падает уверенность (меньше площадь под кривой удаления), тем точнее веса отражают действительную опору решения на признаки.

Обоснование. Первое утверждение непосредственно следует из независимости функции предсказания от  $c_i$ . Мерой верности служит площадь под кривой относительной уверенности  $\{u_0, u_1, \dots\}$ , где  $u_t$  – уверенность после удаления  $t$  наиболее весомых признаков; для верных объяснений последовательность убывает быстрее.

#### 2.4.1 Функция мультимодальной интеграции

Для объединения признаков различных модальностей вводится функция интеграции  $\Phi$ , отображающая представления отдельных модальностей в единое мультимодальное пространство признаков:

$$x^{(u)} = \Phi(x_{t^{(u)}}, x_{v^{(u)}}, x_{b^{(u)}}),$$

где  $x_{t^{(u)}}$ ,  $x_{v^{(u)}}$  и  $x_{b^{(u)}}$  – представления текстовой, визуальной и табличной модальностей соответственно.

В литературе выделяют три основные стратегии мультимодального слияния: раннее, позднее и гибридное [4, 199]. В настоящем исследовании используется гибридное слияние, при котором представления модальностей предварительно отображаются в общее латентное пространство посредством обучаемых преобразований, после чего объединяются механизмом внимания. В результате каждой модальности сопоставляется коэффициент значимости, отражающий её вклад в формирование итогового представления объекта анализа. Формальное описание указанных преобразований – проекций модальностей, межмодального внимания и адаптивного слияния приведено далее в параграфе «Гибридная нейросимвольная мультимодальная модель» настоящей главы (подпараграф «Адаптивное слияние»). Все параметры этих преобразований (матрицы проекций, векторы внимания, параметры слияния) являются обучаемыми и настраиваются совместно с остальными компонентами модели в процессе минимизации общей функции потерь. Развёрнутые вычислительные формулы программной реализации приведены в приложении Д.

Выбор гибридного слияния обусловлен необходимостью совместного учёта информации различных модальностей и возможностью оценки их вклада в формирование итогового представления объекта анализа.

#### 2.4.2 Формальная модель графа знаний

Мультимодальное представление объекта анализа  $x^{(u)}$  используется для формирования графа знаний:

$$G^{(u)} = (N^{(u)}, \mathcal{E}^{(u)}),$$

где  $N^{(u)}$  – множество узлов графа,  $\mathcal{E}^{(u)}$  – множество ребер,  $u$  – индекс объекта анализа.

Состав узлов  $N(u)$  и рёбер  $\mathcal{E}^{(u)}$  соответствует Определению 2: узел объекта анализа, узлы признаков и узлы целевых характеристик, связанные предикатами «имеет\_признак», «ассоциируется\_с» и «коррелирует\_с». Вес рёбер «ассоциируется\_с» инициализируется статистически значимой связью  $r_{ik}$  и далее уточняется как обучаемый параметр графовой нейронной сети.

Таким образом, граф знаний отражает взаимосвязи между объектом анализа, его признаками и целевыми характеристиками в рамках онтологического паттерна «объект анализа – признак – целевая характеристика».

Каждому узлу  $n \in N^{(u)}$  сопоставляется векторное представление

$$h_n \in \mathbb{R}^{d_{\text{hidden}}},$$

где  $h_n$  – эмбединг узла  $n$ ,  $d_{\text{hidden}}$  – размерность скрытого пространства представлений.

Начальные эмбединги узлов задаются следующим образом:

- для узла объекта анализа: начальное представление задаётся проекцией мультимодального вектора  $x(u)$ , формируемого адаптивным слиянием модальностей, в пространство графа знаний;

- для узлов признаков:  $h_{v_{\text{feature}_i}}^{(0)} = z_i$ , где  $z_i$  – проекция извлечённого признака в пространство графа;

- для узлов целевых характеристик:  $h_{v_{\text{trait}_k}}^{(0)} = e_k$ , где  $e_k$  – обучаемые эмбединги  $k$ -й целевой характеристики.

Обучаемые эмбединги целевых характеристик представляют собой векторы параметров модели: они инициализируются случайными значениями и уточняются в процессе обучения совместно с остальными параметрами сети градиентным методом. После обучения вектор содержит обобщённое представление целевой характеристики в едином пространстве представлений графа и используется механизмом внимания для вычисления взаимодействий с узлами признаков. Данные обозначения согласованы с формальной моделью, вводимой далее: проекции признаков соответствуют результатам преобразований уровней 1–2, а эмбединги целевых характеристик – параметрам графовой составляющей модели.

Сформированный граф знаний используется для представления взаимосвязей между признаками объекта анализа и целевыми характеристиками и служит входом графовой нейронной сети.

## 2.5 Типизированное представление входных данных

Для формального описания структуры мультимодального представления объекта  $M(o) = \langle T, V, D \rangle$ , введённого в главе 1, рассматриваются исходные данные объекта анализа  $o$  как конечный набор элементов данных, каждый из которых характеризуется способом получения и способом последующей обработки. Для формализации этой структуры вводится типизированное представление мультимодальных данных.

**Определение 5.** Типизированным элементом мультимодальных данных объекта  $o$  называется величина

$$e_{t,j}^{m,d}, \quad j = 1, \dots, J_t^{m,d},$$

где  $m \in \{1, \dots, M\}$  – модальность (изображение, текст, число и др.);  $d \in \{1, \dots, D_m\}$  – домен, характеризующий источник получения и формат данных конкретной модальности;  $t \in \{1, \dots, T^{m,d}\}$  – тип, представляющий собой задаваемую экспертом категорию элементов одного домена, определяющую единообразный способ их обработки;  $j$  – номер элемента данного типа,  $J_t^{m,d}$  – количество элементов данного типа.

Все элементы вида  $e_{t,*}^{m,d}$  обрабатываются единообразно. Модальности при этом соответствуют компонентам описания из главы 1: текстовая модальность  $T$  объединяет элементы  $\{e_{t,j}^{m_T,d}\}$ , визуальная  $V$  – элементы  $\{e_{t,j}^{m_V,d}\}$ , табличная  $D$  – элементы  $\{e_{t,j}^{m_D,d}\}$ .

Обозначим через  $M^T$ ,  $M^V$ ,  $M^D$  подмножества индексов модальностей  $m$ , относящихся соответственно к текстовой, визуальной и табличной компонентам мультимодального описания. Тогда текстовая, визуальная и табличная компоненты описания объекта представляются множествами типизированных элементов данных

$$T = \{e_{t,j}^{m_T,d} : m \in M^T\}, V = \{e_{t,j}^{m_V,d} : m \in M^V\} \text{ и } D = \{e_{t,j}^{m_D,d} : m \in M^D\}.$$

Все элементы вида  $e_{t,*}^{m,d}$  обрабатываются единообразно одним и тем же набором преобразований. Тем самым способ представления данных отделяется от способа их обработки, который, в свою очередь, определяется типом.

Разбиение на домены и типы, а также состав применяемых к ним преобразований определяются разработчиком конкретной прикладной системы на основе предметной онтологии  $O = (C, R)$ . Предметная онтология задаёт соответствие между типами

данных, извлекаемыми признаками и категориями предметной области, обеспечивая семантическую согласованность процесса обработки.

**Определение 6.** Первичным описанием элемента данных называется тройка

$$\psi = (v, k, \gamma) \in (\mathbb{R}^n \cup \{\emptyset\}) \times (\mathcal{K} \cup \{\emptyset\}) \times (R[0,1] \cup \{\emptyset\}),$$

где  $v$  – вектор-эмбединг;  $k$  – категория из конечного множества  $K$ ;  $\gamma$  – степень уверенности отнесения элемента к категории  $k$ . Отдельные компоненты тройки могут быть пустыми ( $\emptyset$ ), однако не все одновременно.

Предложенное первичное описание охватывает основные варианты обработки мультимодальных данных, возникающие на практике. Так, нейросетевой энкодер текста формирует только эмбединг  $(v, \emptyset, \emptyset)$ ; классификатор изображений – категорию с уверенностью  $(\emptyset, k, \gamma)$  либо полное первичное описание  $(v, k, \gamma)$ ; категориальный табличный атрибут преобразуется в категорию с полной уверенностью  $(\emptyset, k, 1)$ .

## 2.6 Гибридная нейросимвольная мультимодальная модель

Общая схема модели приведена на рисунке 4.

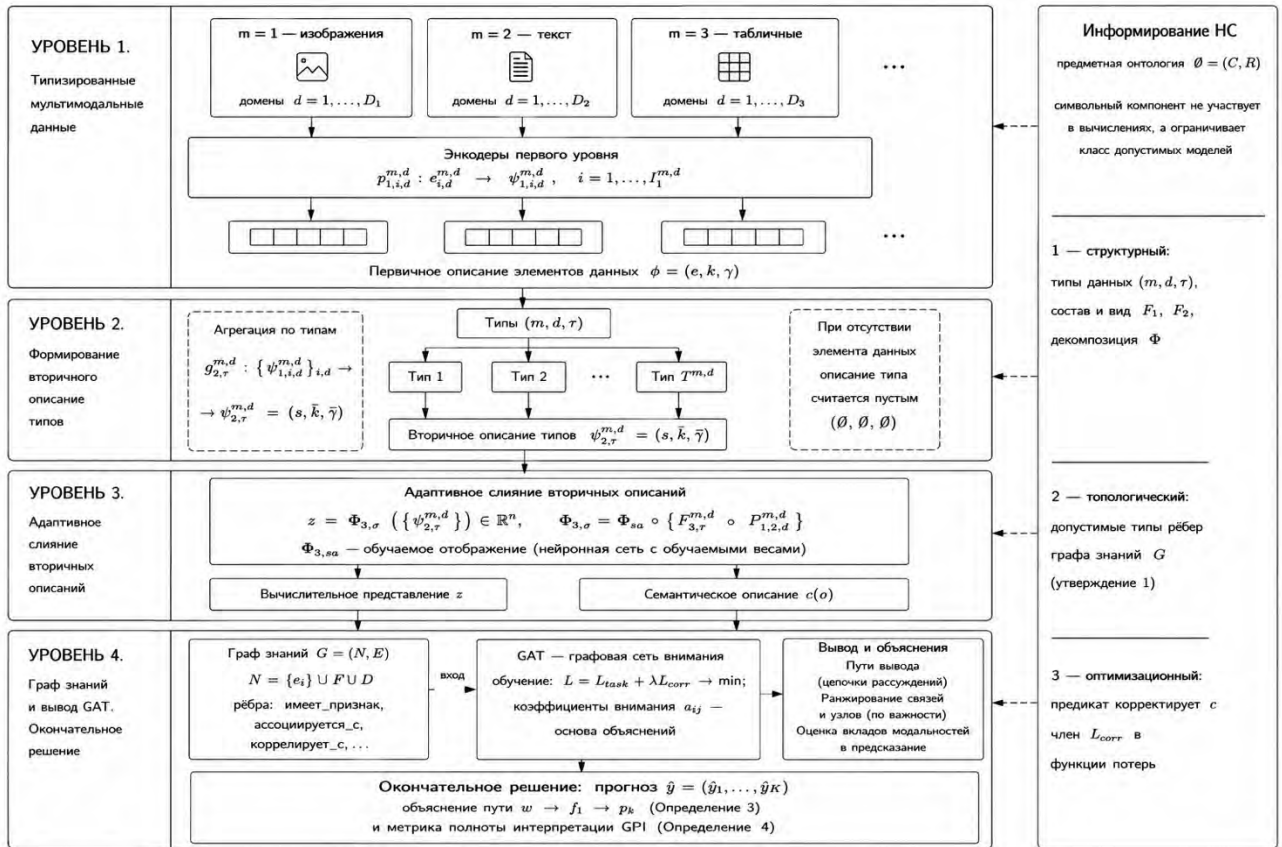


Рисунок 4 – Общая схема гибридной нейросимвольной мультимодальной модели

Модель представляется в виде последовательности основных этапов обработки: формирование первичного описания элементов данных (уровень 1), формирование вторичного описания типов (уровень 2), адаптивное слияние вторичных описаний,

построение графа знаний и вывод графовой сети внимания, завершающийся формированием окончательного решения, а также информирование нейронной сети предметной онтологией, определяющее структуру каждого из этапов.

### 2.6.1 Уровень 1 – Формирование первичного описания

Для каждого типа  $(m, d, t)$  задаётся набор энкодеров первого уровня

$$F_{1,t,i}^{m,d}: e_{t,j}^{m,d} \mapsto \psi_{t,i,j}^{m,d}, i = 1, \dots, I_t^{m,d},$$

каждый из которых отображает элемент данных в его первичное описание (Определение 6). Энкодеры первого уровня выполняют извлечение признаков либо классификацию отдельных элементов данных соответствующего типа. В частном случае энкодер может быть тождественным отображением, например, для готовых числовых атрибутов.

### 2.6.2 Уровень 2 – Формирование вторичного описания

Для каждого типа задаётся преобразование

$$F_{2,t}^{m,d}: \{\psi_{t,i,j}^{m,d}\}_{i,j} \mapsto \Psi_t^{m,d} = (\bar{v}, \bar{k}, \bar{\gamma}),$$

агрегирующее первичные описания элементов данного типа в единое вторичное описание. Преобразование  $F_2$  также может быть тождественным при  $I_t^{m,d} = J_t^{m,d} = 1$ ; основными типами преобразований являются усреднение эмбедингов, голосование по категориям с пересчётом степени уверенности либо иной обучаемый механизм агрегации.

Полученное вторичное описание типа содержит одновременно вычислительное представление объекта (эмбединг), категориальное описание и оценку достоверности. В дальнейшем вторичные описания используются модулем адаптивного слияния модальностей и служат основой для построения графа знаний и выполнения нейросимвольного вывода.

### 2.6.3 Адаптивное слияние вторичных описаний

Вторичные описания типов объединяются модулем адаптивного слияния

$$x = \Phi_{fus}(\{\Psi_t^{m,d}\}) \in \mathbb{R}^n,$$

где  $\Phi_{fus}$  – обучаемое отображение, выполняющее адаптивную интеграцию описаний различных типов данных в единое векторное представление объекта. Отображение реализуется нейронной сетью адаптивного слияния с обучаемыми весами.

Таким образом, введённое в главе 1 отображение

$$x = \Phi(M(o))$$

конкретизируется в виде композиции



$$\Phi = \Phi_{fus} \circ \{F_{2,t}^{m,d} \circ F_{1,t,i}^{m,d}\}.$$

#### 2.6.4 Граф знаний и вывод графовой сети внимания (GAT)

Категориальные компоненты  $(\bar{k}, \bar{\gamma})$  вторичных описаний  $\Psi_t^{m,d}$  образуют интерпретируемое признаковое описание объекта  $c(o)$  (Определение 1), на основе которого формируются узлы признаков  $F$  и рёбра имеет\_признак графа знаний  $G$  (Определение 2).

Тем самым единый двухуровневый конвейер обработки одновременно формирует:

- вычислительное (векторное) представление объекта  $x$ ;
- семантическое (категориальное) представление объекта  $c(o)$ .

Полученные описания используются графовой составляющей гибридной модели: на их основе строится граф знаний  $G = (N, E)$  со множеством узлов.  $N = \{v_o\} \cup F \cup P$  (узел объекта, узлы признаков и узлы целевых характеристик) и рёбрами «имеет\_признак», «ассоциируется\_с» и «коррелирует\_с» (Определение 2). Вывод на графе выполняет графовая сеть внимания GAT, обучаемая минимизацией функционала  $L = L_{MSE} + \lambda L_{corr}$ ; её коэффициенты внимания служат одновременно основой прогноза и основой построения объясняющих путей. Результатом является окончательное решение – прогноз характеристик объекта вместе с его интерпретацией (Определение 3) и оценкой полноты интерпретации метрикой GPI (Определение 4).

Если элемент данных недоступен, соответствующее описание типа считается пустым:

$$\Psi_t^{m,d} = (\emptyset, \emptyset, \emptyset).$$

В этом случае узлы признаков и инцидентные им рёбра для данного типа не создаются, а графовая нейронная сеть выполняет вывод на подграфе, сформированном только доступными данными.

Таким образом, прогноз остаётся определённым, а его интерпретация (Определение 3) строится исключительно по фактически наблюдаемым признакам. Снижение полноты объяснения вследствие отсутствия данных количественно оценивается метрикой GPI (Определение 4).

#### 2.6.5 Информирование нейронной сети предметной онтологией

Сформулируем механизм нейросимвольной интеграции, реализуемый в предлагаемой модели.

Определение 7. Под информированием нейронной сети понимается способ нейросимвольной интеграции, при котором символьный компонент – предметная онтология  $O = (C, R)$  – не участвует непосредственно в вычислении выходов нейронной сети, а определяет структуру пространства допустимых моделей и процесса их обучения посредством трёх механизмов информирования.

- Структурный механизм информирования – онтология определяет типизацию данных  $(m, d, t)$ , состав и вид преобразований  $F_1$  и  $F_2$ , то есть декомпозицию отображения  $\Phi$ .

- Топологический механизм информирования – онтология определяет множество допустимых типов рёбер графа знаний  $G$ , из которых в граф включаются только статистически значимые связи, соответствующие семантическим отношениям предметной области (Определение 2, Утверждение 1).

- Оптимизационный механизм информирования – онтологический предикат коррелирует\_с реализуется корреляционным членом  $\lambda L_{corr}$  функции потерь, обеспечивающим согласованность прогнозируемых характеристик.

Информирование нейронной сети означает, что минимизация функционала

$$L = L_{MSE} + \lambda L_{corr}$$

выполняется не по всему классу возможных моделей, а по его подклассу, структура которого согласована с предметной онтологией.

Ограничение класса моделей может приводить к некоторому изменению точности по сравнению с неограниченной моделью. Количественная оценка этой платы за встроенную интерпретируемость приводится в главе 5 в эксперименте с исключением онтологических ограничений ( $\lambda = 0$ ).

Получаемым преимуществом являются возможность построения семантически интерпретируемых объясняющих путей (Определение 3) и количественный контроль качества интерпретации с использованием метрики GPI (Определение 4).

Тем самым преобразования  $F_1$ ,  $F_2$  и отображение  $\Phi$  задают структуру модели, согласованную с типами данных и категориями предметной области, а сам граф предметной онтологии используется для инициализации графа знаний, на котором выполняется вывод GAT. Информированием нейронной сети описание модели завершается: онтология определяет структуру всех этапов – от формирования первичного описания до получения окончательного решения.

## 2.7 Онтология предметной области

Онтология предметной области задаётся экспертами предметной области на этапе проектирования прикладной системы и фиксируется в машиночитаемом виде до начала обучения модели. Эксперты определяют состав концептов (целевых характеристик и категорий признаков), допустимые типы отношений (предикаты) с их доменами и ко-доменами, а также ограничения целостности; эти элементы в процессе обучения не изменяются – данные уточняют лишь веса допустимых рёбер. В настоящей работе онтология реализована в виде графа свойств под управлением СУБД Neo4j (приложение М); процедура её построения для конкретной предметной области, включая источники экспертных категорий, описана в главе 3. Ниже рассматриваются теоретические основы и структура предлагаемого онтологического паттерна.

### 2.7.1 Теоретические основы онтологического моделирования

Онтология в контексте информационных систем представляет собой формализованную спецификацию концептуализации предметной области – явное определение типов сущностей, их атрибутов и отношений между ними [200, 201]. В настоящем исследовании онтология выполняет роль концептуального каркаса, структурирующего знания о взаимосвязях между объектом анализа, его признаками и целевыми характеристиками. Предложенная онтология первоначально была разработана автором для задачи психологического профилирования пользователей социальной сети [211\*] и впоследствии обобщена до доменно-инвариантного онтологического паттерна проектирования «объект анализа – признак – целевая характеристика» [204\*], допускающего перенос в различные предметные области. Возможность переноса подтверждена в приложениях К и Л на медицинских и текстовых данных.

Граф знаний реализует онтологию в виде направленного графа, в котором узлы соответствуют сущностям предметной области, а рёбра – семантическим отношениям между ними [26, 22]. В отличие от обычного графа данных структура графа знаний определяется онтологией и отражает смысловые взаимосвязи между объектами предметной области.

Базовой единицей представления знаний является триплет (triple), записываемый в формате RDF (Resource Description Framework) [202]:

«Триплет» = («Субъект», «Предикат», «Объект»),

где субъект представляет сущность, о которой формулируется утверждение; предикат определяет тип отношения; объект представляет сущность или значение, связанное с субъектом.

Совокупность триплетов формирует граф знаний, в котором субъекты и объекты соответствуют узлам графа, а предикаты – рёбрам. Семантика отношений задаётся онтологией и определяет допустимые типы связей между концептами предметной области.

#### 2.7.2 Концепты онтологического паттерна

Предлагаемый онтологический паттерн включает три основных типа концептов.

1. Концепт «объект анализа» представляет сущность, для которой выполняется прогнозирование целевых характеристик. Объект анализа описывается совокупностью наблюдаемых данных и извлекаемых признаков. Конкретная природа объекта определяется предметной областью и рассматривается в последующих главах.

2. Концепт «признак» представляет характеристику объекта анализа, полученную непосредственно из данных либо извлечённую методами интеллектуального анализа. Признаки формируются на основе различных модальностей данных и используются в качестве промежуточного уровня представления между объектом анализа и целевыми характеристиками.

3. Концепт «целевая характеристика» представляет прогнозируемое свойство объекта анализа. Целевые характеристики являются результатом работы интеллектуальной системы и используются для поддержки принятия решений в конкретной предметной области.

Таким образом, предложенный онтологический паттерн формализует связь между объектом анализа, наблюдаемыми признаками и целевыми характеристиками предметной области: объект анализа – признак – целевая характеристика.

#### 2.7.3 Типы отношений в онтологии

Связи между концептами формализуются в виде бинарных предикатов.

##### 1. имеет\_признак (has\_feature)

Домен: объект анализа

Ко-домен: признак

Семантика: объект анализа обладает определённым признаком.

##### 2. ассоциируется\_с (associated\_with)

Домен: признак

Ко-домен: целевая характеристика

Семантика: наличие признака связано с формированием соответствующей целевой характеристики.

3. коррелирует\_с (correlates\_with)

Домен: целевая характеристика

Ко-домен: целевая характеристика

Семантика: отражает статистическую или семантическую взаимосвязь между целевыми характеристиками.

Указанные типы отношений образуют базовую структуру графа знаний и используются при формировании нейросимвольной модели анализа. В главе 3 представлена конкретная предметная реализация предложенного онтологического паттерна, включающая состав признаков, целевых характеристик и систему триплетов графа знаний.

#### 2.7.4 Формирование триплетов и семантических связей

Формирование графа знаний в предложенной нейросимвольной модели основано на преобразовании признаков объекта анализа в совокупность триплетов, соответствующих онтологическому паттерну «объект анализа – признак – целевая характеристика». Предложенная нейросимвольная модель, объединяющая графовую сеть внимания и граф знаний для интерпретируемого предсказания, представлена автором в [55\*].

Каждый триплет задаёт семантическое отношение между сущностями предметной области и используется для построения графа знаний. Совокупность триплетов формирует структуру графа, отражающую взаимосвязи между объектом анализа, его признаками и целевыми характеристиками.

В отличие от классических экспертных систем, в которых правила вывода задаются явно, в предложенной архитектуре зависимости между признаками и целевыми характеристиками выявляются автоматически в процессе обучения графовой нейронной сети. При этом граф знаний определяет структуру возможных взаимосвязей, а веса рёбер отражают степень их значимости.

Концептуально вывод целевой характеристики соответствует выявлению в графе знаний путей, связывающих объект анализа с соответствующими целевыми характеристиками через систему признаков. Достоверность такого вывода определяется совокупностью весов рёбер и структурой графа, формируемой в процессе обучения.

Развёрнутый пример формирования графа знаний, набора триплетов и процесса вывода для конкретной предметной области приведён в главе 3.

#### 2.7.5 Формализация онтологии в OWL и RDF

Для формальной спецификации онтологии используются стандарты Semantic Web – OWL (Web Ontology Language) [203] и RDF (Resource Description Framework), обеспечивающие формальное описание концептов, отношений и ограничений предметной области.

В рамках предложенного онтологического паттерна определяются следующие базовые классы:

- ObjectAnalysis – объект анализа;
- Feature – признак;
- TargetCharacteristic – целевая характеристика.

Классы концептов, объектные свойства, ограничения на допустимые связи между сущностями и экземпляры онтологии формализуются средствами OWL и RDF. Формальные спецификации классов, свойств и примеры сериализации онтологии приведены в приложении.

Формализация онтологии средствами OWL и RDF обеспечивает однозначную интерпретацию концептов и отношений предметной области, совместимость с технологиями Semantic Web и возможность использования онтологических знаний в процедурах нейросимвольного вывода.

#### 2.7.6 Визуализация онтологического паттерна

Для наглядного представления структуры разработанного онтологического паттерна и семантических связей между его концептами выполнена графическая визуализация графа знаний.

Визуализация отражает взаимосвязи между объектом анализа, признаками и целевыми характеристиками и демонстрирует структуру организации знаний в графе. На рисунке 5 представлена концептуальная схема онтологического паттерна «объект анализа – признак – целевая характеристика».

Схема отражает:

- представление объекта анализа как центральной сущности графа знаний;
- связь объекта анализа с наблюдаемыми признаками;
- ассоциации признаков с целевыми характеристиками;
- корреляционные связи между целевыми характеристиками.

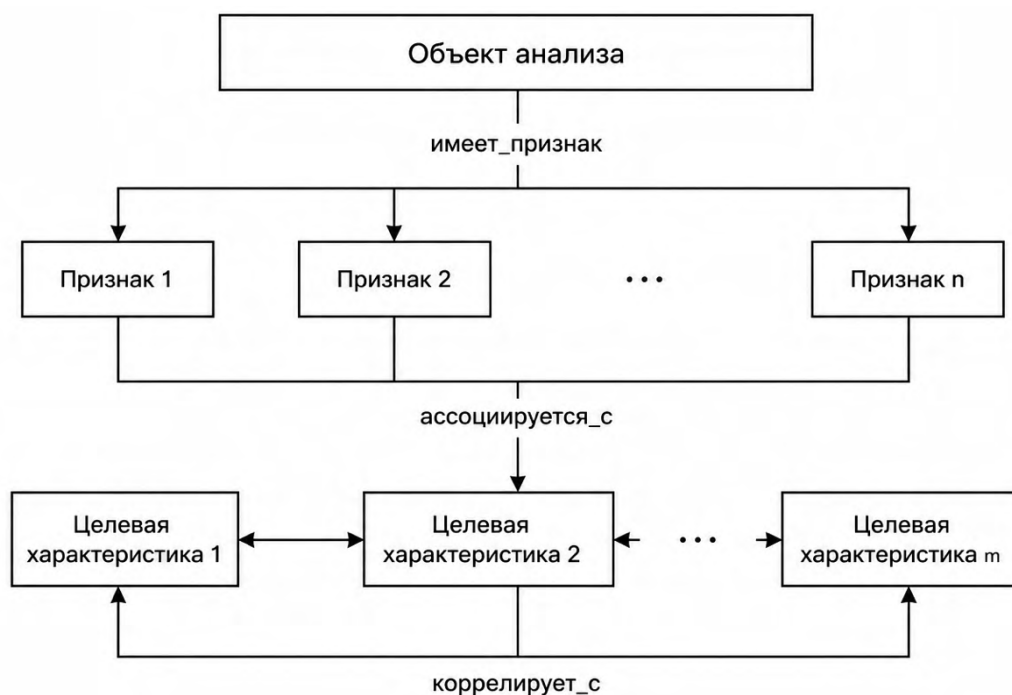


Рисунок 5 – Концептуальная схема онтологического паттерна

Числовые значения на рёбрах отражают веса связей графа знаний. Для отношений типа «ассоциируется\_с» веса формируются в процессе обучения графовой нейронной сети, а для отношений типа «коррелирует\_с» определяются на основе предметных исследований либо статистического анализа данных.

Представленная структура отражает совместное использование знаний, извлекаемых из данных, и знаний предметной области, формализованных в виде графа знаний. Представленная структура используется графовой нейронной сетью для формирования предсказаний и построения интерпретируемых объяснений.

В главе 3 приведена предметная реализация предложенного онтологического паттерна, включающая конкретный состав признаков, целевых характеристик, примеры триплетов графа знаний и количественные характеристики сформированной графовой структуры.

## 2.8 Граф знаний

Граф знаний является главным компонентом предложенного метода, обеспечивающим интеграцию мультимодальных данных, онтологического представления знаний и нейросетевых методов анализа. Концептуально граф знаний представляет собой направленный граф с типизированными узлами и рёбрами, формируемый на основе онтологии предметной области. Узлы графа соответствуют концептам онтологического паттерна («объект анализа», «признак», «целевая характеристика»), а рёбра отражают семантические отношения между ними. Веса

рёбер типа «ассоциируется\_с» являются обучаемыми параметрами и уточняются в процессе обучения графовой нейронной сети, тогда как веса рёбер типа «коррелирует\_с» задаются на основе предметных исследований либо статистического анализа данных. Граф знаний выступает интегрирующей структурой, связывающей признаки, извлекаемые из мультимодальных данных, с целевыми характеристиками посредством явно заданных семантических отношений и обеспечивающей совместное использование нейросетевых и символьных методов анализа.

Процедура формирования графа знаний на основе онтологии и мультимодальных признаков формализована в виде алгоритма 1.

Обозначения:  $G = (N, E)$  – граф знаний;  $N = \{v_o\} \cup F \cup P$  – узлы объекта ( $v_o$ ), признаков  $F = \{f_i\}$  и целевых характеристик  $P = \{p_k\}$ ;  $c(o) = (c_1, \dots, c_d)$  – признаковое описание объекта;  $r_{ik}$  – вес ребра «ассоциируется\_с» ( $f_i \rightarrow p_k$ );  $\rho_{kl}$  – вес ребра «коррелирует\_с»;  $\alpha_{ij}$  – коэффициенты внимания графовой сети внимания (GAT).

В записи алгоритмов символ  $:=$  обозначает операцию присваивания: запись  $a := b$  означает, что переменной или формируемому множеству  $a$  присваивается значение выражения  $b$ . Операции над множествами обозначаются стандартно:  $\in$  – принадлежность,  $\cup$  – объединение,  $\cap$  – пересечение.

Алгоритм 1 – Формирование графа знаний на основе онтологии и мультимодальных признаков

Вход: онтология  $O$  (классы, предикаты, аксиомы); признаки объектов  $\{c(o)\}$ ; множество статистически значимых ассоциаций  $A$ ; матрица корреляций целевых характеристик  $\{\rho_{kl}\}$ .  
(Признаки  $\{c(o)\}$  представляют собой вторичные описания объектов, сформированные преобразованиями уровней 1 и 2 типизированной модели (Определения 5–6).)  
Выход: граф знаний  $G = (N, E)$ .

1.  $N := \emptyset$ ;  $E := \emptyset$ .
2. Для каждого класса целевой характеристики  $p_k$ , определённого онтологией  $O$ :
  - создать узел  $p_k$  с обучаемым эмбедингом;
  - $N := N \cup \{p_k\}$ .
3. Для каждого объекта анализа  $o$ :
  - создать узел объекта  $v_o$ ;
  - $N := N \cup \{v_o\}$ .
4. Для каждого признака  $f_i$  со значением  $c_i$ , допустимого онтологией для данного объекта:
  - создать узел  $f_i$ ;
  - $N := N \cup \{f_i\}$ ;
  - создать ребро  $(v_o, f_i, \text{имеет\_признак}, c_i)$ .
5. Для каждой статистически значимой ассоциации  $(f_i, p_k) \in A$ , допускаемой предикатом онтологии ассоциируется\_с:
  - создать ребро  $(f_i, p_k, \text{ассоциируется\_с}, r_{ik})$ , где  $r_{ik}$  инициализируется значимостью ассоциации.



6. Для каждой пары целевых характеристик  $(p_k, p_l)$ , для которой онтологией допускается предикат коррелирует\_с:
  - создать ребро  $(p_k, p_l, \text{коррелирует\_с}, \rho_{kl})$ .
7. Вернуть граф знаний  $G = (N, E)$ .

Множество допустимых пар  $A$  и начальные значения весов  $r_{ik}$  и  $\rho_{kl}$  формируются на этапе статистической инициализации: меры связи вычисляются по обучающей выборке, после чего проверяется их статистическая значимость с поправкой на множественность проверок; рёбра, не прошедшие проверку либо не допускаемые онтологией, в граф не включаются. Конкретные статистические меры (точечно-бисериальный коэффициент корреляции для рёбер «ассоциируется\_с», коэффициент корреляции Пирсона для рёбер «коррелирует\_с») и формулы их вычисления приведены в главе 3.

Утверждение 4. Алгоритм 1 завершается за конечное число шагов и возвращает граф  $G$ , структура которого допустима онтологией  $O$ : каждое ребро  $G$  соответствует одному из предикатов онтологии, а его концы принадлежат домену и ко-домену этого предиката. Завершаемость следует из конечности множеств классов, объектов, признаков и пар-кандидатов; допустимость обеспечивается формированием рёбер только между допустимыми типами узлов и исключением рёбер, нарушающих ограничения предметной онтологии. Тем самым структурное ограничение гибридной модели выполняется по построению, а не проверяется апостериорно. Вычислительная сложность построения составляет  $O(|F| \cdot |P| + |P|^2)$ , где  $|F|$  – число признаков,  $|P|$  – число целевых характеристик; требуемая память –  $O(|N| + |E|)$ .

Выбор технологической платформы хранения графа, сравнительный анализ графовых СУБД, структура узлов и рёбер конкретной реализации, процедуры наполнения графа, а также количественные характеристики сформированного графа знаний рассмотрены в главе 4.

## 2.9 Нейросимвольная интеграция и механизм взаимодействия компонентов

Нейросимвольная интеграция является основным элементом предложенного метода, объединяющим нейросетевые методы обработки мультимодальных данных и символьное представление знаний в форме графа знаний. В параграфе описан механизм взаимодействия нейросетевого и символьного компонентов, связывающий извлечение признаков из мультимодальных данных с построением графа знаний и формированием интерпретируемых предсказаний.

### 2.9.1 Двухнаправленное взаимодействие данных и знаний

Нейросимвольная интеграция реализуется посредством двухнаправленного взаимодействия между данными и знаниями. Первое направление взаимодействия связано с извлечением признаков из мультимодальных данных с использованием нейронных моделей и последующим включением полученных представлений в структуру графа знаний. Второе направление взаимодействия связано с использованием структуры графа знаний для организации семантических связей между признаками и целевыми характеристиками, формирования предсказаний и их интерпретации. Схематическое представление процесса приведено на рисунке 6.

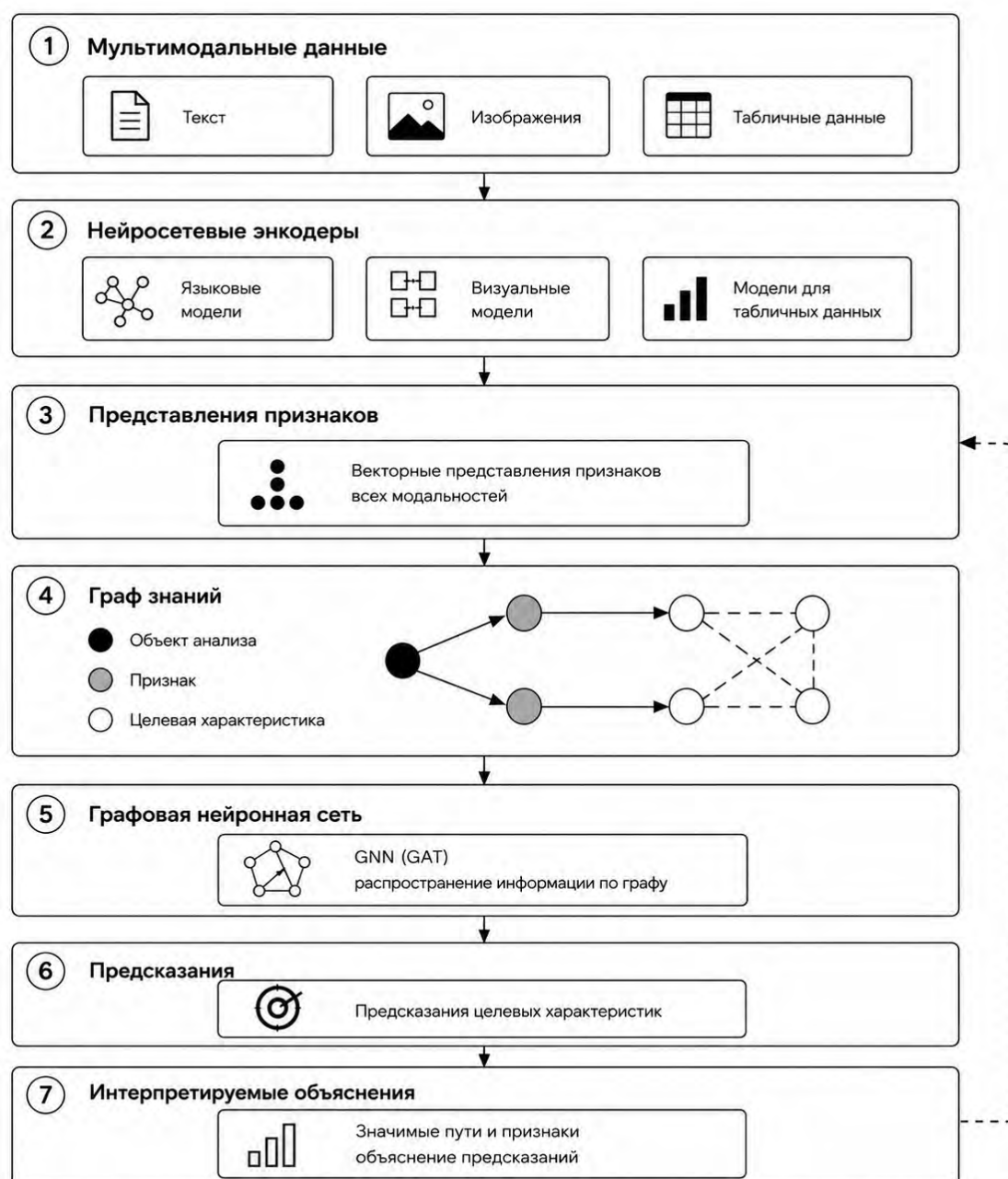


Рисунок 6 – Двухнаправленное взаимодействие данных и знаний в нейросимвольной архитектуре

После извлечения признаки отдельных модальностей объединяются в единое представление объекта анализа с использованием механизма адаптивного слияния модальностей. Сплошными стрелками показан основной поток обработки данных – от мультимодальных источников через нейросетевые энкодеры и граф знаний к формированию предсказаний и интерпретируемых объяснений. Пунктирной линией обозначена обратная связь, отражающая использование структуры графа знаний и результатов интерпретации для уточнения семантических связей между признаками и целевыми характеристиками.

Таким образом, схема иллюстрирует двунаправленный характер нейросимвольной интеграции, при котором данные используются для формирования графовой структуры, а знания, представленные в графе, участвуют в процессе анализа и интерпретации результатов.

Следует отметить, что энкодеры признаков используются в режиме фиксированных параметров; обратный поток «от данных к знаниям» реализуется как уточнение весов допустимых рёбер графа знаний в процессе обучения, то есть адаптации подлежит структура предметных знаний, а не параметры извлечения признаков.

### 2.9.2 Символьный компонент

Символьный компонент представлен графом знаний, который используется для структурирования связей между объектом анализа, признаками и целевыми характеристиками. В отличие от традиционных подходов, где предметные зависимости неявно кодируются в параметрах нейронной сети, в предложенной архитектуре они представлены в явном виде посредством узлов и семантических связей графа.

Граф знаний выполняет три основные функции:

- 1) структурирование предметных знаний в виде семантических связей между сущностями;
- 2) использование предметных зависимостей при формировании предсказаний;
- 3) обеспечение интерпретируемости результатов анализа.

Связи типа «ассоциируется\_с» отражают статистические зависимости между признаками и целевыми характеристиками и могут уточняться в процессе обучения модели. Связи типа «коррелирует\_с» отражают известные взаимосвязи между целевыми характеристиками, полученные из предметных исследований либо статистического анализа данных. Благодаря этому граф знаний содержит как информацию, извлекаемую из данных, так и знания предметной области.

После формирования графа знаний становится возможным использование графовых нейронных сетей для распространения информации между связанными сущностями. Графовая структура учитывает систему взаимосвязей между признаками и целевыми характеристиками, наряду с индивидуальными характеристиками объекта анализа. Проекция разнородных признаков в единое пространство представлений сопровождается снижением их индивидуальной информативности. Это компенсируется тем, что, наряду с векторным представлением, каждый узел признака сохраняет символьное описание – категорию и уверенность её определения (Определение 5), а вклад каждого признака в предсказание дополнительно взвешивается механизмом внимания.

Для формирования предсказаний в работе используется графовая сеть внимания Graph Attention Network (GAT) [14]. Механизм внимания вычисляет веса узлов и связей графа, отражающие их вклад в формирование представлений. В результате каждый узел обновляет своё представление с учётом информации, полученной от связанных сущностей, а итоговые представления используются для вычисления значений целевых характеристик.

Общая последовательность формирования предсказаний и интерпретируемых объяснений на основе графа знаний представлена в алгоритме 2.

#### Алгоритм 2 – Нейросимвольный вывод и формирование объяснений

Вход: граф знаний  $G$ ; обученная GAT с  $L$  слоями.

Выход: вектор предсказаний  $Y$ ; множество объясняющих путей  $I$ .

1.  $H_0$  := начальные представления узлов  $G$ .
2. Для  $l = 1, \dots, L$ :  
 $(H_l, \alpha_l) := \text{GAT}_l(H_{l-1}, G)$ .
3. Для каждой  $p_k \in P$ :  
 $y_k := \text{readout}(h_{\{p_k, L\}})$ ;  
 $Y := (y_1, \dots, y_K)$ .
4.  $\alpha := \alpha_L$ .
5. Для каждой предсказанной  $p_k$ :  
 $\Pi_k := \text{пути } (v_o \rightarrow f_i \rightarrow p_k) \text{ в } G$ .
6. Ранжировать  $\pi \in \Pi_k$  по вкладу  $w(\pi) = \prod_{e \in \pi} \alpha_e$ ; для пути  $v_o \rightarrow f_i \rightarrow p_k$  это произведение  $w(\pi) = \alpha_{oi} \cdot \alpha_{ik}$  коэффициентов внимания его рёбер.
7.  $I := \bigcup_k \{\text{наиболее значимые пути из } \Pi_k\}$ .
8. Вернуть  $(Y, I)$ .

Вклад  $w(\pi)$  согласован с Определениями 3–4: коэффициенты внимания  $\alpha_{ik}$  рёбер «ассоциируется с» служат весами  $w_{ik}$  в компоненте  $\text{GPI}_w$ , поэтому ранжирование путей воспроизводит тот же порядок, по которому оценивается взвешенная согласованность объяснений. Сложность вывода определяется шагом 2 и составляет  $O(L \cdot (|E| \cdot d + |N| \cdot d^2))$ , где  $L$  – число слоёв,  $d$  – размерность представлений;

перечисление и ранжирование путей на шагах 5–7 для паттерна «объект – признак – целевая характеристика» требует  $O(|F| \cdot K)$  операций и не увеличивает асимптотику.

Принципиальной особенностью применения GAT является то, что коэффициенты внимания используются не только при формировании предсказаний, но и при построении интерпретируемых объяснений. Коэффициенты внимания характеризуют вклад узлов и связей графа знаний в формирование результата анализа.

После получения предсказаний выполняется поиск наиболее значимых путей в графе знаний, связывающих наблюдаемые признаки с целевыми характеристиками. Выявленные пути используются для построения объяснений в терминах предметной области. Качество формируемых объяснений оценивается с использованием структурной метрики интерпретируемости путей в графе знаний GPI (Graph Path Interpretability). Процедура извлечения объясняющих путей и вычисления метрики рассмотрена далее в параграфе, посвящённом интерпретации результатов.

Таким образом, граф знаний выполняет роль связующего звена между нейросетевым и символьным уровнями модели: нейронные компоненты формируют признаки на основе мультимодальных данных, граф знаний организует их в единую семантическую структуру, а графовая нейронная сеть использует эту структуру для формирования предсказаний и их интерпретации.

### 2.9.3 Организация обучения нейросимвольной модели

Обучение предложенной модели организовано как процесс совместного уточнения параметров нейронных компонентов и графовой структуры знаний. На первом этапе нейросетевые энкодеры извлекают признаки из мультимодальных данных и формируют представления объектов анализа. На втором этапе графовая нейронная сеть использует структуру графа знаний для формирования предсказаний целевых характеристик и распространения информации между связанными сущностями. В процессе обучения уточняются параметры графовой нейронной сети и веса связей типа «ассоциируется\_с», отражающих зависимости между признаками и целевыми характеристиками.

Процедура совместного нейросимвольного вывода и уточнения параметров графа знаний формализована в виде алгоритма 3.

Алгоритм 3 – Совместный нейросимвольный вывод и уточнение параметров графа знаний

Вход: граф  $G$ ; нейросетевые энкодеры; обучающая выборка  $D = \{(o, y^*)\}$ ; критерий остановки; шаг обучения  $\eta$ .

Выход: обученная GAT; уточнённый граф  $G^*$ .

1. Повторять.
2.  $c(o) := \text{Encode}(o)$  для  $o \in D$ .
3.  $G :=$  Алгоритм 1 на признаках  $\{c(o)\}$ .
4.  $Y := \text{GAT}(G)$ ;  
 $L := \text{Loss}(Y, y^*)$ .
5.  $(\theta_{\text{GAT}}, \{r_{ik}\}) := (\theta_{\text{GAT}}, \{r_{ik}\}) - \eta \cdot \nabla L$ .
6. Оценить качество на валидационной выборке.
7. Пока не выполнен критерий остановки.
8.  $G^* := G$  с уточнёнными весами  $\{r_{ik}\}$ .
9. Вернуть  $(\text{GAT}, G^*)$ .

Поскольку параметры энкодеров в предложенной модели заморожены, признаковые описания  $c(o)$  и состав узлов и рёбер графа не изменяются между итерациями: шаги 2–3 выполняются однократно до начала цикла, а внутри цикла обновляются только параметры  $\theta_{\text{GAT}}$  и веса  $\{r_{ik}\}$  рёбер, допустимых онтологией, состав рёбер в процессе обучения не расширяется. Сложность одной итерации составляет  $O(|D| \cdot L \cdot (|E| \cdot d + |N| \cdot d^2))$ , где  $|D|$  – объём обучающей выборки. Критерием остановки служит отсутствие улучшения ошибки на валидационной выборке в течение заданного числа эпох (ранняя остановка), ограничивающая число итераций и предотвращающая переобучение весов рёбер.

В ходе итерационного обучения статистические зависимости, выявляемые на данных, отражаются в параметрах графовой модели и согласуются со структурой предметных знаний, представленной в графе знаний. В результате граф знаний используется не только как средство представления знаний, но и как активный компонент процесса формирования предсказаний.

Под согласованностью здесь понимается следующее. Во-первых, в процессе обучения уточняются только веса рёбер, допустимых онтологией: состав узлов и рёбер графа не расширяется, поэтому выявляемые на данных зависимости не могут выйти за пределы заданной онтологической структуры. Во-вторых, корреляционный член функции потерь штрафует отклонение корреляций предсказанных целевых характеристик от значений, зафиксированных рёбрами «коррелирует\_с». Таким образом, данные уточняют параметры заранее заданной семантической структуры, но не изменяют её состав.

После обучения модели и получения предсказаний выполняется построение интерпретируемых объяснений на основе структуры графа знаний. Процедура извлечения объясняющих путей и оценки их качества формализована в виде алгоритма 4.

#### Алгоритм 4 – Извлечение объясняющих путей и вычисление метрики GPI

Вход: граф  $G^*$  с коэффициентами внимания  $\alpha$ ;  
предсказанная характеристика  $y$  объекта  $v$ ;  
порог значимости  $\tau_s$ ;  
ограничение длины пути  $L_{max}$ .

Выход: ранжированный набор путей  $\Pi^*$ ; значение GPI.

1.  $\Pi :=$  пути из  $v$  к  $y$  длины  $\leq L_{max}$  по допустимым типам связей.
2. Для каждого  $\pi \in \Pi$ :  
 $s(\pi) := \prod_{e \in \pi} \alpha_e$  – произведение коэффициентов внимания рёбер пути (для базового паттерна  $s(\pi_{ik}) = \alpha_{oi} \cdot \alpha_{ik}$ ).
3.  $\Pi' := \{\pi \in \Pi : s(\pi) \geq \tau_s\}$ ;  
отсортировать  $\Pi'$  по убыванию  $s(\pi)$ .
4.  $\Pi^* := \Pi'$ .
5. GPI := вычислить по  $\Pi^*$  (определение — глава 5).
6. Вернуть  $(\Pi^*, GPI)$ .

Значимость  $s(\pi)$  согласована с Определениями 3–4: для базового паттерна «объект – признак – целевая характеристика» она определяется коэффициентами внимания рёбер пути, и именно эти значения выступают весами  $w_{ik}$  при вычислении компоненты  $GPI_w$ . Число путей ограничено сверху: при  $L_{max} = 2$  их не более  $|F| \cdot K$ , поэтому сложность шагов 1–3 составляет  $O(|F| \cdot K)$ ; при  $L_{max} > 2$  перебор ведётся только по типам связей, допустимым онтологией (включая рёбра «коррелирует\_с» между целевыми характеристиками), что исключает экспоненциальный рост числа маршрутов, характерный для перебора путей в произвольном графе.

Конкретные значения гиперпараметров, объёмы выборок и пороги значимости рассмотрены в главах 3 и 4.

#### 2.9.4 Сравнение способов представления знаний в интеллектуальных системах

Для обоснования выбора графа знаний в качестве основы нейросимвольной архитектуры выполнено сравнение различных способов представления знаний, применяемых в интеллектуальных системах. Результаты сравнения приведены в таблице 4.

Граф знаний обеспечивает совместное представление сущностей предметной области и связей между ними в единой структуре. В отличие от векторных представлений, в которых отношения между объектами кодируются неявно, графовая модель сохраняет семантику связей в явном виде, что делает возможным их использование при анализе и интерпретации результатов. По сравнению с онтологиями граф знаний сохраняет формализованное представление предметной области, но при этом допускает использование методов машинного обучения и графовых нейронных сетей.

Таблица 4 – Сравнение способов представления знаний в интеллектуальных системах

| Аспект                              | Векторные представления                 | Онтологии                 | Граф знаний (предлагаемый подход)                 |
|-------------------------------------|-----------------------------------------|---------------------------|---------------------------------------------------|
| Представление сущностей             | Эмбединги                               | Концепты и классы         | Узлы графа                                        |
| Представление связей                | Неявное                                 | Формализованные отношения | Явные связи с весами                              |
| Интерпретируемость                  | Низкая                                  | Высокая                   | Высокая                                           |
| Работа с данными                    | Непосредственно используется в обучении | Ограничена                | Непосредственно используется в обучении через GNN |
| Масштабируемость                    | Высокая                                 | Средняя                   | Высокая                                           |
| Учёт неопределённости               | Через параметры модели                  | Ограничен                 | Через веса связей                                 |
| Использование предметных знаний     | Неявное                                 | Явное                     | Явное                                             |
| Интеграция с нейросетевыми методами | Полная                                  | Ограниченная              | Полная                                            |

Таким образом, граф знаний объединяет преимущества символического и нейросетевого подходов к представлению знаний. С одной стороны, он обеспечивает явное описание сущностей и их взаимосвязей, а с другой – допускает использование обучаемых представлений и алгоритмов распространения информации по графовой структуре.

Совместное применение графа знаний и графовых нейронных сетей обеспечивает формирование предсказаний с учётом структуры предметной области. В результате предложенная нейросимволическая архитектура сочетает возможности обучения на данных с интерпретируемостью, основанной на явном представлении знаний в графовой форме.

## 2.10 Роль метода в системах поддержки принятия решений

Предложенный метод предназначен для применения в составе систем поддержки принятия решений (СППР). Система поддержки принятия решений не принимает решение автономно, а предоставляет лицу, принимающему решение (ЛПР), обоснованную информацию для решения. На выходе метода формируется не только прогноз целевой характеристики объекта анализа, но и проверяемое объяснение этого прогноза – объясняющие пути в графе знаний с количественной оценкой их качества метрикой интерпретируемости путей (GPI). Тем самым ЛПР получает обоснованный результат, который можно соотнести с предметными знаниями и при необходимости оспорить.

В этом состоит принципиальное отличие интерпретируемого подхода от моделей типа «чёрный ящик»: в ответственных задачах принятия решений (медицинская диагностика, кадровый отбор, модерация контента) прогноз без



обоснования неприменим, поскольку специалист не может оценить его достоверность и нести ответственность за решение. Интерпретируемый вывод даёт ЛПР основания доверять результату: специалист рассматривает прогноз, объясняющие пути и вклад отдельных модальностей и на их основе принимает либо корректирует решение. Примеры подобного сценария приведены в кросс-доменных приложениях – поддержка дерматоскопической диагностики (приложение К) и поддержка решения о модерации токсичного контента и классификации эмоций (приложение Л).

## 2.11 Выводы по главе 2

Основным результатом настоящей главы является доменно-инвариантный метод построения гетерогенного графа знаний для описания мультимодальных данных и нейросимвольная архитектура интерпретируемого вывода (графовая сеть внимания над графом знаний) с метрикой интерпретируемости путей GPI; перечисленные ниже выводы конкретизируют этот результат.

1. Разработан метод нейросимвольной интеграции для интерпретируемого анализа мультимодальных данных, основанный на совместном использовании мультимодальных нейросетевых энкодеров и графа знаний. Предложена четырёхуровневая архитектура, включающая извлечение признаков, их семантическую интеграцию, нейросимвольный вывод и формирование интерпретируемых объяснений.

2. Выполнена математическая формализация задачи интерпретируемого анализа мультимодальных данных. Определены представление мультимодальных данных, графовая модель знаний, отображение анализа и механизм гибридной интеграции модальностей на основе механизма внимания.

3. Разработана онтологическая модель предметной области, включающая концепты «объект анализа», «признак» и «целевая характеристика», а также систему семантических отношений между ними. Онтология формализована средствами OWL/RDF и служит основой построения графа знаний.

4. Граф знаний определён как интегрирующая структура, обеспечивающая совместное представление объектов анализа, признаков и целевых характеристик в единой семантической модели. Формализована процедура инстанцирования графа знаний на основе мультимодальных данных и онтологии предметной области.

5. Разработан механизм нейросимвольной интеграции, обеспечивающий преобразование мультимодальных данных в элементы графа знаний и использование графовой структуры при формировании предсказаний. Показано, что граф знаний

одновременно выполняет функции представления предметных знаний и формирования предсказаний в составе нейросимвольной архитектуры.

6. Предложен подход к формированию интерпретируемых объяснений на основе анализа путей в графе знаний, связывающих наблюдаемые признаки с целевыми характеристиками. Объяснения формируются одновременно с предсказаниями и опираются на структуру графовой модели знаний.

7. Разработанные процедуры построения графа знаний, обучения нейросетевых компонентов, нейросимвольного вывода и формирования интерпретируемых объяснений формализованы в виде алгоритмов.

8. Показана предметная независимость предложенного метода. Онтологический паттерн «объект анализа – признак – целевая характеристика», механизм нейросимвольной интеграции и подход к построению объяснений образуют инвариантное ядро, допускающее перенос между различными предметными областями.

### 3 Применение метода к задаче интерпретируемого анализа цифрового следа

#### 3.1 Постановка прикладной задачи

В данной главе рассматривается применение разработанного нейросимвольного метода к задаче интерпретируемого анализа цифрового следа пользователей социальной сети. Разработанный метод является доменно-инвариантным и может применяться в различных предметных областях, требующих интерпретируемого анализа мультимодальных данных в системах поддержки принятия решений.

Настоящая глава представляет собой конкретизацию общего доменно-инвариантного метода (глава 2) для данной предметной области: здесь определяются состав мультимодальных признаков, предметная онтология и структура графа знаний, которые затем используются в экспериментальной проверке метода (глава 5).

В качестве предметной области выбран анализ личности по цифровому следу пользователя социальной сети. Данная задача рассматривается как пример прикладной реализации разработанного метода в системах поддержки принятия решений; результаты анализа могут использоваться специалистами (кадровыми, психологическими, медицинскими) для принятия обоснованных решений на основе интерпретируемых прогнозов и выявленных семантических связей. На данном примере последовательно рассматриваются основные этапы предложенного метода: извлечение признаков из различных модальностей данных, построение онтологии и графа знаний, формирование предсказаний и построение интерпретируемых объяснений. Практическая реализация метода для данной предметной области и её последующее обобщение до доменно-инвариантного онтологического паттерна проектирования представлены автором в [204\*].

Анализ личности в цифровой среде требует интеграции нескольких модальностей цифрового следа:

- Текстовая модальность (никнеймы, статусы, посты) – вербальное самовыражение, осознанные установки и мировоззрение. Текст является наиболее явной формой самопрезентации, легко поддающейся рефлексии пользователем и, следовательно, потенциально подверженной влиянию социальной желательности [205].

- Визуальная модальность (аватары, фотографии профиля) – невербальные сигналы, включая возраст, пол, эмоциональную экспрессию и социальный контекст. Визуальный канал передаёт информацию, в меньшей степени контролируемую

сознанием пользователя: микровыражения эмоций [206], особенности визуальной самопрезентации (селфи или фотографии с другими людьми) [80].

– Табличная модальность (данные цифровой активности) – характеристики активности пользователя, временные параметры публикаций и метаданные профиля, отражающие особенности цифрового поведения. Частота публикаций, время активности и продолжительность сессий могут быть связаны с особенностями темперамента и регуляторных функций.

Целевыми характеристиками прикладной задачи являются личностные черты, измеряемые психометрическими методиками: 20 факторов опросника Кэттелла (общительность (А), интеллект (В), эмоциональная устойчивость (С) и др.), а также 6 обобщённых факторов, полученных на основе 18 психологических методик (невротичность, рефлексивность, интеллектуальность, социальность, целеустремлённость и конформность). Целевые значения определяются по результатам психологического тестирования. Объём размеченной выборки составляет 1 722 пользователя социальной сети ВКонтакте.

Выбор данной предметной области обусловлен необходимостью одновременного анализа текстовых, визуальных и табличных данных, а также наличием измеряемых целевых характеристик. Это делает возможным последовательное рассмотрение всех этапов предложенного метода: извлечение признаков, построение графа знаний, нейросимвольный вывод и формирование интерпретируемых объяснений.

Важным требованием к решению задачи является возможность проследить, какие признаки цифрового следа и какие семантические связи графа знаний обусловили итоговую оценку личностной характеристики.

Формальная постановка прикладной задачи представлена на рисунке 7.

На вход модели поступает мультимодальное описание пользователя

$$M(o) = \langle T, V, D \rangle,$$

где  $T$  – текстовые признаки,  $V$  – визуальные признаки,  $D$  – признаки цифровой активности.

Результатом анализа является вектор целевых характеристик

$$Y = (Y_1, Y_2, \dots, Y_{26}),$$

содержащий 20 первичных факторов опросника Кэттелла и 6 интегральных факторов личности.

Таким образом, задача заключается в построении отображения

$$F: X \rightarrow Y,$$

определяющего оценку значений личностных характеристик по мультимодальным данным цифрового следа пользователя. В рамках предложенного метода формирование предсказаний выполняется с использованием графа знаний и механизмов нейросимвольного вывода, что обеспечивает интерпретируемость результатов анализа.

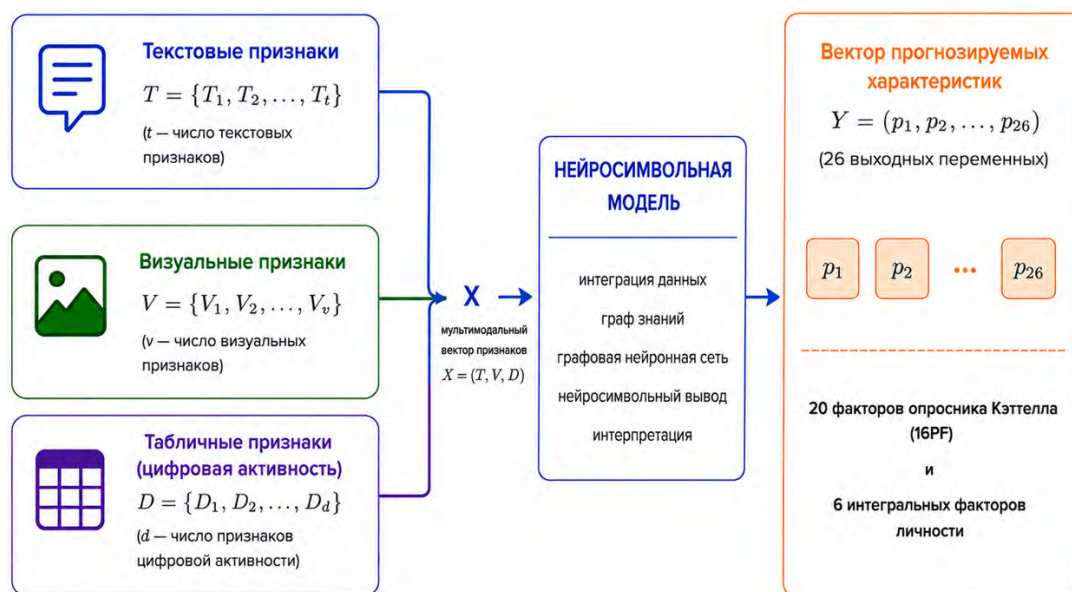


Рисунок 7 – Формальная постановка задачи анализа личности по цифровому следу

Дополнительным результатом работы модели является множество объяснений  $I$ , представляющее собой набор семантических путей в графе знаний, связывающих признаки цифрового следа с предсказанными характеристиками личности.

### 3.2 Описание данных

Для реализации предложенного нейросимвольного метода сформирована база данных, объединяющая результаты психометрического тестирования и мультимодальные цифровые следы пользователей социальной сети ВКонтакте. Типизация мультимодальных данных профиля пользователя для рассматриваемой задачи, в соответствии с формальной моделью главы 2 (Определения 5-6), приведена в таблице 5.

Таблица 5 – Типизация мультимодальных данных профиля социальной сети «ВКонтакте»

| Индекс модальности $m$ | Домен $d$ | Содержание                         | Примеры типов $t$                                |
|------------------------|-----------|------------------------------------|--------------------------------------------------|
| 1                      | 1         | аватар пользователя (фото профиля) | фотография человека; рисунок/арт; логотип        |
| 1                      | 2         | изображения публикаций             | личные фотографии; цветовая гамма; пейзажи; мемы |
| 2                      | 1         | никнейм                            | настоящее имя; псевдоним                         |

Продолжение таблицы 5

|   |   |                                            |                                        |
|---|---|--------------------------------------------|----------------------------------------|
| 2 | 2 | статус                                     | цитата; авторский текст                |
| 2 | 3 | посты и комментарии                        | собственные публикации;<br>комментарии |
| 3 | 1 | социально-демографические атрибуты профиля | категориальные, числовые признаки      |
| 3 | 2 | числовые показатели активности             | количество друзей, постов, лайков      |

Здесь  $m = 1$  – визуальная модальность (изображения),  $m = 2$  – текстовая модальность,  $m = 3$  – табличные данные; соответственно  $M^V = \{1\}$ ,  $M^T = \{2\}$ ,  $M^D = \{3\}$ . Разбиение на домены и типы определяется составом доступных данных социальной сети.

### 3.2.1 Сбор и описание выборки

Формирование базы данных осуществлялось в период с 2007 по 2025 гг. и включало несколько этапов сбора психометрических и цифровых данных пользователей социальной сети ВКонтакте. На первом этапе проводился набор добровольцев и их психометрическое обследование. Набор участников осуществлялся на базе Самарского национального исследовательского университета имени академика С.П. Королёва при участии Самарского государственного социально-педагогического университета и кадрового агентства «Интеллект».

Всего 821 участник прошёл полную батарею из 18 психологических методик, ещё 901 человек – 16-факторный личностный опросник Кэттелла. Итоговый объём выборки с психометрическими метками составил 1 722 пользователя. Тестирование проводилось посредством веб-интерфейса и занимало от 4 до 5 часов.

Основным инструментом оценки личности выступал 16-факторный личностный опросник Кэттелла (16PF), предназначенный для оценки 20 первичных и 6 глобальных факторов личности. Выбор данной методики обусловлен возможностью получения более детализированного описания структуры личности по сравнению с моделями, основанными на меньшем числе факторов. Благодаря этому становится возможным анализ более тонких взаимосвязей между особенностями цифровой активности и личностными характеристиками. Психометрической основой служат модель и опросник 16PF [95, 96] и исследования устойчивости личностных черт [208]; связь цифрового поведения и самопрезентации личности рассматривается в [209, 210].

После прохождения тестирования участники предоставляли информированное согласие на обработку их публичных данных из социальной сети ВКонтакте. Сбор данных осуществлялся посредством VK API (users.get, wall.get, photos.get, friends.get) с

соблюдением действующих ограничений по частоте запросов. Перечень извлекаемых типов данных представлен в таблице 6.

Для обучения моделей извлечения признаков объём размеченной выборки оказался недостаточным. В связи с этим был выполнен дополнительный сбор данных из открытых профилей пользователей социальной сети ВКонтакте без прохождения психометрического тестирования. В результате была сформирована расширенная выборка объёмом 31 286 пользователей.

Таблица 6 – Извлекаемые данные о каждом пользователе

| Тип данных                 | Метод                  | Скачиваемая информация                                                                 |
|----------------------------|------------------------|----------------------------------------------------------------------------------------|
| Метаданные профиля         | users.get, поля fields | Возраст, пол, город, семейное положение, образование                                   |
|                            |                        | Число друзей, подписчиков, сообществ                                                   |
|                            |                        | Дата регистрации, последняя активность                                                 |
| Текстовые данные           | wall.get               | Посты на стене пользователя (текст, дата публикации, число лайков и репостов)          |
|                            |                        | Статус профиля                                                                         |
|                            |                        | Раздел «О себе» (при наличии доступа)                                                  |
| Визуальные данные          | photos.get             | Главная фотография профиля (аватар)                                                    |
|                            |                        | Фотографии из альбома «Фотографии со мной»                                             |
|                            |                        | Обложка профиля (при наличии)                                                          |
| Данные цифровой активности | wall.get               | Временные ряды активности, построенные по временным меткам публикаций                  |
|                            |                        | Соотношение различных типов контента (посты, репосты, комментарии)                     |
|                            |                        | Метрики взаимодействия (среднее число лайков на публикацию, коэффициент вовлечённости) |
|                            | friends.get            | Число друзей                                                                           |

Полученная расширенная выборка использовалась исключительно для предобучения моделей извлечения признаков без доступа к психометрическим меткам. Обучение моделей предсказания личностных характеристик и оценка их качества выполнялись только на размеченной выборке из 1 722 пользователей.

Таким образом, были сформированы две взаимосвязанные выборки: размеченная выборка пользователей с психометрическими метками, используемая для обучения и оценки моделей предсказания целевых характеристик, и расширенная неразмеченная выборка, предназначенная для предобучения моделей извлечения признаков из мультимодальных данных.

После завершения сбора данных была выполнена их предварительная очистка. Из расширенной выборки исключались аккаунты, демонстрирующие признаки неаутентичного поведения, включая аномально высокую активность публикаций, значительную долю повторяющегося контента и недостаточную заполненность профиля.

В результате было отфильтровано 13 318 аккаунтов (42% от исходной расширенной выборки). Итоговый объём датасета, использованного для предобучения моделей извлечения признаков, составил 17 968 пользователей.

Структура сформированного датасета представлена на рисунке 8. Сформированный датасет включает метаданные профилей пользователей, текстовые публикации, изображения профиля и результаты психометрического тестирования. Для пользователей размеченной выборки доступны значения целевых характеристик личности, тогда как расширенная выборка содержит только данные цифрового следа и используется на этапе предобучения моделей извлечения признаков.

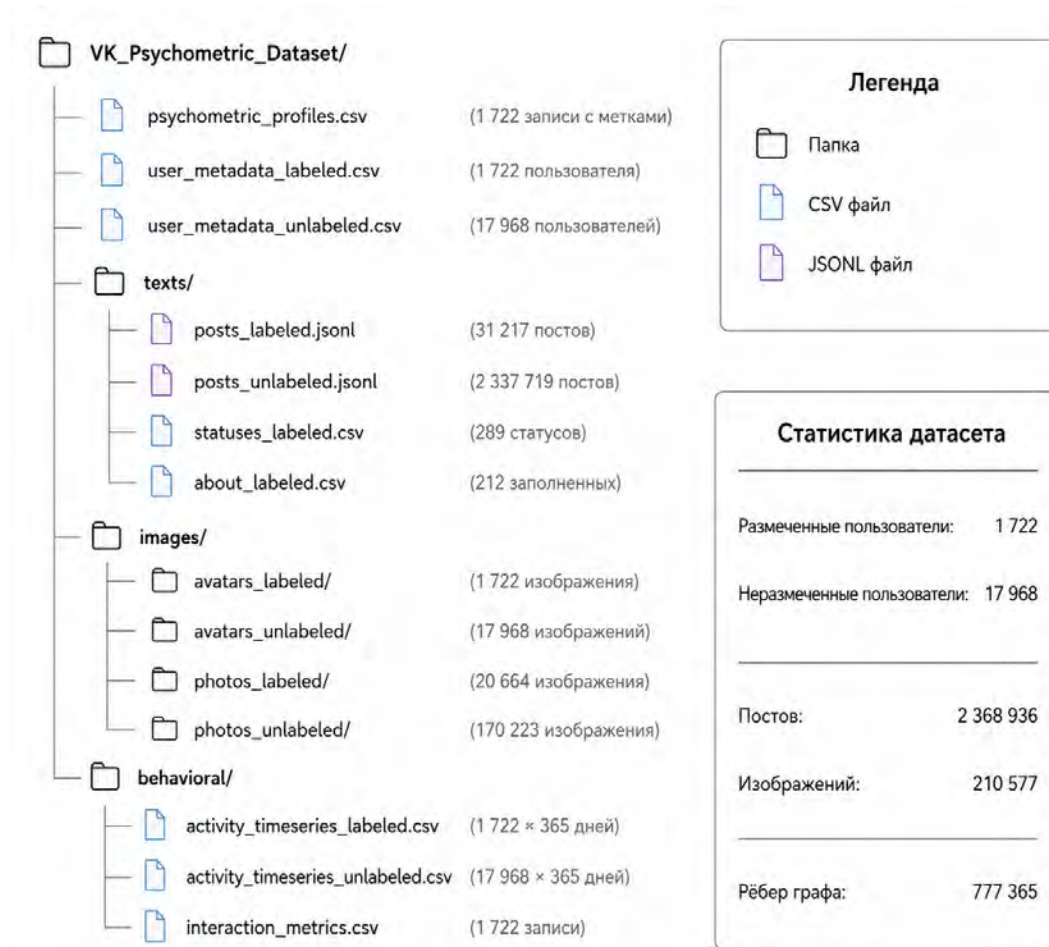


Рисунок 8 – Структура и состав датасета VK\_Psychometric\_Dataset

Размеченная часть датасета содержит данные 1 722 пользователей и более 31 тыс. публикаций. Расширенная неразмеченная выборка включает данные 17 968 пользователей и более 2,3 млн публикаций. Собранные данные охватывают текстовую, визуальную и табличную модальности, необходимые для реализации предложенного метода мультимодального анализа.

Далее был проведён анализ основных характеристик сформированных выборок. Демографические характеристики пользователей представлены в таблице 7.



Таблица 7 – Демографические характеристики выборок

| Характеристика | Выборка с психометрическими метками | Расширенная выборка                       |
|----------------|-------------------------------------|-------------------------------------------|
| Размер выборки | 1 722 пользователя                  | 17 968 пользователей                      |
| Пол            | 42,4 % женщины, 57,6 % мужчины      | 57,0 % женщины, 43,0 % мужчины            |
| Возраст        | 27,1 ± 8,2 года                     | 25,8 ± 7,4 года                           |
| География      | преимущественно Самарская область   | пользователи из различных регионов РФ     |
| Образование    | 68,3 % высшее образование           | данные доступны не для всех пользователей |

Выборка с психометрическими метками преимущественно представлена пользователями молодого и среднего возраста, имеющими высшее или неоконченное высшее образование. Это обусловлено особенностями формирования выборки на базе образовательных организаций и кадрового агентства.

Несмотря на региональную концентрацию участников психометрического обследования, расширенная выборка охватывает пользователей из различных регионов Российской Федерации. Это повышает разнообразие цифровых данных, используемых на этапе предобучения моделей извлечения признаков.

Сформированные выборки используются совместно в рамках двухэтапной схемы обучения, описанной в главе 2: расширенная выборка применяется для предобучения моделей извлечения признаков, а выборка с психометрическими метками – для обучения и оценки моделей предсказания целевых характеристик.

Дополнительно был проведён анализ распределений психометрических характеристик и основных параметров мультимодальных данных. Для большинства психометрических факторов наблюдались распределения, близкие к нормальным. Лишь для фактора В (интеллект) был выявлен незначительный положительный скос распределения, что может быть связано с высоким образовательным уровнем участников исследования. Доля экстремальных значений оставалась незначительной и не требовала дополнительной коррекции данных. Основные статистические характеристики мультимодальных данных представлены в таблице 8.

Таблица 8 – Основные характеристики мультимодальных данных

| Модальность                | Характеристика                                                               |
|----------------------------|------------------------------------------------------------------------------|
| Текстовая                  | публикации пользователей, статусы профиля и текстовые описания               |
| Визуальная                 | фотографии профиля и пользовательские изображения                            |
| Данные цифровой активности | показатели активности, взаимодействия и социальной связанности пользователей |

Анализ показал существенную вариативность параметров цифрового следа как между пользователями, так и между различными модальностями данных. Наличие вариативности является важным условием для выявления статистических

зависимостей между особенностями цифрового поведения и личностными характеристиками.

Для предварительной оценки взаимосвязей между цифровыми признаками и психометрическими характеристиками был выполнен корреляционный анализ с использованием коэффициента ранговой корреляции Спирмена. Полученные результаты использовались при формировании связей между признаками и целевыми характеристиками в графе знаний, рассматриваемом далее.

Для предварительного выявления статистических зависимостей между характеристиками цифрового следа и личностными особенностями был проведён корреляционный анализ. Рассматривались 39 признаков цифровой активности и 20 факторов опросника Кэттелла, что соответствует 780 анализируемым парам признаков.

В таблице 9 приведены наиболее значимые корреляционные зависимости, выявленные в ходе анализа. Большинство статистически значимых связей находятся в диапазоне  $|r| = 0,20-0,45$ , соответствующем типичным значениям корреляций между поведенческими индикаторами и личностными характеристиками в психологических исследованиях.

Таблица 9 – Наиболее значимые корреляции между признаками цифрового следа и факторами личности

| Признак цифрового следа                   | Фактор личности                | $r$   | p-value |
|-------------------------------------------|--------------------------------|-------|---------|
| Частота публикаций                        | F2 (экстраверсия)              | 0,42  | <0,001  |
| Число друзей                              | A (общительность)              | 0,38  | <0,001  |
| Позитивная эмоциональная окраска статуса  | C (эмоциональная устойчивость) | 0,35  | <0,001  |
| Разнообразие тематик интересов            | Q1 (радикализм)                | 0,33  | <0,001  |
| Число подписок на сообщества              | B (интеллект)                  | 0,29  | <0,001  |
| Эмоция радости на фотографиях             | F (беспечность)                | 0,27  | <0,001  |
| Коэффициент кластеризации социальной сети | Q2 (самостоятельность)         | -0,26 | <0,001  |
| Особенности литературного стиля никнейма  | I (чувствительность)           | 0,24  | <0,001  |
| Длина раздела «О себе»                    | M (абстрактность)              | 0,22  | <0,001  |
| Среднее число отметок «Нравится»          | H (социальная активность)      | 0,21  | <0,001  |

Направление выявленных зависимостей согласуется с существующими представлениями психологии личности. Так, экстраверсия положительно связана с частотой публикаций и интенсивностью цифровой активности, общительность – с размером социальной сети пользователя, а самостоятельность – с меньшей плотностью социальных связей. Выявленные статистически значимые зависимости использовались при формировании семантических связей между признаками и целевыми характеристиками в графе знаний, рассматриваемом далее.

Наличие статистически значимых зависимостей для признаков текстовой, визуальной и табличной модальностей подтверждает целесообразность использования мультимодального подхода для анализа цифрового следа пользователей.

Для обучения и оценки моделей выборка из 1 722 пользователей с психометрическими метками была разделена на обучающую, валидационную и тестовую подвыборки в соотношении 70%, 15% и 15% соответственно. Разделение выполнялось с использованием стратифицированной выборки, обеспечивающей сохранение распределений основных характеристик пользователей во всех подвыборках. В качестве признаков стратификации использовались пол, возраст и показатели социальной активности.

В результате были сформированы:

- обучающая выборка: 1 206 пользователей;
- валидационная выборка: 258 пользователей;
- тестовая выборка: 258 пользователей.

Сформированное разделение сохраняет распределения основных демографических характеристик и параметров цифровой активности во всех подвыборках, снижая риск систематического смещения при оценке качества моделей.

Дополнительно расширенная неразмеченная выборка, включающая 17 968 пользователей, использовалась на этапе предобучения нейросетевых энкодеров в соответствии с двухэтапной схемой обучения, описанной в главе 2. На данном этапе выполнялось обучение моделей извлечения признаков для текстовой, визуальной и табличной модальностей без использования психометрических меток.

Предварительное обучение энкодеров на расширенной неразмеченной выборке способствует формированию более устойчивых признаковых представлений и снижению риска переобучения при последующем обучении моделей предсказания личностных характеристик.

### 3.2.2 Этические аспекты обработки данных

Все участники исследования предоставили информированное согласие на использование результатов психометрического тестирования и данных цифрового следа в научных целях. Участникам разъяснялись состав собираемых данных, цели их использования и порядок обработки персональной информации.

В исследовании использовались исключительно данные, доступные пользователям в открытой части профилей социальной сети ВКонтакте. К ним

относились текстовые публикации, изображения профиля, сведения о социальной активности и другие общедоступные характеристики цифрового поведения.

Для защиты конфиденциальности участников идентифицирующая информация была исключена из рабочей версии датасета и заменена обезличенными идентификаторами. Результаты исследования представляются только в агрегированном виде без возможности идентификации конкретных пользователей.

При формировании датасета были реализованы меры по предотвращению деанонимизации пользователей, включая удаление прямых персональных идентификаторов и обобщение редких категорий социально-демографических признаков.

Обработка данных осуществлялась в соответствии с требованиями законодательства Российской Федерации в области защиты персональных данных. Использование обезличенных данных и наличие информированного согласия участников обеспечивали соблюдение этических требований при проведении исследования.

### 3.2.3 Ограничения датасета

Несмотря на значительный объем собранных данных, сформированный датасет имеет ряд ограничений, которые необходимо учитывать при интерпретации результатов исследования.

Выборка с психометрическими метками характеризуется региональной концентрацией: большинство участников проживают в Самарской области. В связи с этим выявленные закономерности могут частично отражать особенности региональной аудитории социальной сети и требуют дополнительной проверки на выборках с более широким географическим охватом.

Возрастная структура выборки также является неоднородной. Основная часть пользователей относится к возрастной группе от 18 до 52 лет, тогда как представители старших возрастных категорий представлены в меньшей степени. Это может ограничивать качество обобщения моделей на пользователей старшего возраста.

Дополнительным фактором является эффект самоотбора участников исследования. Поскольку прохождение психометрического обследования требовало значительных временных затрат, в выборке могут быть представлены пользователи с повышенной мотивацией к участию в исследованиях и более высоким уровнем образования по сравнению со средними показателями популяции.

Следует также учитывать временной аспект формирования датасета. Сбор данных осуществлялся на протяжении длительного периода, в течение которого изменялись особенности цифровой среды, форматы пользовательского контента и механизмы взаимодействия в социальных сетях. В связи с этим при практическом применении моделей может потребоваться их периодическая актуализация на более современных данных.

Наконец, данные были собраны исключительно из социальной сети ВКонтакте. Особенности аудитории и характер цифровой активности пользователей других социальных платформ могут отличаться, что ограничивает прямой перенос полученных результатов на иные цифровые среды.

Указанные ограничения не препятствуют использованию датасета для апробации предложенного метода, однако определяют направления дальнейшего расширения выборки и проверки обобщающей способности разработанных моделей.

### 3.3 Текстовая модальность

В рамках рассматриваемой предметной области текстовая модальность выступает одним из источников признаков для последующего формирования графа знаний. К текстовым данным относятся никнеймы пользователей, статусы профиля и текстовые публикации, отражающие особенности самопрезентации и цифрового поведения.

В соответствии с методом, представленным в главе 2, обработка текстовой модальности включает извлечение признаков с использованием нейросетевых моделей и их последующее преобразование в семантические элементы графа знаний. В качестве текстовых признаков используются как непосредственно наблюдаемые характеристики текста, так и результаты автоматической классификации его содержательных и структурных особенностей.

#### 3.3.1 Функциональная типология никнеймов

Никнейм пользователя рассматривается как краткая форма цифровой самопрезентации и может содержать сведения о предпочтениях, интересах и особенностях самовыражения. Для преобразования никнеймов в структурированные признаки была разработана типология, учитывающая особенности русскоязычного сегмента социальных сетей.

На основе анализа 19 690 никнеймов пользователей ВКонтакте (включая 1 722 пользователя с психометрическими метками) сформирована система категорий, описывающих содержательные, словообразовательные и графические особенности

никнеймов. Система включает 6 смысловых, 8 словообразовательных и 4 графические категории. Полученная категориальная разметка используется в качестве источника признаков текстовой модальности и позволяет преобразовать неструктурированные текстовые данные в формализованное представление для последующего включения в граф знаний.

Автоматическая классификация никнеймов выполняется с использованием трансформерной модели RoBERTa, обученной по двухэтапной схеме [212\*], описанной в главе 2. Полученные категории используются в качестве признаков текстовой модальности и далее включаются в граф знаний посредством связей между признаками и целевыми характеристиками. Подробное описание категорий, параметров обучения модели и результатов сравнительного анализа приведено в приложениях. Векторное представление текстовой модальности формируется как:

$$e_T = f_{RoBERTa}(T) \in \mathbb{R}^{768},$$

$$e_T = f_{RuBERT}(T) \in \mathbb{R}^{768},$$

где  $T$  – текстовые данные объекта анализа;  $f_{RoBERTa}, f_{RuBERT}$  – предобученный энкодер, используемый в режиме фиксированных параметров;  $e_T$  – вектор признаков представлений.

### 3.3.2 Типология и тематика пользовательских статусов

Статусы пользователей представляют собой более динамичный элемент цифровой самопрезентации по сравнению с никнеймами и отражают актуальные интересы, эмоциональное состояние и особенности мировоззрения. Для преобразования статусов в формализованные признаки выполняется их автоматическая классификация по нескольким группам характеристик.

Разметка обучающей выборки никнеймов и статусов по категориям типологии выполнялась двумя независимыми экспертами; достигнута высокая согласованность разметки по коэффициенту  $\kappa$  Козна, расхождения разрешались консенсусом.

В рамках анализа выделяются функциональные характеристики статусов, их тематическая направленность и мировоззренческая окраска. Такой подход учитывает как содержательные характеристики текста, так и особенности его использования пользователем в процессе цифровой самопрезентации.

Тематическая классификация включает 10 категорий, а мировоззренческая – 4 типа установок, отражающих особенности восприятия и интерпретации окружающей действительности.

Дополнительно анализируются эмодзи, рассматриваемые как самостоятельные элементы передачи эмоциональной информации. Включение эмодзи в процесс обработки позволяет учитывать особенности эмоциональной окраски коротких текстов и повышает качество их интерпретации.

Автоматическая классификация статусов выполняется с использованием модели RuBERT, адаптированной для обработки русскоязычных текстов. Полученные категории используются в качестве признаков текстовой модальности и далее включаются в граф знаний посредством связей между признаками и целевыми характеристиками. Подробное описание категорий, параметров обучения и результатов сравнительного анализа приведено в приложениях.

Таким образом, анализ никнеймов и пользовательских статусов формирует совокупность семантических признаков текстовой модальности, характеризующих особенности цифровой самопрезентации пользователя и используемых при последующем построении графа знаний.

### 3.4 Визуальная модальность

Визуальная модальность является вторым источником признаков цифрового следа пользователя. В соответствии с предложенным методом изображения профиля обрабатываются нейросетевыми моделями, а результаты анализа преобразуются в категориальные признаки, используемые при построении графа знаний.

Из изображений профиля извлекаются несколько групп признаков, характеризующих демографические особенности пользователя, тип визуальной самопрезентации и эмоциональное содержание изображения. Тип визуального контента описывается 6 категориями, семантика изображений животных представлена 10 категориями, а эмоциональная экспрессия определяется по 7 базовым эмоциям.

К первой группе относятся демографические признаки, включающие оценку возраста и пола пользователя. Данные характеристики используются как дополнительные признаки при анализе взаимосвязей между цифровым поведением и личностными особенностями.

Вторая группа признаков описывает тип визуального контента. Выделяются 6 категорий, включающих индивидуальные фотографии, групповые изображения, фотографии с животными, изображения животных, пейзажи и объекты, а также отсутствие изображения профиля. Эти признаки отражают особенности визуальной самопрезентации пользователя.

Третья группа связана с анализом семантического содержания изображений животных, присутствующих на аватарах пользователей. Для описания данного типа контента используются 10 категорий животных, выступающих дополнительными признаками визуальной модальности. Выделенные категории используются для сопоставления особенностей визуального и текстового самовыражения пользователя.

Четвёртая группа признаков характеризует эмоциональную экспрессию, определяемую по изображению лица. Анализ выполняется по 7 базовым эмоциям, отражающим аффективные особенности визуальной самопрезентации пользователя. Полученные эмоциональные характеристики включаются в граф знаний и используются при формировании мультимодального представления объекта анализа.

Извлечение признаков выполняется с использованием свёрточных нейронных сетей, обученных на специализированных наборах изображений. Подробные сведения о моделях, параметрах обучения и сравнительном анализе представлены в приложениях. Векторное представление визуальной модальности:

$$e_V = f_{ResNet50}(V) \in \mathbb{R}^{2048},$$

где  $V$  – визуальные данные;  $f_{ResNet50}$  – предобученный энкодер, используемый в режиме фиксированных параметров;  $e_V$  – вектор признаковых представлений.

Таким образом, визуальная модальность формирует совокупность признаков, описывающих демографические характеристики, особенности визуальной самопрезентации и эмоциональную экспрессию пользователя. Полученные признаки включаются в граф знаний и участвуют в последующей мультимодальной интеграции.

### 3.5 Табличная (числовая) модальность

Третьей модальностью, рассматриваемой в рамках апробации предложенного метода, являются данные цифровой активности пользователей. В отличие от текстовой и визуальной модальностей, данный тип данных представлен в виде структурированного набора числовых и категориальных признаков, характеризующих особенности поведения пользователя в социальной сети.

Целью обработки табличной модальности является формирование признакового описания цифровой активности, используемого при последующем построении графа знаний и формировании мультимодального представления объекта анализа. Табличная модальность образует числовой вектор признаков:

$$e_D = f_{tab}(x_D) \in \mathbb{R}^{d_D}, x_D \in \mathbb{R}^{d_D},$$

где  $x_D$  – вектор табличных признаков, характеризующих цифровую активность объекта;  $f_{tab}$  – предобученный энкодер, используемый в режиме фиксированных



параметров;  $e_D$  – вектор признаков представлений;  $d_D$  – размерность пространства табличных признаков.

### 3.5.1 Признаковое описание цифровой активности

На основе данных профиля пользователя формируется набор признаков, отражающих различные аспекты цифрового поведения.

К первой группе относятся показатели активности пользователя, включая характеристики публикационной активности, частоту размещения контента и регулярность взаимодействия с социальной сетью.

Вторая группа включает показатели вовлечённости, характеризующие реакцию аудитории на публикуемый контент и интенсивность социальных взаимодействий пользователя.

Третья группа представлена метаданными профиля, отражающими структуру социальных связей пользователя и особенности его присутствия в цифровой среде.

Отдельную группу образуют временные характеристики цифровой активности, описывающие распределение активности пользователя во времени и рассматриваемые далее в подпараграфе 3.5.2.

Помимо количественных признаков, формируется категориальное описание цифрового поведения пользователя. Для этого используется типология активности, разработанная на основе анализа цифрового следа пользователей социальной сети и представленная в подпараграфе 3.5.3. Полученный тип активности рассматривается как самостоятельный признак и включается в граф знаний в виде отдельного узла, связанного с другими признаками и целевыми характеристиками объекта анализа.

Таким образом, табличная модальность формирует совокупность признаков, характеризующих особенности цифрового поведения пользователя и дополняющих информацию, извлекаемую из текстовой и визуальной модальностей. Полученные признаки используются при последующей мультимодальной интеграции и формировании графа знаний.

### 3.5.2 Темпоральный анализ активности

Помимо содержательных характеристик цифрового следа, информативными являются временные особенности активности пользователей. Время публикаций, регулярность появления контента и распределение активности в течение суток отражают устойчивые особенности цифрового поведения и могут рассматриваться как дополнительный источник признаков при анализе личности.

Для каждого пользователя вычисляются характеристики публикационной активности, включающие частоту публикаций, показатели регулярности, распределение активности по времени суток и дням недели. Полученные признаки формируют темпоральное описание цифрового поведения пользователя.

Для выявления типичных стратегий цифровой активности применяется кластеризация пользователей в пространстве темпоральных признаков. Выделенные типы активности используются как категориальные признаки и включаются в граф знаний в виде узлов, связанных с другими признаками и целевыми характеристиками объекта анализа.

В результате кластеризации были выделены пять типов темпоральной активности пользователей, представленные в таблице 10.

Таблица 10 – Типы темпоральной активности пользователей

| Тип активности     | Размер | Характеристика                                  |
|--------------------|--------|-------------------------------------------------|
| Утренние активные  | 16 %   | Преобладание активности в утренние часы         |
| Вечерние умеренные | 36 %   | Основная активность в вечернее время            |
| Ночные активные    | 8 %    | Преобладание ночных публикаций                  |
| Пассивные          | 28 %   | Низкая частота публикаций                       |
| Высокоактивные     | 12 %   | Высокая интенсивность публикационной активности |

Выделенные типы активности отражают различные стратегии цифрового поведения пользователей и используются при формировании мультимодального представления объекта анализа.

Анализ выявил статистически значимые различия между выделенными типами активности по ряду личностных характеристик. В связи с этим темпоральные признаки были включены в состав признакового описания пользователя и использованы при построении графа знаний.

Темпоральные признаки и типы активности включаются в граф знаний наряду с признаками текстовой и визуальной модальностей и участвуют в формировании мультимодального представления объекта анализа.

### 3.5.3 Типологизация активности пользователя

Одним из категориальных признаков табличной модальности является тип активности пользователя. Для его определения используется типология активности, разработанная и апробированная в работе [64\*, 211\*].

Типология включает четыре категории: позитивную, девиантную, делинквентную и нейтральную активность. Позитивный тип характеризуется просоциальным поведением и конструктивным взаимодействием с другими пользователями. Девиантный тип отражает отклонение от неформальных норм

общения при сохранении соответствия правилам платформы. Делинквентный тип связан с поведением, нарушающим правила социальной сети. Нейтральный тип соответствует стандартной пользовательской активности без выраженной поведенческой направленности.

Автоматическое определение типа активности выполняется на основе данных цифрового следа пользователя. Подробное описание процедуры разметки, обучения классификатора и оценки качества представлено в приложении.

Полученный тип активности рассматривается как самостоятельный категориальный признак и включается в граф знаний в виде узла-признака, связанного с другими признаками и целевыми характеристиками объекта анализа.

Таким образом, табличная модальность формирует признаки, отражающие количественные характеристики цифрового поведения, временные особенности активности и обобщённые типы пользовательского поведения, которые используются при построении мультимодального представления объекта анализа.

### 3.6 Предметная онтология и граф знаний

В данном параграфе выполняется переход от общей онтологической модели, представленной в главе 2, к её реализации для задачи анализа цифрового следа пользователей социальной сети. Абстрактные концепты «объект анализа», «признак» и «целевая характеристика» получают конкретное содержательное наполнение, соответствующее рассматриваемой предметной области.

В качестве объекта анализа рассматривается пользователь социальной сети. Роль признаков выполняют характеристики цифрового следа, извлечённые из текстовой, визуальной и табличной модальностей. В качестве целевых характеристик выступают факторы личности, определяемые на основе психометрического тестирования.

Конкретная онтология факторов личности и граф знаний для данной предметной области впервые представлены автором в работе [204\*]. Ниже рассматриваются основные концепты, отношения и структура графа знаний, используемые при реализации предложенного метода.

#### 3.6.1 Объект анализа

Объектом анализа выступает виртуальная личность (Virtual Identity), представляющая пользователя социальной сети как совокупность его цифровых проявлений.

Описание виртуальной личности формируется на основе двух групп характеристик. Первая группа включает атрибуты профиля пользователя, получаемые

из социальной сети: демографические сведения, характеристики активности, параметры цифрового присутствия и другие метаданные профиля. Вторая группа представлена признаками, автоматически извлекаемыми из текстовой, визуальной и табличной модальностей цифрового следа с использованием нейросетевых моделей.

В результате описание виртуальной личности объединяет 28 характеристик профиля, получаемых из API социальной сети, и 13 признаков, автоматически извлекаемых из мультимодальных данных. Представление объединяет исходные данные профиля и результаты их интеллектуального анализа в рамках единой модели пользователя.

Таким образом, виртуальная личность рассматривается как интегральное представление пользователя, объединяющее исходные данные профиля и результаты мультимодального анализа.

Агрегированные характеристики цифрового окружения пользователя (число друзей, подписчиков, сообществ, показатели активности) входят в состав табличной модальности; отдельная социальная модальность не выделяется. Явные связи между пользователями в структуру графа знаний не включаются, поскольку целью исследования является анализ взаимосвязей между признаками цифрового следа и личностными характеристиками, а не моделирование процессов распространения информации в социальной сети.

### 3.6.2 Целевые характеристики

Целевые характеристики представлены психологическими и социально-психологическими конструктами, измеряемыми с использованием психометрических методик.

Основу системы целевых характеристик составляют 20 факторов опросника Кэттелла, включающие первичные и вторичные факторы личности. Дополнительно используются 6 интегральных факторов личности, полученных на основе результатов комплекса психологических методик.

Формально целевые характеристики представляются вектором психологических показателей, значения которых используются при обучении и оценке моделей предсказания личности. Полный перечень факторов и их описание приведены в приложении.

В структуре графа знаний целевые характеристики представлены отдельными узлами, связанными между собой отношениями корреляции и ассоциации,

отражающими выявленные статистические и семантические зависимости между психологическими конструктами.

### 3.6.3 Типы отношений и примеры триплетов

Связи между объектом анализа, признаками и целевыми характеристиками формализуются в виде ориентированных триплетов графа знаний. В рамках рассматриваемой предметной области используются следующие основные типы отношений:

- имеет\_признак – связывает объект анализа с наблюдаемыми или извлечёнными признаками цифрового следа;
- ассоциируется\_с – отражает статистически значимые связи между признаками и целевыми характеристиками;
- коррелирует\_с – описывает взаимосвязи между целевыми характеристиками.

Начальные значения весов рёбер графа знаний определяются статистическими мерами связи между признаками и целевыми характеристиками. Для исключения случайных зависимостей выполняются проверка статистической значимости и поправка на множественность проверок.

Состав узлов и допустимые типы рёбер графа при этом полностью определяются онтологией предметной области (глава 2): статистические меры задают лишь начальные веса рёбер, разрешённых доменами и ко-доменами онтологических предикатов, и не влияют на структуру графа: ребро, не предусмотренное онтологией, не может быть добавлено ни при каком значении статистической меры.

Для рёбер типа «ассоциируется\_с» ( $f_i \rightarrow p_k$ ) вычисляется точечно-бисериальный коэффициент корреляции:

$$r_{ik} = \frac{M_1 - M_0}{s_k} \sqrt{\frac{n_1 n_0}{n^2}}.$$

Статистическая значимость связи проверяется с использованием  $t$ -критерия Стьюдента:

$$t_{ik} = r_{ik} \sqrt{\frac{n-2}{1-r_{ik}^2}}.$$

По вычисленному значению статистики определяется достигаемый уровень значимости  $p_{ik}$ . После поправки Бонферрони ребро включается в граф при выполнении условия:

$$p_{ik} < \alpha'.$$

При этом начальный вес ребра определяется как

$$w_{ik} = \begin{cases} r_{ik}, & p_{ik} < \alpha', \\ 0, & \text{иначе.} \end{cases}$$

Для рёбер типа «коррелирует\_с» ( $p_k \rightarrow p_l$ ) используется коэффициент корреляции Пирсона

$$\rho_{kl} = \frac{\text{cov}(p_k, p_l)}{\sigma_k \sigma_l},$$

где  $f_i$  –  $i$ -й признак объекта анализа;  $p_k$  и  $p_l$  –  $k$ -я и  $l$ -я целевые характеристики соответственно;  $M_1$  и  $M_0$  – средние значения характеристики  $p_k$  в группах объектов соответственно с наличием и отсутствием признака  $f_i$ ;  $s_k$  – стандартное отклонение характеристики  $p_k$ ;  $n_1$  и  $n_0$  – число объектов в соответствующих группах;  $n = n_1 + n_0$  – общий объём выборки;  $t_{ik}$  – значение статистики Стьюдента для проверки значимости коэффициента корреляции  $r_{ik}$ ;  $p_{ik}$  – достигаемый уровень статистической значимости;  $\alpha$  – уровень значимости;  $\alpha' = \alpha/m$  – уровень значимости после поправки Бонферрони;  $m$  – число одновременно проверяемых статистических гипотез;  $w_{ik}$  – начальный вес ребра типа «ассоциируется\_с»;  $\rho_{kl}$  – коэффициент корреляции Пирсона между целевыми характеристиками  $p_k$  и  $p_l$ ;  $\sigma_k$  и  $\sigma_l$  – стандартные отклонения соответствующих целевых характеристик.

Полученные значения используются для формирования начальной взвешенной структуры графа знаний, которая затем уточняется в процессе обучения графовой нейронной сети.

Примеры триплетов представлены ниже.

```
имеет_признак:  
(Пользователь_123, имеет_признак, Никнейм_литературный_персонаж)  
ассоциируется_с:  
(Никнейм_литературный_персонаж, ассоциируется_с, Радикализм)  
коррелирует_с:  
(Экстраверсия, коррелирует_с, Нейротизм)
```

На основе указанных типов отношений формируется граф знаний, объединяющий объект анализа, признаки цифрового следа и целевые характеристики личности в единую семантическую структуру. Полученный граф используется для последующего нейросимвольного вывода и построения интерпретируемых объяснений.

#### 3.6.4 Пример формирования графа знаний для пользователя

Процесс формирования графа знаний для конкретного пользователя соответствует общей процедуре, описанной в главе 2, и включает последовательное

извлечение признаков из текстовой, визуальной и табличной модальностей с последующим отображением полученных признаков в структуру графа знаний. Признаки модальностей проецируются в общее латентное пространство:

$$\begin{aligned} X(o) &= \{X_m\}_{m \in \mathcal{M}}, \mathcal{M} = \{T, V, D\}, \\ e_m &= f_m(X_m), \\ h_m &= W_m e_m + b_m, \\ h &= \Phi(\{h_m\}_{m \in \mathcal{M}}). \end{aligned}$$

где  $X_m$  – данные  $m$ -й модальности;  $f_m(\cdot)$  – предобученный энкодер соответствующей модальности, используемый в режиме фиксированных параметров;  $e_m$  – вектор признакового представления модальности;  $W_m$  и  $b_m$  – обучаемые параметры линейной проекции в общее латентное пространство;  $h_m \in \mathbb{R}^{512}$  – проекция признакового представления  $m$ -й модальности;  $\Phi(\cdot)$  – функция мультимодального слияния;  $h$  – итоговое мультимодальное представление объекта анализа.

$$h = \lambda_T h_T + \lambda_V h_V + \lambda_D h_D,$$

где  $\lambda_T, \lambda_V, \lambda_D$  – обучаемые коэффициенты значимости модальностей, удовлетворяющие условиям  $\lambda_T + \lambda_V + \lambda_D = 1$  и  $\lambda_i \geq 0$ .

Рассмотрим пример пользователя, для которого в результате анализа цифрового следа были выявлены следующие признаки:

- никнейм, отнесённый к категории «литературный персонаж»;
- статус с позитивной эмоциональной окраской;
- фотография профиля романтического содержания;
- высокий уровень публикационной активности.

На первом этапе для пользователя создаётся узел объекта анализа и узлы соответствующих признаков, связанные отношением «имеет\_признак»:

```
(Пользователь_001, имеет_признак, Никнейм_литературный_персонаж)
(Пользователь_001, имеет_признак, Статус_позитивная_тональность)
(Пользователь_001, имеет_признак, Фото_романтическое)
(Пользователь_001, имеет_признак, Активность_высокая)
```

Далее на основе графа знаний устанавливаются связи между признаками и целевыми характеристиками:

```
(Никнейм_литературный_персонаж, ассоциируется_с, Радикализм)
(Статус_позитивная_тональность, ассоциируется_с, Беспечность)
(Фото_романтическое, ассоциируется_с, Общительность)
(Активность_высокая, ассоциируется_с, Экстраверсия)
```

Кроме того, между факторами личности присутствуют связи типа «коррелирует\_с», отражающие статистически значимые зависимости между психологическими характеристиками.

Полученная структура графа передаётся графовой сети внимания, которая использует как признаки цифрового следа, так и систему связей между признаками и психологическими конструктами для формирования предсказаний. Одновременно коэффициенты внимания используются для определения наиболее значимых путей, объясняющих результат анализа.

Пример фрагмента графа знаний, построенного на данных пользователей социальной сети «ВКонтакте», представлен на рисунке 9.

На рисунке представлены узлы объекта анализа, признаков цифрового следа и психологических конструктов, связанные семантическими отношениями различного типа.

Таким образом, граф знаний выступает связующим звеном между мультимодальными данными и целевыми характеристиками личности, обеспечивая их совместное использование в процессе нейросимвольного вывода и построения интерпретируемых объяснений.

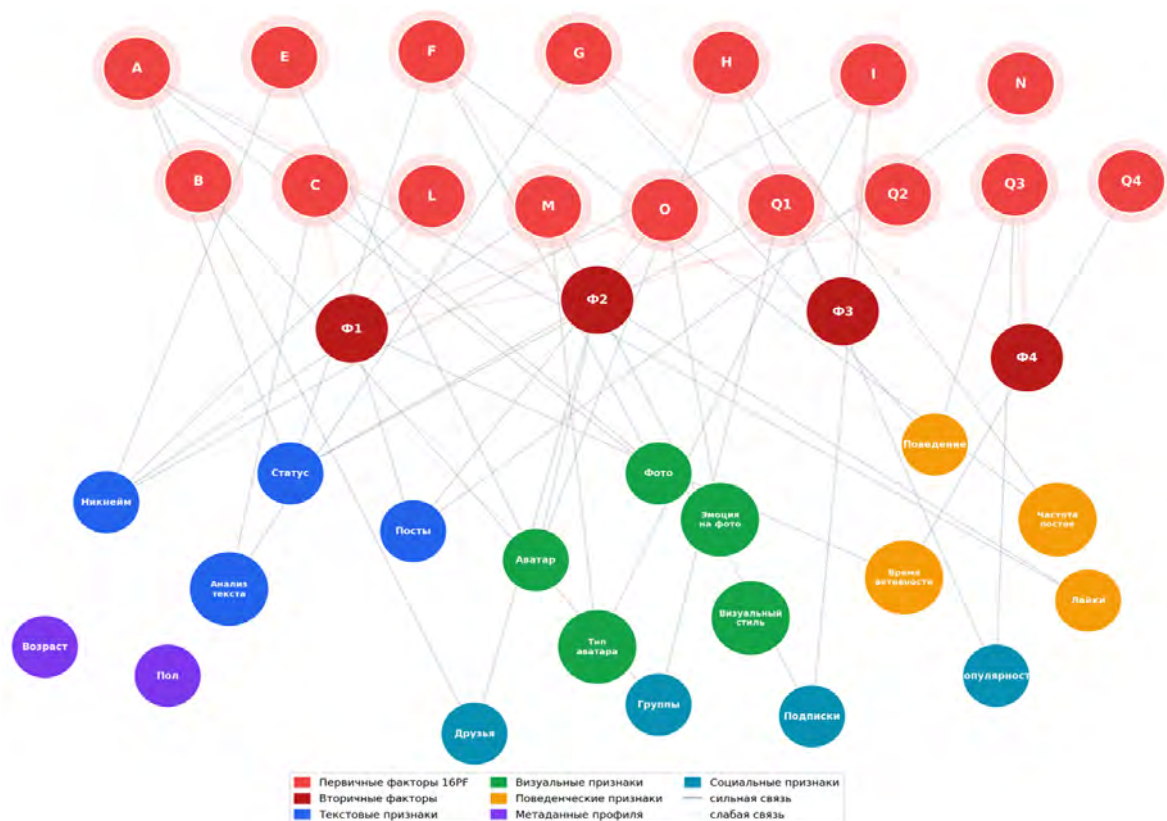


Рисунок 9 – Пример фрагмента графа знаний, построенного на данных пользователей социальной сети «ВКонтакте»



### 3.7 Состав и источники знаний графа

Знания, формирующие граф, имеют два источника. Символьная структура графа (типы признаков и предикаты онтологии) задаётся экспертами априори и определяет, что может быть связано; веса связей определяются статистически по обучающей выборке и определяют, что действительно связано и насколько сильно. В этом состоит нейросимвольная природа метода. Сводная информация об источниках знаний для каждой модальности представлена в таблице 11.

Таблица 11 – Состав и источники знаний графа по модальностям

| Модальность               | Признаки<br>(извлекает нейросеть)                                      | Категории/правила<br>(эксперты, априори)                                                                                                                            | Связи<br>(из обучающей<br>выборки)                                                         |
|---------------------------|------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Текстовая                 | эмбединг<br>RoBERTa/RuBERT $\in \mathbb{R}^{768}$ ;<br>тематика постов | типология никнеймов: 18<br>категорий (6 смысловых, 8<br>словообразовательных, 4<br>графические, на базе 19 690<br>никнеймов); тематика<br>статусов: 10 тематических | значимые рёбра<br>«ассоциируется_с»:<br>никнейм/тема –<br>фактор ( $r_{ik}, p < \alpha'$ ) |
| Визуальная                | эмбединг ResNet50 $\in \mathbb{R}^{2048}$                              | тип визуального контента: 6<br>категорий, 10 категорий<br>животных, демография,<br>эмоциональная экспрессия                                                         | значимые рёбра:<br>тип аватара –<br>фактор ( $r_{ik}, p < \alpha'$ )                       |
| Табличная                 | 39 признаков цифровой<br>активности                                    | 5 типов темпоральной<br>активности                                                                                                                                  | значимые рёбра:<br>активность –<br>фактор ( $r_{ik}, p < \alpha'$ )                        |
| Между<br>характеристиками | –                                                                      | предикат «коррелирует_с»<br>(психологическая теория)                                                                                                                | $\rho_{kl}$ – корреляции<br>факторов из<br>психометрической<br>выборки                     |

Символьная структура графа включает около 49 экспертных категорий и 3 предиката онтологии («имеет\_признак», «ассоциируется\_с», «коррелирует\_с»); веса рёбер определяются статистически значимыми связями после поправки Бонферрони. Таким образом, эксперт задаёт структуру графа, а данные – силу связей.

### 3.8 Интерпретируемость результатов прикладного анализа

Одним из основных требований рассматриваемой прикладной задачи является возможность объяснения получаемых предсказаний. Для практического применения результатов анализа существенное значение имеет возможность установить, какие признаки цифрового следа и в какой степени повлияли на формирование прогнозируемых личностных характеристик.

Механизм построения интерпретируемых объяснений был формализован в главе 2 и основан на извлечении объясняющих путей из графа знаний. В рамках рассматриваемой предметной области такие пути связывают признаки цифрового

следа пользователя с соответствующими психологическими характеристиками через систему семантических отношений графа.

Качество формируемых объяснений оценивается метрикой интерпретируемости путей GPI (Graph Path Interpretability). В настоящем параграфе приведены две агрегированные компоненты метрики; её полное описание представлено в приложении П.

Покрывание статистически значимыми объясняющими путями определяется следующим образом:

$$GPI_{cov} = \frac{1}{U} \sum_{u=1}^U \mathbf{1} (P_u^* \neq \emptyset),$$

где  $P_u^*$  – множество статистически значимых объясняющих путей для объекта (решения)  $u$ ;  $U$  – число анализируемых объектов (решений);  $\mathbf{1}(\cdot)$  – индикаторная функция, принимающая значение 1 при наличии хотя бы одного статистически значимого объясняющего пути и 0 – в противном случае.

Взвешенная согласованность объясняющих путей определяется как:

$$GPI_w = \frac{\sum_{\pi \in \Pi_{sig}} w(\pi)}{\sum_{\pi \in \Pi} w(\pi)},$$

где вес пути вычисляется по формуле:

$$w(\pi) = \prod_{e_{ij} \in \pi} \alpha_{ij}.$$

где  $\Pi$  – множество всех объясняющих путей для рассматриваемого объекта;  $\Pi_{sig} \subseteq \Pi$  – множество путей, содержащих только статистически значимые рёбра, определённые согласно условию (3.6);  $w(\pi)$  – вес пути;  $\alpha_{ij}$  – коэффициент внимания графовой нейронной сети GAT для ребра  $e_{ij}$ . Значения  $GPI_{cov}$  и  $GPI_w$  принадлежат интервалу  $[0, 1]$ .

Представленные показатели используются для количественной оценки интерпретируемости результатов анализа и позволяют сопоставлять качество формируемых объяснений при сравнении различных моделей.

### 3.8.1 Механизм формирования объяснения

Одним из требований рассматриваемой прикладной задачи является возможность интерпретации получаемых предсказаний. Для практического применения результатов анализа важно установить, какие признаки цифрового следа и в какой степени повлияли на формирование прогнозируемых личностных характеристик.

Объяснение предсказания формируется на основе объясняющих путей графа знаний, отражающих вклад отдельных признаков цифрового следа в прогнозируемые характеристики личности. В процессе нейросимвольного вывода графовая нейронная сеть определяет наиболее значимые связи между признаками и целевыми характеристиками. На основе коэффициентов внимания выделяются пути, оказывающие наибольшее влияние на формирование предсказания.

Для каждого предсказания анализируются связи между признаками цифрового следа пользователя и соответствующей целевой характеристикой. В результате предсказание сопровождается объяснением, показывающим наиболее значимые признаки цифрового следа и их вклад в формирование итоговой оценки.

Важным преимуществом графового представления является возможность учитывать как прямые, так и косвенные зависимости между признаками и психологическими характеристиками. Например, признак, редко встречающийся в обучающей выборке, может быть связан в графе знаний с более общими признаками и психологическими конструктами, что позволяет учитывать его влияние при формировании предсказания.

Объяснение строится непосредственно на структуре графа знаний и результатах нейросимвольного вывода. Граф знаний обеспечивает интерпретацию предсказаний через явное представление связей между признаками цифрового следа и прогнозируемыми характеристиками личности.

Пример формирования объясняющего подграфа и интерпретации результата для конкретного пользователя приведён в подпараграфе 3.8.2.

### 3.8.2 Пример объясняющего подграфа

Для иллюстрации работы предложенного механизма интерпретации рассмотрим пример пользователя, для которого модель предсказала высокое значение фактора М (Абстрактность). На основе коэффициентов внимания графовой нейронной сети были выделены наиболее значимые объясняющие пути, связывающие признаки цифрового следа пользователя с прогнозируемой характеристикой личности. Результаты извлечения наиболее значимых путей представлены в таблице 12.

Таблица 12 – Наиболее значимые объясняющие пути для предсказания фактора «Абстрактность»

| Ранг | Объясняющий путь                                                | Вес пути |
|------|-----------------------------------------------------------------|----------|
| 1    | Литературный никнейм → Абстрактность                            | 0,35     |
| 2    | Книги на аватаре → Литературный никнейм → Абстрактность         | 0,24     |
| 3    | Литературный никнейм → Открытость → Абстрактность               | 0,15     |
| 4    | Утренняя активность → Нормативность → Абстрактность             | 0,14     |
| 5    | Нейтральный статус → Эмоциональная устойчивость → Абстрактность | 0,02     |

Анализ выделенных путей показывает, что наибольший вклад в формирование предсказания фактора М вносят признаки, связанные с литературным типом никнейма пользователя. Дополнительное влияние оказывают особенности визуальной самопрезентации и характеристики цифровой активности. Наиболее значимые пути связаны с признаками, отражающими интерес к литературе, интеллектуальную направленность и особенности организации цифровой активности, что согласуется с содержательной интерпретацией фактора М (Абстрактность).

На основе выделенных путей формируется интерпретация результата, связывающая прогнозируемую характеристику личности с наиболее значимыми признаками цифрового следа пользователя. В рассматриваемом примере высокая оценка фактора «Абстрактность» связана с литературным типом никнейма пользователя, особенностями визуального контента профиля и дополнительными связями через психологические конструкторы, ассоциированные с открытостью новому опыту и интеллектуальной направленностью.

Структура объясняющего подграфа для рассматриваемого пользователя представлена на рисунке 10.

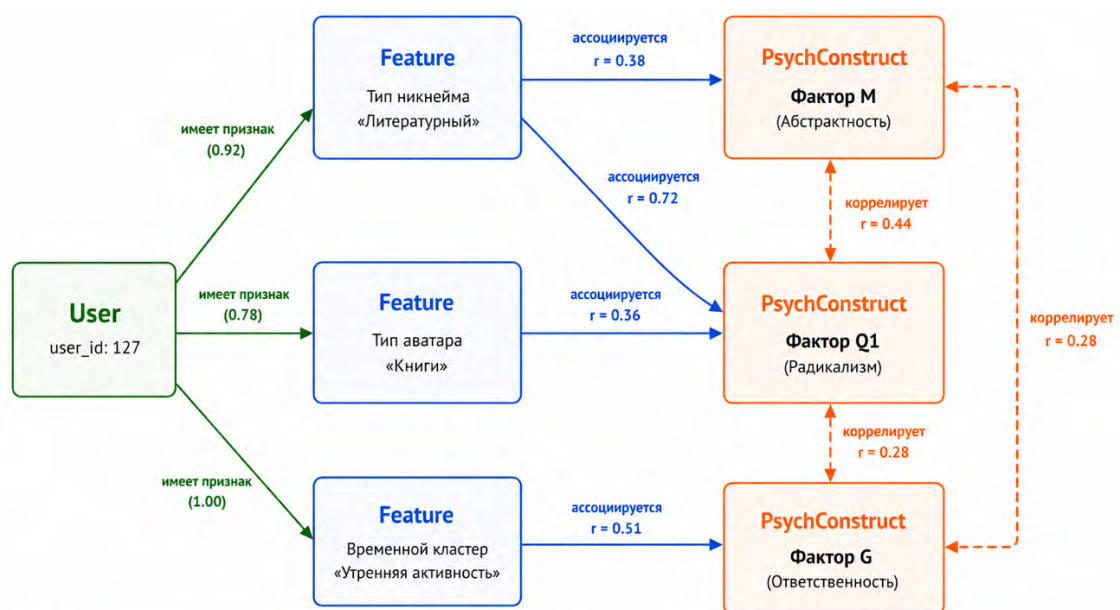


Рисунок 10 – Пример объясняющего подграфа для предсказания фактора «Абстрактность»

На рисунке показаны признаки цифрового следа пользователя, промежуточные психологические конструкторы и целевая характеристика, связанные объясняющими путями. Подписи на рёбрах отражают веса ассоциаций и корреляций, используемые при формировании объясняющего подграфа.

Представленный подграф показывает, каким образом признаки различных модальностей объединяются в графе знаний и участвуют в формировании интерпретируемого предсказания.

### 3.8.3 Вклад модальностей и ограничения интерпретации

Рассмотренный пример показывает, что формирование объяснения опирается на признаки всех трёх модальностей: текстовой (тип никнейма, характеристики статуса), визуальной (тип аватара, содержание изображения) и табличной (характеристики цифровой активности, включая темпоральные паттерны поведения). Совместное использование текстовой, визуальной и табличной модальностей позволяет учитывать различные аспекты цифрового поведения пользователя и выявлять взаимно подтверждающие признаки. Так, литературный тип никнейма и визуальные признаки, связанные с интересом к чтению, согласованно указывают на близкие психологические характеристики, усиливая итоговое предсказание.

Вместе с тем информативность различных признаков и модальностей может существенно различаться. Для части пользователей цифровой след содержит ограниченное количество признаков либо признаки, связанные с различными психологическими характеристиками. В подобных случаях объясняющие пути имеют меньшие веса, а предсказания оказываются ближе к средним значениям по выборке, отражая более высокий уровень неопределённости.

Данное обстоятельство представляет собой одно из ограничений предложенного подхода. Не все личностные характеристики имеют устойчивые и однозначно интерпретируемые проявления в цифровом следе пользователя. В таких ситуациях механизм интерпретации не формирует искусственно убедительных объяснений, а отражает фактическую степень обоснованности полученного результата через структуру объясняющего подграфа и веса связей между его элементами.

Таким образом, интерпретация предсказаний основывается на явных семантических связях между признаками цифрового следа и прогнозируемыми характеристиками личности, представленными в графе знаний. Объясняющий подграф отображает процесс формирования итоговой оценки и вклад отдельных признаков и модальностей.

## 3.9 Выводы по главе 3

Основным результатом настоящей главы, полученным лично автором, является применение предложенного доменно-инвариантного метода к задаче анализа личности по цифровому следу и его программная реализация; приводимые далее конкретные предметные понятия (поля данных, категории признаков, элементы онтологии) являются деталями этого применения, а не самостоятельным результатом.

1. Выполнена предметная реализация разработанного нейросимвольного метода для задачи анализа личности по цифровому следу пользователей социальной сети ВКонтакте. На едином сквозном примере рассмотрены все этапы предложенного подхода: от извлечения признаков из мультимодальных данных до формирования интерпретируемых объяснений результатов анализа. Полученные результаты подтверждают возможность переноса метода на другие предметные области за счёт использования доменно-инвариантного онтологического паттерна. Сформирована мультимодальная база данных, объединяющая психометрические профили пользователей и данные их цифровой активности, а также выполнен анализ ограничений сформированной выборки и особенностей её использования в задачах интеллектуального анализа данных.

2. Выполнено единообразное преобразование текстовой, визуальной и табличной модальностей цифрового следа в категориальные и числовые признаки, используемые при построении графа знаний. Извлечённые признаки отражают различные аспекты цифровой самопрезентации и поведения пользователя и служат основой для последующего анализа.

3. Для текстовой модальности сформировано категориальное описание никнеймов и статусов пользователей, учитывающее семантические, функциональные и тематические характеристики текстового контента и используемое в качестве источника признаков для графа знаний.

4. Для табличной модальности сформировано признаковое описание цифровой активности, включающее статистики активности, показатели вовлечённости, темпоральные паттерны и метаданные профиля, используемые при построении графа знаний.

5. Выполнена предметная конкретизация онтологии и графа знаний для задачи анализа цифрового следа: определены объект анализа, признаки цифрового следа, целевые характеристики и типы отношений между ними. Формирование графа знаний пользователя выполнено в соответствии с общим подходом к построению графов знаний, разработанным в главе 2.

6. Показано, что результаты анализа могут быть представлены в интерпретируемой форме на основе графа знаний. Объясняющий подграф отражает взаимосвязи между признаками цифрового следа, психологическими конструктами и прогнозируемыми характеристиками личности.

## 4 Архитектура системы

### 4.1 Общая архитектура клиент-серверной системы

Система реализована на основе трёхуровневой клиент-серверной архитектуры, включающей уровень представления (веб-интерфейс), уровень приложения (обработка данных и бизнес-логика) и уровень данных (хранилища и модели). Этапы извлечения признаков, формирования графа знаний и прогнозирования относятся к уровню приложения; визуализация результатов и представление объяснений – к уровню представления. Уровень данных обеспечивает хранение графа знаний, метаданных пользователей и промежуточных результатов.

Общая схема архитектуры представлена на рисунке 11. В качестве хранилища знаний используется графовая база данных Neo4j, содержащая 63 161 узел и 777 365 рёбер. Извлечение признаков выполняется моделями обработки текста, изображений и данных цифровой активности; прогнозирование – графовой нейронной сетью внимания GAT приблизительно с 3,5 млн обучаемых параметров.

Коммуникация между уровнями организована по протоколу HTTPS (TLS 1.3) с JWT-аутентификацией для взаимодействия клиента и сервера; сервер взаимодействует с хранилищами данных через соответствующие драйверы (Neo4j – Bolt, PostgreSQL – psycopg3, Redis – redis-py), а модели машинного обучения загружаются в оперативную память при запуске системы. Технологический стек системы представлен в таблице 13.

Таблица 13 – Технологический стек системы

| Уровень               | Компонент           | Технология         | Назначение                              |
|-----------------------|---------------------|--------------------|-----------------------------------------|
| Уровень представления | Веб-интерфейс       | React              | Пользовательский интерфейс              |
|                       | Визуализация данных | D3.js, Recharts    | Отображение графов знаний и результатов |
| Уровень приложения    | API-сервер          | FastAPI            | REST API и обработка запросов           |
|                       | Очередь задач       | Celery             | Асинхронная обработка                   |
|                       | Брокер/кэш          | Redis              | Очереди и промежуточные результаты      |
| Модели ML             | Текст               | RoBERTa (PyTorch)  | Извлечение текстовых признаков          |
|                       | Визуал              | ResNet50 (PyTorch) | Извлечение визуальных признаков         |
|                       | Табличная           | XGBoost            | Анализ цифровой активности              |
|                       | Графовая            | GAT (DGL, PyTorch) | Прогнозирование на графе знаний         |
| Уровень данных        | Граф знаний         | Neo4j              | Хранение онтологии и графа              |
|                       | Реляционная БД      | PostgreSQL         | Метаданные и результаты                 |
|                       | Кэш                 | Redis              | Распределённая обработка                |

Выбранная трёхуровневая архитектура разделяет ответственность уровней: представление отвечает за интерфейс, приложение – за вычисления и координацию, данные – за хранение. Это упрощает независимое развитие и адаптацию системы к

новым задачам без перестройки ядра. Интеграция мультимодальных данных, графа знаний и GAT даёт высокую точность прогнозов и интерпретируемые объяснения через пути в графе знаний.

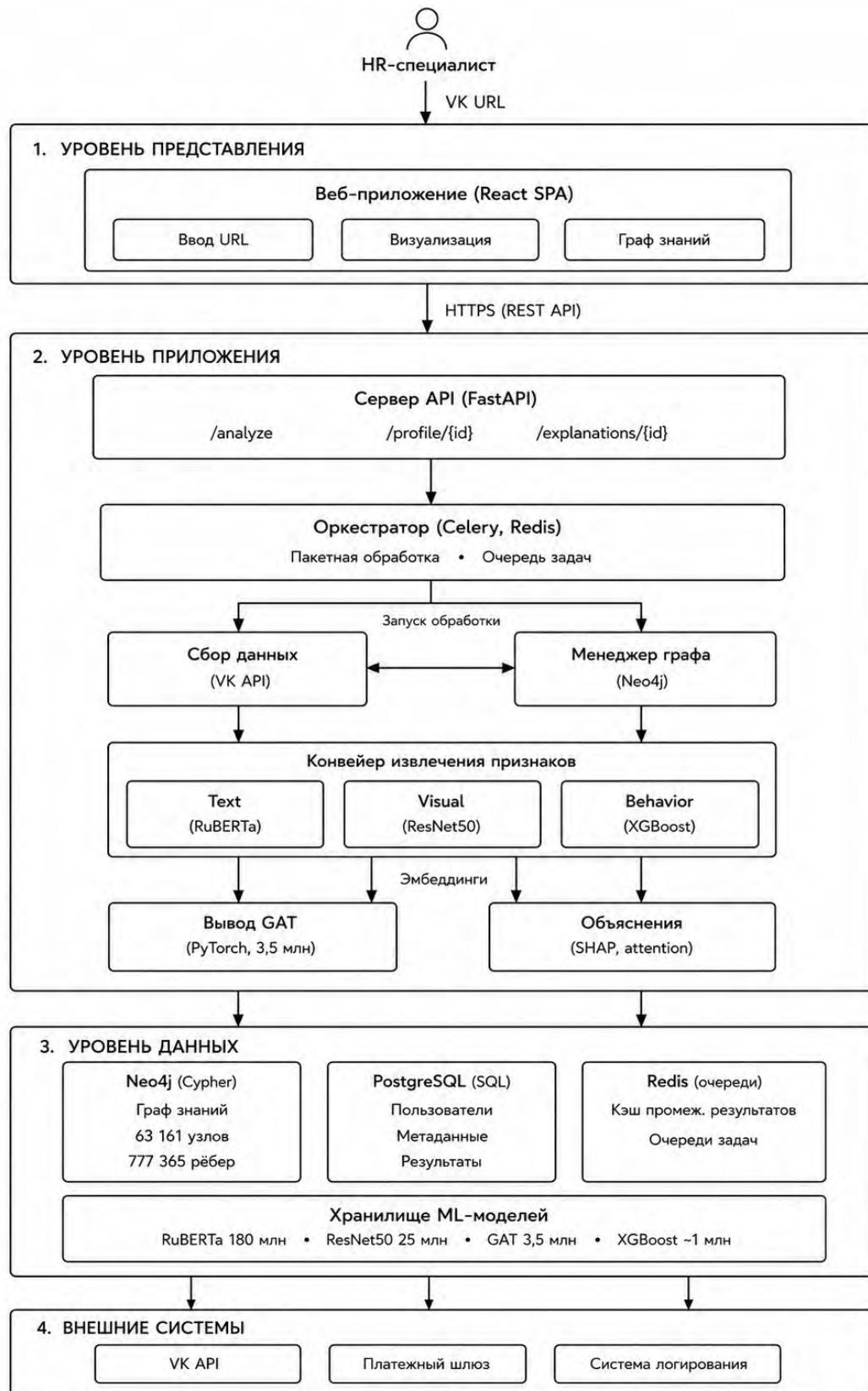


Рисунок 11 – Трёхуровневая клиент-серверная архитектура системы



#### 4.1.1 Подсистемы и их функции

Система включает шесть подсистем, соответствующих этапам обработки данных.

1. Модуль сбора данных – интеграция с VK API: параллельный сбор пяти типов данных (информация о пользователе, посты, фотографии, друзья, группы), обработка исключений (закрытые/удалённые профили, rate limit 20 запросов/с). Реализация – VKDataCollector (приложение А, листинг А.1).

2. Подсистема извлечения признаков – три экстрактора на базе моделей из главы 3:

- текстовый (RoBERTa, RuBERT): никнеймы, статусы, посты – категория никнейма (6 классов), тематика статусов (10 категорий), тип поведения (4 категории), эмбединг 768;

- визуальный (ResNet50): аватар – возраст и тип аватара (6 классов), VGG16 – определение пола, эмбединг 2048; предобработка: детекция лица YOLOv5x, нормализация;

- активности (XGBoost, K-Means): метаданные публикаций – тип поведения (4), временной (5) и сетевой (4) кластеры.

Реализации – TextFeatureExtractor, VisualFeatureExtractor, BehavioralFeatureExtractor (приложение А, листинги А.3–А.5). Модули зарегистрированы как программа для ЭВМ [247\*].

3. Менеджер графа знаний (Neo4j) – наполнение графа признаками и выполнение Cypher-запросов [214]: создание узлов User и Feature, рёбер HAS\_FEATURE, сохранение эмбедингов (768 для текста, 2048 для визуала и активности), извлечение подграфов радиуса k (по умолчанию k = 2). Реализация – KnowledgeGraphManager (приложение А, листинг А.6). Схема графа – на рисунке 10 (полное определение – приложение М). Граф включает узлы User, Feature, PsychConstruct и связи HAS\_FEATURE, ASSOCIATED\_WITH, CORRELATES\_WITH.

4. Модуль графового вывода (GAT) – предсказание 26 факторов личности: трёхслойная GAT (8, 8 и 4 головы внимания, остаточные связи), слой Linear (512,26), сохранение весов внимания. Процедура графового вывода включает последовательное выполнение этапов: извлечение из графа знаний подграфа радиуса 2 (2-hop), преобразование полученной структуры в формат библиотеки DGL с отображением идентификаторов узлов Neo4j на внутренние индексы графовой библиотеки, инициализация векторных представлений узлов (для узла User – конкатенация

извлечённых мультимодальных признаков; для узлов Feature и PsychConstruct – обучаемые векторные представления), прямой проход по графовой нейронной сети с сохранением коэффициентов внимания и формирование предсказания для целевого узла User. Программные реализации классов GATModel (на базе PyTorch с использованием слоёв GATConv) и GATInferenceEngine представлены в приложении А (листинг А.7). Модуль зарегистрирован как программа для ЭВМ [248\*].

5. Генератор объяснений – формирование интерпретаций на основе графа знаний, весов внимания и вклада признаков: извлечение объясняющих путей, оценка вклада модальностей, генерация текстовых объяснений. Формирует локальные объяснения (для конкретного пользователя) и структурные объяснения (процесс принятия решения моделью).

6. Веб-интерфейс – визуализация результатов для конечного пользователя (подробно – параграф 4.4): ввод профиля, отображение предсказанных характеристик, вклад модальностей, текстовые объяснения, интерактивная визуализация графа знаний и объясняющих подграфов.

#### 4.1.2 Интеграция компонентов и последовательность обработки запроса

Последовательность взаимодействия компонентов при обработке запроса показана на рисунке 12. Пользователь вводит ссылку на профиль через веб-интерфейс; сервер создаёт асинхронную задачу и передаёт её в оркестратор. Оркестратор запускает конвейер: сбор данных через VK API, извлечение признаков обученными моделями, сохранение результатов в граф знаний, извлечение подграфа и прогнозирование с использованием GAT. На основе графового вывода формируются интерпретируемые объяснения с учётом структуры графа и коэффициентов внимания.

Результаты сохраняются в базе данных; клиент отслеживает статус задачи и после завершения получает результаты и объяснения для визуализации. Таким образом, обработка запроса реализована как последовательный конвейер: сбор данных – извлечение признаков – построение графа – прогнозирование – интерпретация.

Представленная архитектура подтверждает практическую реализацию нейросимвольного метода и возможность совместного использования мультимодальных данных, графа знаний и GAT для интерпретируемого анализа.

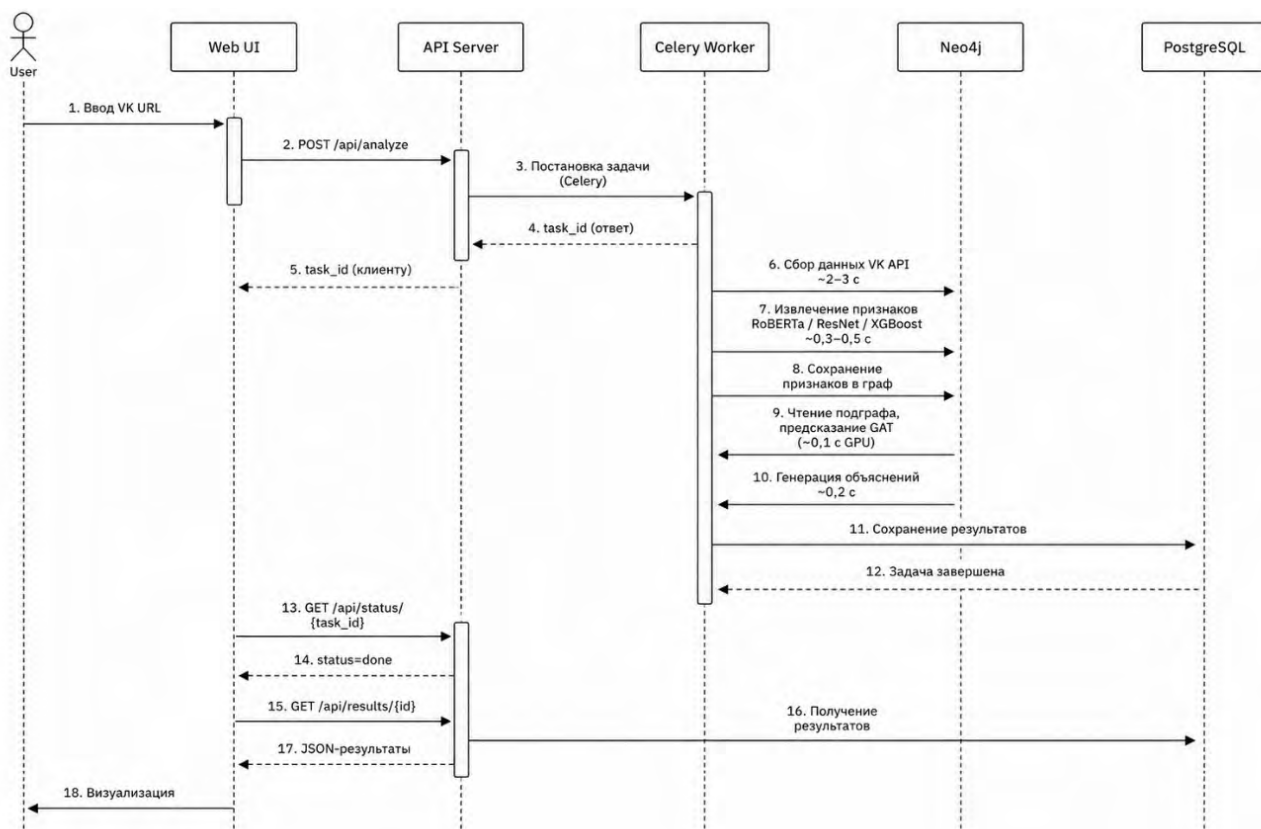


Рисунок 12 – Диаграмма последовательности обработки запроса

## 4.2 Технологическая реализация графа знаний

Граф знаний служит центральным компонентом нейросимвольной архитектуры, интегрирующим мультимодальные данные, предметные знания и графовые методы машинного обучения. В параграфе рассмотрены архитектура графа, выбор технологической платформы, схема данных и процесс наполнения графа. Нейросимвольная интеграция GAT и графа знаний предложена автором в [55\*].

Граф знаний специфицирован в терминах разработанной онтологии и реализован в Neo4j. Экземплярный граф охватывает 19 690 пользователей очищенной выборки (в том числе 1 722 с психометрическими метками) и содержит 63 161 узел и 777 365 рёбер. Хранилище интегрируется с PyTorch Geometric и DGL для обучения GNN.

В ходе проектирования проведён сравнительный анализ графовых СУБД. Выбор Neo4j обусловлен поддержкой модели property graph, развитыми средствами визуализации и интеграцией с графовыми нейронными сетями, что соответствует требованиям нейросимвольной архитектуры.

Результаты сравнительного анализа представлены в таблице 14.

Таблица 14 – Сравнение графовых баз данных для хранилища графа знаний

| Критерий              | Neo4j               | RDF-хранилища       | Amazon Neptune |
|-----------------------|---------------------|---------------------|----------------|
| Язык запросов         | Cypher              | SPARQL              | Gremlin/SPARQL |
| Интеграция с GNN      | Высокая             | Ограниченная        | Ограниченная   |
| Поддержка онтологий   | Через neosemantics  | Нативная            | Через SPARQL   |
| Визуализация и анализ | Встроенные средства | Внешние инструменты | Ограниченная   |

#### 4.2.1 Схема узлов и рёбер

Структура графа знаний реализует онтологическую модель (глава 3) и включает три типа узлов: пользователь (User), признак (Feature) и психологический конструкт (PsychologicalConstruct).

Узлы User соответствуют виртуальной личности пользователя и содержат характеристики профиля из VK API: демографические данные, показатели активности, метаданные (всего 28 атрибутов). Узлы Feature представляют автоматически извлечённые признаки трёх модальностей: категории никнеймов, тематику статусов, тип аватара, эмоции, поведенческие и темпоральные паттерны (всего 13 признаков). Атрибуты: наименование, значение и модальность (text, visual, behavioral). Узлы PsychologicalConstruct соответствуют целевым характеристикам: 20 факторов 16PF, 6 интегральных факторов личности и дополнительные социально-психологические показатели.

Между узлами заданы три типа отношений: HAS\_FEATURE (User → Feature) – наличие признака у пользователя; ASSOCIATED\_WITH (Feature → PsychologicalConstruct) – статистически значимая связь между признаком и конструктом; CORRELATES\_WITH (PsychologicalConstruct ↔ PsychologicalConstruct) – взаимосвязи между конструктами.

Структура графа описывается как:

$$G = (N, E, w),$$

где:

- $N = N_{User} \cup N_{Feature} \cup N_{Construct}$ ;
- $E = E_{HAS\_FEATURE} \cup E_{ASSOCIATED\_WITH} \cup E_{CORRELATES\_WITH}$ ;
- $w: E \rightarrow \mathbb{R}$  – веса рёбер.

Таким образом, граф связывает цифровые признаки с психологическими конструктами через семантически интерпретируемые отношения и служит основой для вывода и объяснений. Для производительности запросов в Neo4j созданы индексы по ключевым атрибутам (идентификатор пользователя, модальность, код конструкта); полные листинги схемы – в приложении А.

#### 4.2.2 Наполнение графа из данных

Наполнение графа знаний выполняется по ETL-схеме (Extract, Transform, Load) на основе данных цифрового следа пользователей ВКонтакте и результатов психометрического тестирования.

Extract – для каждого пользователя собираются: характеристики профиля (возраст, пол, город, образование), текстовые данные (никнейм, статус, публикации), визуальные данные (аватар, фотографии), показатели цифровой активности (частота публикаций, лайки, комментарии, репосты) и характеристики окружения (число друзей, подписчиков, сообществ). Дополнительно загружаются результаты психометрического тестирования.

Transform – на основе трёх модальностей выполняется автоматическое извлечение признаков с использованием моделей из главы 3 (таблица 15). Признаки сохраняются в промежуточном табличном представлении.

Таблица 15 – Примеры автоматически извлекаемых признаков цифрового следа

| Признак                         | Модель                 | Входные данные             | Результат   |
|---------------------------------|------------------------|----------------------------|-------------|
| Тип никнейма                    | RoBERTa                | Никнейм пользователя       | 6 категорий |
| Тональность статуса             | RuBERT                 | Текст статуса              | 3 категории |
| Тематика публикаций             | RuBERT                 | Тексты публикаций          | 4 категории |
| Тип фото профиля                | ResNet50               | Изображение профиля        | 6 категорий |
| Возрастная категория            | ResNet50               | Изображение профиля        | 4 категории |
| Эмоциональное состояние         | CNN                    | Изображение профиля        | 7 категорий |
| Тип активности                  | XGBoost                | Данные цифровой активности | 4 категории |
| Темпоральный паттерн активности | Анализ временных рядов | Временные метки публикаций | 5 категорий |

Load – загрузка в Neo4j включает: создание узлов User и Feature, формирование рёбер HAS\_FEATURE, создание узлов PsychologicalConstruct, загрузку рёбер ASSOCIATED\_WITH (по статистическому анализу) и CORRELATES\_WITH (по выявленным взаимосвязям). После загрузки выполняется проверка целостности графа. Полные листинги запросов – в приложении А (параграф А.6).

На момент экспериментов граф содержал 63 161 узел и 777 365 рёбер, охватывая 19 690 пользователей (включая 1 722 с психометрическими метками). ETL-процесс формализован в алгоритме 5.

Алгоритм 5 – ETL-наполнение графа знаний из данных социальной сети

Вход: профили пользователей VK API; результаты психометрии; ML-модели извлечения признаков; онтология  $\mathcal{O}$ .

Выход: наполненный граф знаний в хранилище Neo4j.

1. Extract: для каждого пользователя  $u$  получить из VK API профиль, тексты, изображения, показатели активности и окружения; загрузить психометрию.

2. Transform:  $c(u) := \text{ML-модели(текст, изображения, табличные данные)}$  – признаки трёх модальностей; сохранить в промежуточном табличном представлении.
3. Load:
  - создать узлы  $\text{User}(u)$ ;
  - создать узлы  $\text{Feature}(f_i)$  и рёбра  $\text{HAS\_FEATURE}(u, f_i)$ ;
  - создать узлы  $\text{PsychologicalConstruct}(p_k)$ ;
  - создать рёбра  $\text{ASSOCIATED\_WITH}(f_i, p_k)$  по статистическому анализу и экспертным знаниям;
  - создать рёбра  $\text{CORRELATES\_WITH}(p_k, p_l)$  по выявленным взаимосвязям.
4. Проверить целостность графа (связность, отсутствие изолированных вершин и дублей).
5. Вернуть наполненный граф.

#### 4.2.3 Количественные параметры наполненного графа

На момент экспериментов граф включал 19 690 пользователей очищенной выборки (из них 1 722 с психометрическими метками) и содержал 63 161 узел трёх типов (User, Feature, PsychologicalConstruct) и 777 365 рёбер трёх типов (HAS\_FEATURE, ASSOCIATED\_WITH, CORRELATES\_WITH). В граф включены только статистически значимые связи между признаками и психологическими конструктами. Сформированное графовое представление используется для обучения GAT и формирования интерпретируемых объяснений.

### 4.3 Алгоритмическое обеспечение конвейера обработки профиля

#### 4.3.1 Конвейер обработки профиля пользователя

Обработка профиля выполняется как последовательный конвейер, объединяющий этапы сбора данных, извлечения признаков, построения графа, прогнозирования и формирования объяснений. Этапы имеют формализованные входы и выходы, что позволяет разрабатывать и тестировать компоненты независимо. Извлечение признаков и графовый вывод поддерживают вычисления на GPU, кэширование промежуточных представлений сокращает время обработки. Последовательность этапов показана на рисунке 13.

Этап 1 – Сбор данных из VK API (2–3 с). Параллельный сбор характеристик профиля, текстов, изображений и показателей активности с контролем скорости запросов и обработкой исключений (закрытые/удалённые профили, rate limit).

Этап 2 – Предобработка данных (0,1–0,2 с). Очистка текстов, подготовка изображений, нормализация метаданных, формирование промежуточного представления (приложение А, листинг А.2).

Этап 3 – Извлечение признаков (0,3–0,5 с). Текстовая модальность (RoBERTa, RuBERT): никнеймы, статусы; визуальная (ResNet50): аватар; табличная:

статистические показатели активности. Формируется мультимодальное представление пользователя.

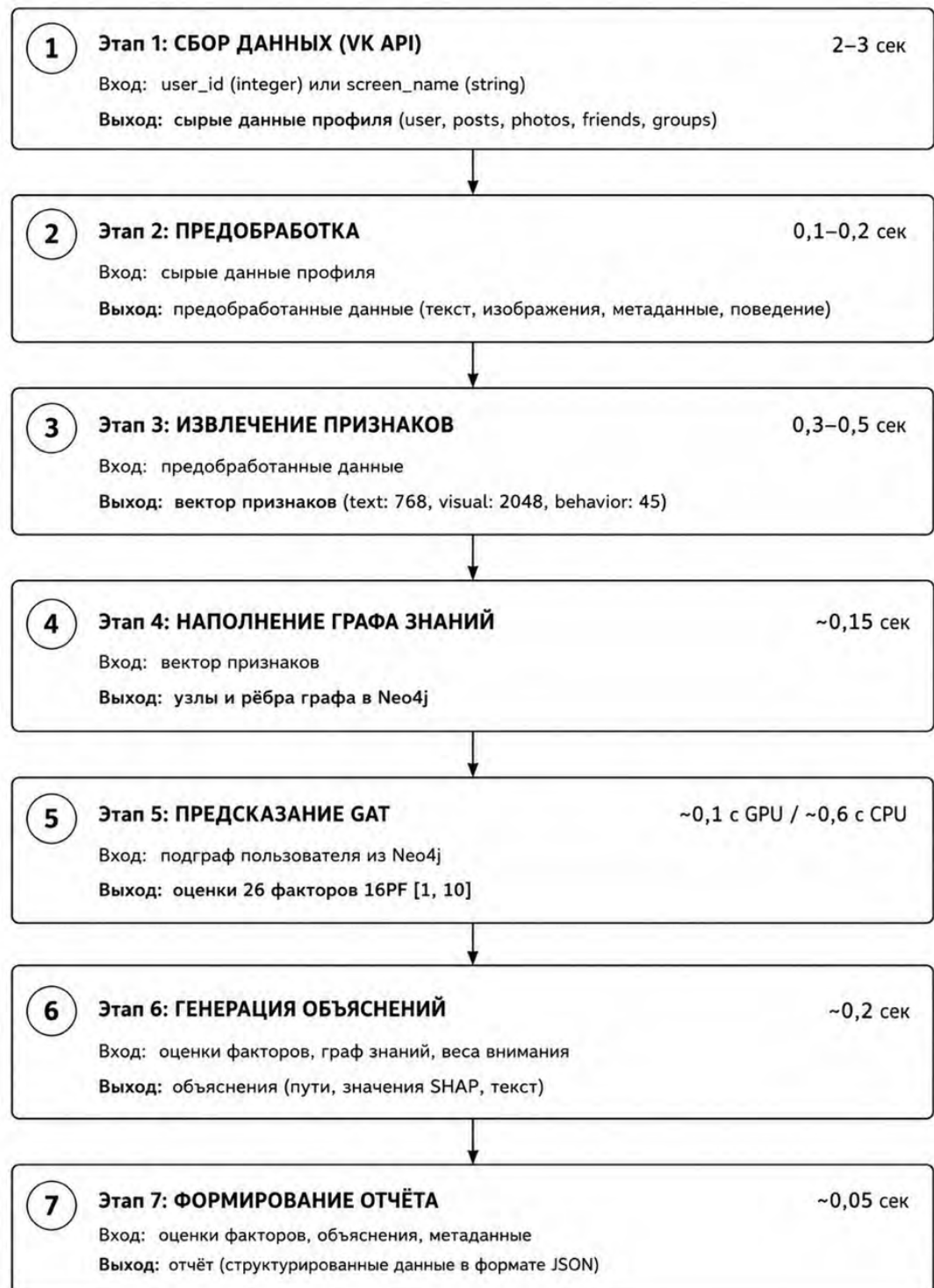


Рисунок 13 – Алгоритмический конвейер обработки профиля

Этап 4 – Наполнение графа Neo4j (~0,15 с). Создание узлов User и Feature, формирование рёбер HAS\_FEATURE, ASSOCIATED\_WITH и CORRELATES\_WITH, сохранение подграфа.

Этап 5 – Прогнозирование GAT ( $\approx 0,1$  с на GPU, 0,6 с на CPU). Извлечение подграфа пользователя, преобразование в формат DGL, инициализация узлов, прямой проход с сохранением внимания, получение прогноза.

Этап 6 – Генерация объяснений ( $\approx 0,2$  с). Извлечение объясняющих путей, вычисление GPI (приложение П), оценка вклада модальностей, формирование подграфа и текстовой интерпретации.

Этап 7 – Формирование отчёта ( $\approx 0,05$  с). Структурированное представление результатов с прогнозами, объяснениями и служебной информацией.

Сквозной конвейер формализован в алгоритме 6.

Алгоритм 6 – Конвейер обработки профиля пользователя в режиме исполнения

Вход: ссылка на профиль; обученные ML-модели; графовая сеть GAT; граф знаний  $G$ .

Выход: структурированный отчёт с прогнозами характеристик и объяснениями.

1.  $data := VK\ API(\text{профиль})$  — профиль, тексты, изображения, активность.
2.  $data := preprocess(data)$  — очистка, нормализация.
3.  $c(u) := extract\_features(data)$  — признаки трёх модальностей обученными моделями.
4. Сформировать фрагмент графа: узел  $User(u)$ , узлы  $Feature$ , рёбра  $HAS\_FEATURE$ .
5.  $G_u :=$  подграф пользователя;  $(Y, \alpha) := GAT(G_u)$  с сохранением внимания.
6.  $(\Pi^*, GPI) := \text{Алгоритм } 4(G_u, Y, \alpha)$ ; определить вклад модальностей.
7. Сформировать отчёт по  $(Y, \Pi^*, GPI, \text{вклад модальностей})$
8. Вернуть отчёт.

#### 4.3.2 Оптимизация производительности

Базовый конвейер обрабатывает один профиль за 3–4 секунды на GPU. Для повышения эффективности реализованы четыре стратегии оптимизации.

1. Кэширование эмбедингов в Redis. Извлечение признаков (RuBERT, ResNet50) занимает около 350 мс; при повторных запросах эмбединги кэшируются на 24 часа. При доле повторных запросов  $>70\%$  время обработки сокращается до  $\approx 1$  с.

2. Параллельная пакетная обработка. Последовательная обработка 100 профилей занимает  $\approx 350$  с. Использование пула из 8 рабочих процессов даёт  $\approx 130$  профилей/мин, из 16 –  $>200$  профилей/мин.

3. Пакетная обработка на GPU. Обработка профилей пакетами повышает загрузку GPU: время извлечения признаков для одного профиля сокращается с  $\approx 300$  мс до 10–15 мс (текст) и аналогично для визуальных данных.

4. Предзагрузка моделей в память. Модели загружаются при запуске сервера, исключая задержки инициализации (5–10 с) при каждом запросе.

В итоге система достигает  $\approx 1$  с на профиль при кэшировании и  $>200$  профилей/мин при пакетной обработке. Реализация – в приложении А.



## 4.4 Пользовательский интерфейс и REST API

### 4.4.1 Веб-интерфейс для HR-специалистов

Для визуализации результатов анализа цифрового следа и их интерпретации разработан веб-интерфейс, ориентированный на специалистов в области HR-аналитики и психологии. Интерфейс предоставляет доступ к прогнозируемым характеристикам личности, объясняющим подграфам графа знаний, а также к сведениям о вкладе отдельных признаков и модальностей в формирование предсказания.

Функционально интерфейс включает:

- ввод идентификатора профиля пользователя и запуск анализа;
- отображение прогнозируемых личностных характеристик (26 факторов 16PF) в виде радиальной диаграммы с цветовым кодированием уровней выраженности;
- визуализацию объясняющих путей в графе знаний, связывающих признаки цифрового следа с прогнозируемыми факторами;
- представление вклада текстовой, визуальной и табличной модальностей в итоговое предсказание;
- пакетный режим для сравнительного анализа множества профилей.

Главный экран системы представлен на рисунке 14.

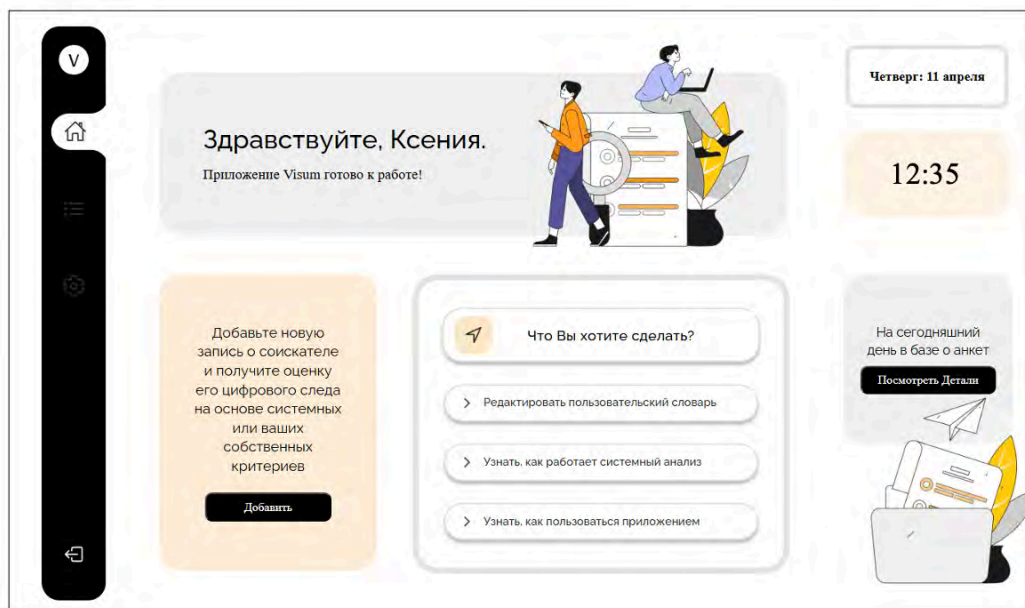


Рисунок 14 – Главный экран системы анализа цифрового профиля

Экран результатов анализа профиля представлен на рисунке 15. После завершения обработки пользователю предоставляется интегрированный отчёт, объединяющий сведения о цифровой активности, прогнозируемые характеристики и их интерпретации.

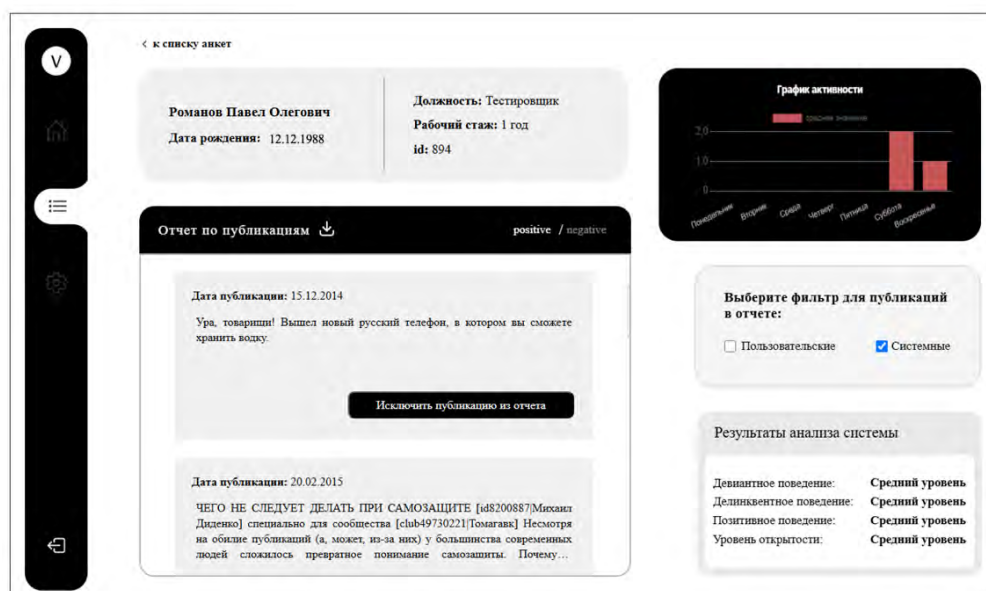


Рисунок 15 – Экран результатов анализа профиля

Визуализация личностного профиля в пространстве факторов 16PF представлена на рисунке 16.

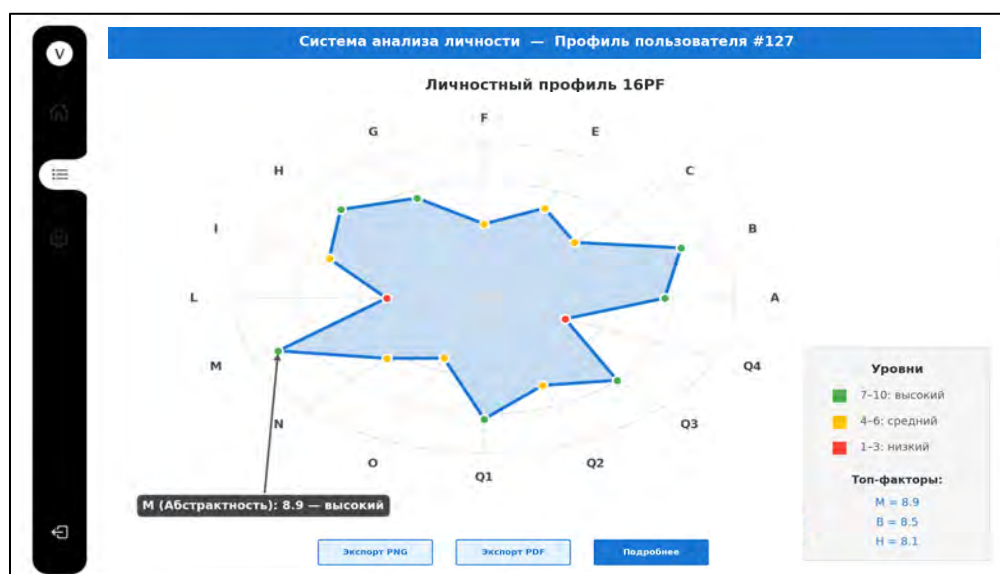


Рисунок 16 – Визуализация личностного профиля в пространстве факторов 16PF

Для отображения прогнозируемых характеристик личности используется радиальная диаграмма, каждая ось которой соответствует одному из 26 факторов 16PF. Цветовое кодирование отражает уровень выраженности фактора (низкий, средний, высокий). Интерактивные элементы диаграммы обеспечивают доступ к численному значению, содержательной интерпретации фактора и сведениям о признаках цифрового следа, внесших наибольший вклад в предсказание. Каждая оценка может быть связана с объясняющим подграфом, отражающим взаимосвязи между признаками, промежуточными психологическими конструктами и прогнозируемой характеристикой.

Визуализация графа знаний пользователя представлена на рисунке 17. Для анализа взаимосвязей между признаками цифрового следа и психологическими характеристиками используется интерактивное представление графа знаний в виде силовой диаграммы. Граф включает узлы пользователя, признаков (Feature) и психологических конструктов (PsychConstruct), а также связи HAS\_FEATURE, ASSOCIATED\_WITH и CORRELATES\_WITH. Вес связей соответствует их статистической значимости и учитывается при формировании объяснений.

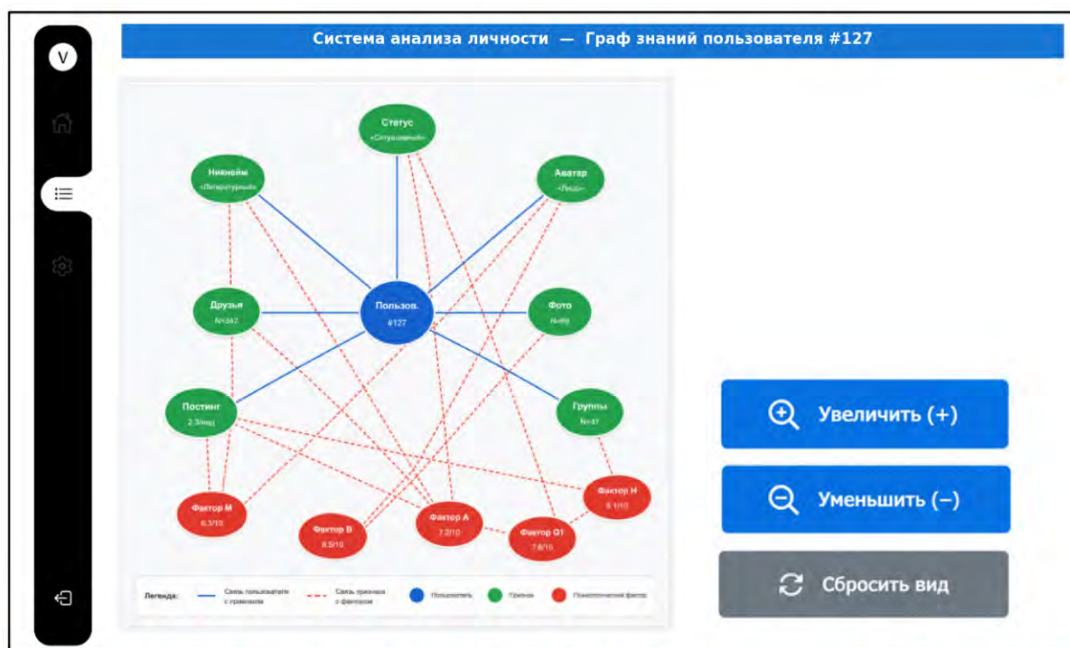


Рисунок 17 – Визуализация графа знаний пользователя  
Интерпретация предсказаний представлена на рисунке 18.

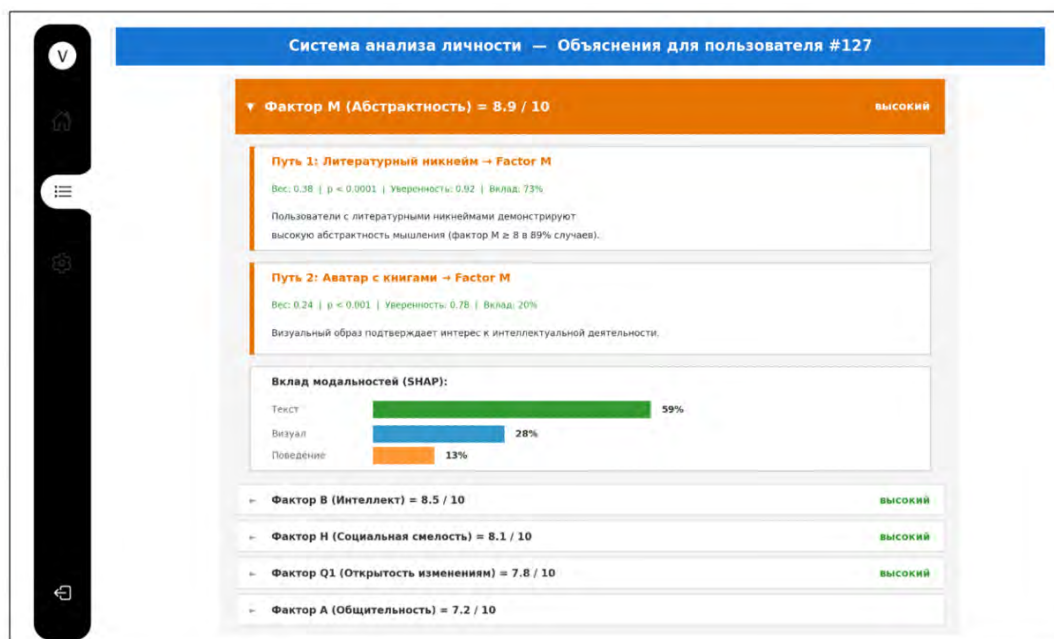


Рисунок 18 – Панель объяснений предсказаний с объясняющими путями и оценкой вклада модальностей

Основным элементом является объясняющий подграф для выбранной характеристики, содержащий наиболее значимые признаки и связи. Предусмотрены механизмы фильтрации по модальностям и уровню значимости.

Панель объяснений объединяет результаты прогнозирования с объясняющими путями и подграфами графа знаний, позволяя установить признаки и связи, оказавшие наибольшее влияние на формирование каждой прогнозируемой характеристики (подробно – в подпараграфе 4.4.4).

Сравнительный анализ цифровых профилей представлен на рисунке 19.

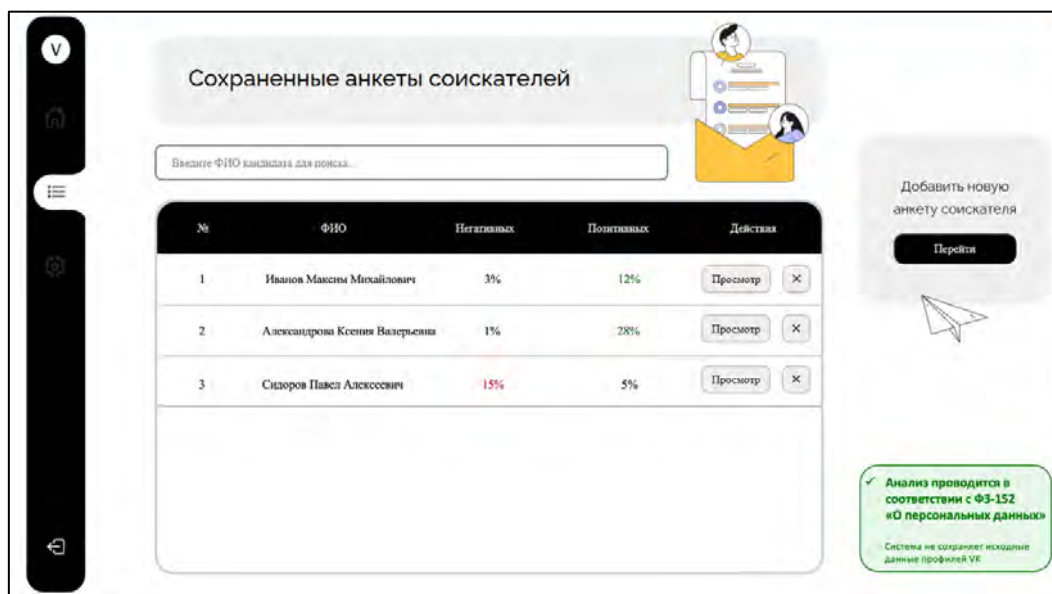


Рисунок 19 – Панель сравнительного анализа цифровых профилей

В пакетном режиме обеспечивается одновременный анализ множества профилей и их сопоставление в едином признаковом пространстве с возможностью сортировки, фильтрации и экспорта результатов.

#### 4.4.2 Программный интерфейс системы

Для интеграции с внешними информационными ресурсами реализован программный интерфейс в виде REST API, предоставляющий доступ к основным этапам обработки данных в асинхронном режиме:

- запуск анализа профиля по идентификатору пользователя или ссылке;
- получение информации о состоянии выполнения задачи;
- получение результатов прогнозирования и объяснений для отдельных характеристик личности;
- пакетная обработка множества профилей с агрегацией результатов.

Возвращаемые данные имеют унифицированную структуру и включают метаинформацию о профиле, прогнозируемые характеристики, объясняющие пути и вклад модальностей, а также служебную информацию о параметрах модели и времени

выполнения вычислений. Единый формат представления упрощает интеграцию с внешними системами.

API поддерживает аутентификацию пользователей, ограничение интенсивности запросов и защиту каналов передачи данных. Программная реализация интерфейса приведена в приложении А. Апробация системы на задаче анализа цифровых профилей, анализ ошибок и сравнение с альтернативными подходами представлены в главе 5.

#### 4.5 Ограничения предложенного подхода и направления дальнейших исследований

##### 4.5.1 Текущие ограничения

Предложенный подход имеет следующие ограничения.

1. Ограниченность выборки. Исследование выполнено на 1722 пользователях одной социальной сети, преимущественно из Самарского региона. Редкие категории признаков представлены ограниченным числом наблюдений, что снижает устойчивость предсказаний для отдельных групп. Результаты требуют дополнительной проверки на данных других социальных платформ и культурных контекстов.

2. Корреляционный характер зависимостей. Выявляемые связи носят статистический, а не причинно-следственный характер. Прогнозы следует интерпретировать как вероятностные оценки, а не прямое измерение личностных качеств пользователя.

3. Неполнота цифрового профиля. Качество анализа зависит от полноты цифрового следа: отсутствие отдельных модальностей снижает информативность входных данных. Кроме того, анализ статичен — профиль рассматривается как фиксированное состояние, без учёта временной динамики поведения пользователя.

4. Этические ограничения. Использование цифрового следа требует соблюдения законодательства о персональных данных. Результаты анализа зависят от особенностей самопрезентации пользователя, что может приводить к расхождениям между наблюдаемыми признаками и реальными характеристиками личности.

##### 4.5.2 Направления дальнейших исследований

Полученные результаты определяют следующие направления дальнейшего развития.

1. Федеративное обучение. В практических условиях доступ к данным ограничен требованиями конфиденциальности. Перспективно применение Federated Learning для обучения графовых нейронных сетей без централизации данных, включая согласование локальных подграфов и сохранение семантической целостности графа знаний.

2. Генерация объяснений на естественном языке. Структурные объяснения (пути и подграфы) могут быть преобразованы в интерпретации на естественном языке с использованием языковых моделей, где граф знаний выступает источником фактов и ограничений для обеспечения согласованности.

3. Интеграция данных из разных источников. Цифровой профиль распределён между несколькими платформами. Актуальна задача объединения мультимодальных данных из социальных сетей, образовательных и профессиональных сервисов с разрешением сущностей и устранением противоречий.

4. Адаптация к новым культурным контекстам. Результаты получены на выборке из определённой культурной среды. Необходима разработка методов адаптации графовых представлений и механизмов внимания к новым областям и культурам при ограниченных размеченных данных.

5. Динамические графы знаний. В работе профиль статичен, тогда как цифровая активность изменяется во времени. Перспективен переход к динамическим моделям, учитывающим эволюцию цифрового следа и изменение структуры графа, что позволит исследовать долгосрочные закономерности поведения и личностного развития.

#### 4.6 Выводы по главе 4

В данной главе представлена программная реализация разработанного нейросимвольного метода анализа цифрового следа пользователей социальных сетей.

В ходе исследования получены следующие основные результаты:

1. Разработана программная реализация нейросимвольной архитектуры анализа цифрового следа, объединяющая средства обработки мультимодальных данных, граф знаний и графовую нейронную сеть в рамках единой вычислительной системы.

2. Реализован граф знаний, обеспечивающий интеграцию цифровых признаков и предметных знаний, а также разработан алгоритм его автоматического наполнения на основе данных цифровой активности пользователей.

3. Реализован алгоритмический конвейер обработки цифрового профиля, включающий этапы сбора данных, извлечения признаков, построения графового представления, графового вывода и формирования интерпретируемых объяснений.

4. Реализованы средства визуализации и интерпретации результатов анализа, обеспечивающие представление предсказанных характеристик, структуры графа знаний и объясняющих зависимостей.

5. Выполнен анализ ограничений предложенного подхода и определены перспективные направления его дальнейшего развития.

## 5 Экспериментальная проверка разработанных методов и моделей

В настоящей главе приведены результаты экспериментальной проверки разработанных методов и моделей: определены метрики оценки и методика валидации, выполнена психометрическая валидация признаков, представлены результаты мультимодальной интеграции, сравнение графовой нейронной сети с базовыми моделями, валидация метрики интерпретируемости GPI и апробация системы на прикладной задаче. Экспериментальная база соответствует опубликованным результатам автора [211\*].

### 5.1 Программа экспериментальных исследований

Экспериментальная проверка организована в два уровня. Первый уровень – основная комплексная задача анализа личности по цифровому следу пользователя (три модальности, граф знаний, нейросимвольный вывод). Второй уровень – проверка адекватности и переносимости метода на независимые предметные области, где сохраняется инвариантное ядро метода, а изменяются предметное наполнение онтологии, состав признаков и целевых характеристик.

Уровень 1. Основная задача – анализ личности по цифровому следу (ВКонтакте). Программа экспериментов уровня 1 представлена в таблице 16  
Таблица 16 – Программа экспериментов уровня 1: анализ личности по цифровому следу

| Эксперимент                   | Что сравнивается                                                                                                                                   | Метрики                                         | Цель                           |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------|--------------------------------|
| Сравнение с базовыми моделями | линейная регрессия; сквозная мультимодальная сеть (RoBERTa, RuBERT, ResNet50, MLP); система экспертных правил; предложенная нейросимвольная модель | MAE, Pearson r, F1                              | качество прогноза              |
| Стратегии слияния             | раннее / позднее / гибридное слияние                                                                                                               | MAE, r, F1                                      | обоснование слияния            |
| Вклад модальностей (абляция)  | текст / визуал / табличная по отдельности и вместе                                                                                                 | MAE, r                                          | роль каждой модальности        |
| Валидация GPI                 | нейросимвольная модель и «чёрный ящик»                                                                                                             | GPI_cov, GPI_w; корреляция с экспертной оценкой | измеримость интерпретируемости |
| Тест удаления концептов       | падение уверенности при удалении значимых признаков                                                                                                | площадь под кривой                              | верность объяснений            |

Уровень 2. Проверка переносимости метода на другие предметные области. Программа кросс-доменных экспериментов приведена в таблице 17.



Сначала подтверждается адекватность метода на областях с известным эталонным результатом, а затем происходит раскрытие подхода в полном объеме на основной комплексной задаче, объединяющей три модальности.

Таблица 17 – Программа кросс-доменной проверки переносимости метода

| Область                        | Набор данных                    | Задача                                 | Метрики       | Приложение |
|--------------------------------|---------------------------------|----------------------------------------|---------------|------------|
| Дерматоскопическая диагностика | HAM10000 (ISIC 2018)            | классификация новообразований кожи     | точность, GPI | К          |
| Токсичные комментарии (англ.)  | HateXplain                      | классификация токсичности              | точность, GPI | Л          |
| Токсичные комментарии (рус.)   | Russian Language Toxic Comments | классификация токсичности              | точность, GPI | Л          |
| Эмоции (многоцелевая)          | CEDR                            | классификация эмоций (все 3 предиката) | точность, GPI | Л          |

## 5.2 Метрики оценки и методика валидации

Основной задачей исследования является предсказание 26 психометрических факторов личности (20 факторов 16PF Кэттелла и 6 интегральных факторов) на основе мультимодальных данных цифрового следа пользователей социальных сетей. Значения факторов представлены непрерывными величинами в диапазоне от 1 до 10 стенов, что позволяет рассматривать задачу как задачу множественной регрессии.

Для оценки качества моделей используются метрики среднеквадратичной ошибки (MSE), корня из среднеквадратичной ошибки (RMSE), средней абсолютной ошибки (MAE) и коэффициента детерминации ( $R^2$ ).

### 5.2.1 Метрики точности предсказания

Метрика 1 – среднеквадратичная ошибка (Mean Squared Error, MSE) и корень из среднеквадратичной ошибки (Root Mean Squared Error, RMSE)

$$MSE = \frac{1}{U} \sum_{u=1}^U \frac{1}{K} \sum_{k=1}^K (\hat{P}_{uk} - P_{uk})^2,$$

$$RMSE = \sqrt{MSE},$$

где:

$U$  – число пользователей в тестовой выборке;

$K = 26$  – число факторов;

$\hat{P}_{uk}$  – предсказанное значение  $k$ -го фактора для пользователя  $u$ ;

$P_{uk}$  – истинное значение из психометрического теста.

Метрики MSE и RMSE характеризуют среднюю величину отклонения прогнозируемых значений от истинных. RMSE измеряется в стенох и позволяет интерпретировать ошибку непосредственно в единицах психометрической шкалы.



Метрика 2 – средняя абсолютная ошибка (Mean Absolute Error (MAE))

$$MAE = \frac{1}{U} \sum_{u=1}^U \cdot \frac{1}{K} \sum_{k=1}^K | \hat{P}_{uk} - P_{uk} |.$$

Средняя абсолютная ошибка характеризует среднее отклонение предсказаний от истинных значений и является менее чувствительной к выбросам по сравнению с MSE.

Метрика 3 – коэффициент детерминации ( $R^2$ )

$$R_k^2 = 1 - \frac{\sum_{u=1}^U (\hat{P}_{uk} - P_{uk})^2}{\sum_{u=1}^U (P_{uk} - P_k)^2},$$

где  $P_k$  – среднее значение  $k$ -го фактора в выборке.

Коэффициент детерминации показывает долю дисперсии целевой характеристики, объясняемую моделью. Данная метрика характеризует долю вариативности психологического фактора, объясняемую используемыми признаками цифрового следа.

В качестве ориентиров для интерпретации результатов использовались данные опубликованных исследований по прогнозированию личностных характеристик на основе цифрового следа пользователей социальных сетей. Согласно результатам метаанализа, значения коэффициента детерминации для подобных задач обычно находятся в диапазоне 0,29–0,40, что соответствует умеренному уровню объясняемой дисперсии психологических характеристик.

### 5.2.2 Метрики валидности корреляций

Помимо точности прогнозирования отдельных факторов личности, важным аспектом является сохранение структуры взаимосвязей между психологическими характеристиками. В психометрии многие факторы обладают устойчивыми корреляционными связями, отражающими внутреннюю организацию личностного пространства. Поэтому дополнительно оценивается согласованность корреляционной структуры предсказанных факторов с корреляционной структурой исходных психометрических данных.

Метрика 4 – индекс сохранения корреляций (Correlation Preservation Score, CPS)

$$CPS = \frac{1}{M} \sum_{i=1}^M \mathbb{1}[\text{sign}(\rho_{ij}^{pred}) = \text{sign}(\rho_{ij}^{true})] \cdot w_{ij},$$

где:

- $M$  – число пар факторов с известными корреляциями из психометрической теории;
- $\rho_{ij}^{pred}$  – корреляция между предсказанными значениями факторов  $i$  и  $j$ ;

- $\rho_{ij}^{true}$  – истинная корреляция (из теста на обучающей выборке или из литературы);
- $w_{ij} = |\rho_{ij}^{true}|$  – вес (сильные корреляции важнее);
- $1[\cdot]$  – индикаторная функция (1, если знаки совпадают).

Метрика CPS характеризует степень сохранения направленности корреляционных связей между психологическими факторами в предсказаниях модели. Более высокие значения CPS соответствуют лучшему сохранению структуры психометрического пространства.

Метрика 5 – канонический корреляционный анализ (Canonical Correlation Analysis, CCA).

Для оценки согласованности многомерной структуры предсказанных и истинных факторов используется канонический корреляционный анализ:

$$\rho_{CCA} = \max_{w_P, w_{\hat{P}}} \text{corr}(w_P^T P, w_{\hat{P}}^T \hat{P}),$$

где:

- $P$  – матрица истинных психометрических факторов;
- $\hat{P}$  – матрица предсказанных факторов;
- $w_P^T$  – канонические коэффициенты.

Каноническая корреляция позволяет оценить степень соответствия общей структуры предсказанного и реального психометрического пространства. Высокие значения свидетельствуют о сохранении системой взаимосвязей между психологическими характеристиками пользователей.

### 5.2.3 Метрики интерпретируемости

Одним из требований к разрабатываемой системе является возможность интерпретации результатов прогнозирования. Помимо точности предсказаний, необходимо оценивать согласованность объяснений модели с используемыми данными и структурой предметной области.

Метрика 6 – согласованность весов внимания (Attention Weights Consistency, AWC).

Для моделей, использующих механизм внимания при объединении мультимодальных данных, дополнительным объектом анализа являются веса внимания, характеризующие относительный вклад различных модальностей в формирование предсказания.

Для фактора  $k$  согласованность весов внимания определяется следующим образом:

$$AWC_k = \frac{\sum_{u=1}^U \lambda_{m^*}^{(u,k)}}{\sum_{u=1}^U \sum_m \lambda_m^{(u,k)}}$$

где:

- $\lambda_m^{(u,k)}$  – вес модальности  $m$  при предсказании фактора  $k$  для пользователя  $u$ ;
- $m^* = \operatorname{argmax}_m E[\lambda_m^{(u,k)}]$  – доминирующая модальность для фактора  $k$ .

Интегральная оценка вычисляется как среднее значение по всем прогнозируемым факторам:

$$AWC = \frac{1}{K} \sum_{k=1}^K \mathbb{1}[m_k^* = m_k^{expected}],$$

где:

- $K$  – число факторов личности;
- $m_k^{expected}$  – модальность, предположительно наиболее информативная для данного фактора согласно результатам психометрического анализа и экспертной оценке.

Метрика AWC характеризует степень согласованности распределения внимания модели с ожидаемой информативностью различных модальностей. Высокие значения метрики свидетельствуют о том, что механизм внимания использует источники информации, наиболее релевантные прогнозируемым психологическим характеристикам.

Метрика 7 – интерпретируемость путей в графе GPI (Graph Path Interpretability).

Для оценки качества объяснений, формируемых на основе графа знаний, используется метрика интерпретируемости путей GPI, предложенная в параграфе 2.4.

Для каждого предсказания формируется объясняющий подграф, включающий наиболее значимые пути, связывающие признаки цифрового следа пользователя с прогнозируемой характеристикой личности. Метрика GPI оценивает степень согласованности объясняющего подграфа с внутренними механизмами модели и структурой графа знаний.

Обобщённая оценка интерпретируемости вычисляется как среднее значение GPI по всем пользователям и прогнозируемым характеристикам:

$$GPI = \frac{1}{U \cdot K} \sum_{u=1}^U \sum_{k=1}^K \mathbb{1}[\exists \text{ path from } u \text{ to } P_k \text{ with } p < 0,05],$$

где:

- $U$  – число пользователей;
- $K$  – число прогнозируемых характеристик.

Значение метрики интерпретируемости для объяснения  $k$ -й характеристики пользователя  $u$ .

Высокие значения GPI соответствуют объяснениям, основанным на структурно согласованных путях графа знаний и отражающим значимые зависимости между признаками цифрового следа и психологическими характеристиками. Низкие значения GPI свидетельствуют о слабой согласованности объясняющего подграфа с внутренней логикой модели.

Метрика GPI используется для сравнения различных вариантов графовых моделей и оценки качества формируемых объяснений наряду с традиционными показателями точности прогнозирования. Полное определение метрики GPI как системы – основная метрика покрытия путями (GPI\_cov), вспомогательная метрика взвешенной согласованности (GPI\_w) и проверка верности методом удаления концептов – приведено в Приложении П, а её расширенное применение для сравнения рассуждающих архитектур по критерию интерпретируемости показано в кросс-доменных приложениях К и Л. В основной экспериментальной части обозначение GPI без индекса используется для основной компоненты GPI\_cov.

#### 5.2.4 Базовые модели

Для оценки эффективности предложенного нейросимвольного метода выполнено сравнение с тремя альтернативными подходами, отражающими различные стратегии анализа данных: статистическую, нейросетевую и экспертно-символьную.

Базовая модель 1 – линейная регрессия.

В качестве статистического базового метода используется линейная регрессия с L2-регуляризацией (Ridge Regression). Входными данными являются агрегированные признаки цифрового профиля пользователя, извлекаемые из VK API, включая характеристики социальной активности, публикационной активности и параметры профиля. Для каждого из 26 психометрических факторов обучается отдельная регрессионная модель.

Поскольку модель не использует графовое представление данных и механизмы нейросимвольной интеграции, её результаты могут рассматриваться в качестве базового уровня для последующего сравнения с предлагаемым подходом.

Базовая модель 2 – сквозная мультимодальная нейронная сеть.

В качестве нейросетевого базового метода используется мультимодальная архитектура, объединяющая модели обработки текста, изображений и данных цифровой активности.

Текстовая модальность обрабатывается моделью BERT, визуальная — сверточной нейронной сетью ResNet, а признаки цифровой активности поступают на вход многослойного персептрона (MLP). Объединение признаков выполняется на уровне скрытых представлений без использования графа знаний и онтологической информации.

Данная модель используется для оценки влияния графа знаний и механизмов нейросимвольной интеграции на качество прогнозирования по сравнению с традиционным мультимодальным обучением.

Базовая модель 3 – система экспертных правил.

В качестве символьного базового метода используется система правил, построенная на основе экспертных психологических гипотез о взаимосвязи цифрового поведения и личностных характеристик пользователей.

Правила задаются вручную и связывают наблюдаемые характеристики цифрового профиля с отдельными психологическими факторами. Например, высокая социальная активность и большое число социальных связей могут интерпретироваться как признаки экстраверсии. Всего в системе используется 50 экспертных правил, охватывающих основные группы признаков цифрового следа.

Подход характеризуется высокой интерпретируемостью, однако не обладает механизмами обучения на данных и адаптации к новым закономерностям.

Предлагаемый подход основан на совместном использовании графа знаний и графовой нейронной сети внимания. В отличие от базовых моделей он обеспечивает интеграцию мультимодальных данных и предметных знаний в рамках единого графового представления, а также формирование интерпретируемых объяснений на основе структуры графа знаний.

Сравнение моделей выполняется по метрикам RMSE, MAE,  $R^2$ , CPS, CCA и GPI, описанным в параграфе 5.2.

#### 5.2.5 Статистическая значимость и кросс-валидация

Для оценки статистической значимости различий между предложенной нейросимвольной моделью и базовыми моделями используется парный t-критерий Стьюдента по значениям ошибки прогнозирования, рассчитанным для пользователей тестовой выборки.

Нормальность распределения попарных разностей ошибок прогнозирования предварительно проверялась по критерию Шапиро-Уилка; значимых отклонений от

нормальности, препятствующих применению параметрического критерия, не выявлено.

Статистика критерия определяется следующим образом:

$$t = \frac{MSE_{proposed} - MSE_{baseline}}{SE(MSE_{proposed} - MSE_{baseline})},$$

где SE – стандартная ошибка разности.

Формулируются гипотезы:

Нулевая гипотеза  $H_0: MSE_{proposed} = MSE_{baseline}$

Альтернативная гипотеза  $H_1: MSE_{proposed} < MSE_{baseline}$ .

Уровень значимости устанавливается на уровне  $\alpha = 0,01$ .

С учётом трёх попарных сравнений применяется поправка Бонферрони:

$$\alpha' = \frac{0,01}{3} = 0,0033.$$

Для оценки устойчивости результатов используется 5-кратная перекрёстная проверка (5-fold cross-validation). Выбор данной схемы обусловлен объёмом размеченной выборки (1722 пользователя) и необходимостью обеспечения достаточного числа наблюдений как на этапе обучения, так и на этапе тестирования.

На каждой итерации четыре фолда используются для обучения модели, а оставшийся фолд – для тестирования. Для каждой итерации вычисляются значения метрик RMSE, MAE,  $R^2$ , CPS, CCA, AWC и GPI.

Перекрёстная проверка используется для оценки устойчивости модели к вариациям обучающей выборки и снижения риска случайной переоценки качества.

Таким образом, сочетание статистических критериев и перекрёстной проверки повышает надёжность выводов о качестве прогнозирования и интерпретируемости предложенного подхода.

### 5.3 Психометрическая валидация признаков

Одним из основных этапов исследования является проверка психологической валидности признаков цифровой идентичности, то есть их связи с личностными характеристиками, измеренными стандартизированными психометрическими методиками. Психологическая валидность текстовых элементов цифрового профиля является необходимым условием их использования в качестве признаков при прогнозировании личностных характеристик.

В рамках исследования проведён анализ взаимосвязей между семантическими категориями никнеймов и психологическими факторами по методике 16PF Кэттелла.

На основе теоретических представлений о цифровой идентичности и результатов ранее опубликованных исследований были сформулированы следующие гипотезы:

- Н1: никнеймы с литературными референсами связаны с фактором М (мечтательность);
- Н2: никнеймы, содержащие имена знаменитостей, связаны с фактором Н (социальная смелость);
- Н3: никнеймы с зоонимами связаны с фактором I (чувствительность);
- Н4: метафорические никнеймы связаны с фактором Q1 (радикализм);
- Н5: абстрактные никнеймы связаны с фактором О (тревожность);
- Н6: никнеймы с топонимами связаны с фактором А (общительность).

Исследование выполнено на выборке из 1722 пользователей социальной сети ВКонтакте (возраст 18–52 года, 57,6 % мужчин), для которых были доступны результаты психометрического тестирования.

Для проверки гипотез рассчитывались коэффициенты корреляции Пирсона между категориями никнеймов и психологическими факторами. Для контроля множественных сравнений использовалась поправка Бонферрони ( $p < 0,0004$ ). Дополнительно выполнялся регрессионный анализ с контролем возраста и пола. Для проверки мультиколлинеарности использовался коэффициент инфляции дисперсии (VIF). Подробности статистического анализа приведены в Приложении Г.

В результате выявлены статистически значимые связи между отдельными категориями никнеймов и психологическими факторами ( $p < 0,0004$ ). Наиболее выраженные ассоциации представлены в таблице 18.

Таблица 18 – Статистически значимые связи между категориями никнеймов и психологическими факторами

| Категория никнейма    | Фактор | $r$  | $d$ Коэна |
|-----------------------|--------|------|-----------|
| Литературный персонаж | М      | 0,38 | 0,52      |
| Знаменитость          | Н      | 0,42 | 0,58      |
| Зооним                | I      | 0,44 | 0,61      |
| Метафора              | Q1     | 0,36 | 0,49      |

Наиболее сильная связь обнаружена между категорией «Зооним» и фактором I ( $r = 0,44$ ;  $d = 0,61$ ). Также высокие значения корреляции наблюдаются для категорий «Знаменитость» и «Литературный персонаж».

#### 5.4 Мультимодальная интеграция признаков

Результаты лингвистического и визуального анализа, представленные в параграфах 3.2 и 3.3, показали, что различные компоненты цифровой идентичности содержат психологически значимую информацию о личности пользователя. Поэтому

следующей задачей является их совместное использование при прогнозировании личностных характеристик.

Преимущество мультимодального подхода заключается в том, что различные модальности дополняют друг друга, компенсируют недостаток информации и совместно повышают качество прогнозирования.

Для оценки эффективности предложенного подхода выполнено сравнение трёх стратегий мультимодальной интеграции:

- раннего слияния (Early Fusion);
- позднего слияния (Late Fusion);
- гибридного слияния (Hybrid Fusion).

В качестве предлагаемого решения используется архитектура гибридного слияния, включающая механизм межмодального внимания и механизм адаптивного объединения признаков, описание которых приведено в приложении Д. Архитектура учитывает взаимодействия между текстовыми и визуальными признаками, а также динамически определяет вклад отдельных модальностей в зависимости от их информативности для конкретного пользователя.

Архитектура предлагаемой модели включает пять основных компонентов, представленных в таблице 19. Механизм межмодального внимания используется для выявления соответствий между текстовыми и визуальными признаками. Например, при согласованности никнейма и изображения профиля модель усиливает вклад соответствующих признаков в итоговое предсказание.

Таблица 19 – Архитектура мультимодальной модели интеграции

| Этап | Компонент                             | Описание                                                                                                                                                                                  |
|------|---------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1    | Модальные энкодеры                    | RoBERTa для текстовой модальности: никнейм и статус, 768 признаков; ResNet50 для визуальной модальности: аватар, 2048 признаков. Выполняется извлечение эмбедингов из каждой модальности. |
| 2    | Проекция                              | Полносвязные слои переводят модальные представления в общее латентное пространство размерности 512.                                                                                       |
| 3    | Двунаправленное межмодальное внимание | Механизм межмодального внимания сопоставляет текстовые и визуальные признаки и формирует взаимосвязанные представления модальностей. Остаточные связи сохраняют исходную информацию.      |
| 4    | Адаптивное слияние модальностей       | Обучаемый коэффициент $\lambda$ $[0,1]$ определяет относительный вклад текстовой и визуальной модальностей. Значение $\lambda$ настраивается в процессе обучения модели.                  |
| 5    | Выходной слой                         | Полносвязный слой (Linear (512,26)) формирует значения 26 психологических факторов. Функция потерь включает ошибку предсказания и корреляционный штраф.                                   |

Адаптивное слияние модальностей учитывает различную информативность источников данных для конкретного пользователя. Если текстовая модальность содержит мало информации, больший вклад получает визуальная модальность; если



изображение профиля неинформативно, больший вес может быть присвоен текстовым признакам. Это особенно важно при анализе данных социальных сетей, для которых характерны неполнота и неоднородность информации.

Предлагаемая архитектура гибридного слияния валидирована на выборке из 1722 пользователей ВКонтакте с полными психологическими профилями. Для пользователей были доступны никнеймы, статусы и изображения профиля; доля доступных статусов составила 91%. Выборка разделялась на обучающую, валидационную и тестовую части в соотношении 70%/15%/15% со стратификацией по полу.

В таблице 20 представлены результаты сравнения предложенной архитектуры гибридного слияния с базовыми моделями. Технические детали расчёта метрик и процедуры разбиения выборки приведены в Приложении Д.

Таблица 20 – Сравнение архитектуры гибридного слияния с базовыми моделями

| Модель                       | Модальности        | MAE           | Коэффициент Пирсона $r$ | F1-score | Число параметров |
|------------------------------|--------------------|---------------|-------------------------|----------|------------------|
| Baseline 1: случайная модель | –                  | $4,9 \pm 0,1$ | $0,02 \pm 0,03$         | 0,33     | –                |
| Baseline 2: среднее значение | –                  | $3,2 \pm 0,0$ | 0,00                    | 0,41     | –                |
| Текстовая модель (RoBERTa)   | Никнейм, статус    | $3,8 \pm 0,4$ | $0,62 \pm 0,05$         | 0,71     | 125 млн          |
| Визуальная модель (ResNet50) | Аватар             | $4,2 \pm 0,5$ | $0,55 \pm 0,06$         | 0,68     | 25 млн           |
| Раннее слияние               | Текст, изображения | $3,6 \pm 0,4$ | $0,65 \pm 0,05$         | 0,74     | 150 млн          |
| Позднее слияние              | Текст, изображения | $3,5 \pm 0,3$ | $0,67 \pm 0,04$         | 0,76     | 150 млн          |
| Гибридное слияние            | Текст, изображения | $3,1 \pm 0,3$ | $0,72 \pm 0,04$         | 0,81     | 158–205 млн      |

Для гибридной модели приведено число обучаемых параметров; полный размер модели с учётом замороженных предобученных энкодеров составляет около 205 млн.

Наилучшие значения всех рассматриваемых метрик получены для модели гибридного слияния, что свидетельствует о целесообразности совместного использования текстовых и визуальных признаков.

По сравнению с базовыми моделями получены следующие улучшения качества:

- по сравнению с текстовой моделью: MAE уменьшилась на 18%, коэффициент Пирсона увеличился на 16%, значение F1-score – на 14%;
- по сравнению с визуальной моделью: MAE уменьшилась на 26%, коэффициент Пирсона увеличился на 31%, значение F1-score – на 19%;
- по сравнению с ранним слиянием: MAE уменьшилась на 14%, коэффициент Пирсона увеличился на 11%, значение F1-score – на 9%;

– по сравнению с поздним слиянием: MAE уменьшилась на 11%, коэффициент Пирсона увеличился на 7%, значение F1-score – на 7%.

Статистическая значимость различий между гибридной и текстовой моделями проверялась с использованием парного t-критерия Стьюдента по значениям метрик для 26 факторов личности. Полученные результаты показывают статистически значимое превосходство гибридной модели:  $t(25) = 4,82$ ,  $p < 0,0001$ .

Полученные результаты подтверждают, что объединение текстовых и визуальных признаков формирует более полное представление цифровой идентичности пользователя по сравнению с использованием отдельных модальностей. Дополнительный прирост качества достигается за счёт механизма межмодального внимания, учитывающего взаимосвязи между различными типами признаков, и механизма адаптивного объединения, динамически определяющего вклад каждой модальности. В проведённых экспериментах значение коэффициента адаптивного объединения изменялось в диапазоне от 0,12 до 0,82, отражая различную значимость отдельных модальностей для разных пользователей.

Таким образом, использование механизмов межмодального внимания и адаптивного объединения признаков обеспечивает более эффективную интеграцию разнородных данных цифрового следа.

Для оценки вклада отдельных компонентов архитектуры проведён абляционный анализ, заключающийся в поочерёдном исключении элементов модели. Удаление механизма межмодального внимания привело к увеличению MAE на 13%. Отключение адаптивного объединения модальностей ухудшило качество на 10% по сравнению с вариантом с фиксированным коэффициентом  $\lambda = 0,5$ . Исключение остаточных связей увеличило ошибку на 6%, а отказ от корреляционного штрафа – на 3%. Технические детали экспериментов приведены в Приложении Е.

Результаты абляционного анализа показывают, что наибольший вклад в качество прогнозирования вносит механизм межмодального внимания, обеспечивающий учёт взаимосвязей между текстовыми и визуальными признаками.

Наибольший эффект мультимодальной интеграции наблюдается для факторов, одновременно проявляющихся в текстовых и визуальных компонентах цифрового профиля. Для фактора I (Чувствительность) прирост коэффициента корреляции составил  $\Delta r = 0,16$ , для фактора М (Мечтательность) –  $\Delta r = 0,14$ , для фактора Н (Социальная смелость) –  $\Delta r = 0,11$ , для фактора Q1 (Радикализм) –  $\Delta r = 0,11$ . Для факторов, слабо представленных в статических характеристиках профиля, улучшение

оказалось менее выраженным  $\Delta r = 0,02-0,05$ . Полные результаты представлены в Приложении Е.

Таким образом, архитектура гибридного слияния демонстрирует улучшение качества прогнозирования по сравнению как с одномодальными моделями, так и с традиционными схемами интеграции признаков. По сравнению с текстовой моделью значение MAE снизилось с 3,8 до 3,1, а коэффициент корреляции Пирсона увеличился с 0,62 до 0,72. По сравнению с визуальной моделью значение MAE уменьшилось с 4,2 до 3,1, а коэффициент корреляции Пирсона увеличился с 0,55 до 0,72.

Для анализа причин формирования предсказаний используются SHAP-значения и карты внимания, отражающие вклад отдельных модальностей и признаков в итоговый результат. В среднем вклад текстовой модальности составил 54 %, визуальной – 46 %, при этом распределение вкладов варьировалось в зависимости от прогнозируемого психологического фактора.

Таким образом, применение механизмов межмодального внимания и адаптивного объединения признаков является одним из главных факторов повышения качества мультимодального анализа цифрового следа. Разработанная мультимодальная модель формирует нейросетевую подсистему предложенной нейросимвольной архитектуры. Извлечённые признаки цифровой идентичности используются при построении графа знаний и в дальнейшем участвуют в графовом выводе на основе графовой нейронной сети, рассматриваемом в следующем параграфе.

#### 5.4.1 Интерпретируемость мультимодальной модели

Высокая точность прогнозирования представляет практический интерес только при наличии возможности интерпретации получаемых результатов. В связи с этим для анализа работы мультимодальной модели использовались два взаимодополняющих механизма интерпретации: механизм межмодального внимания и SHAP-анализ.

Механизм межмодального внимания обеспечивает встроенную интерпретируемость модели, позволяя анализировать взаимодействие текстовых и визуальных признаков в процессе формирования предсказания. Дополнительно для количественной оценки вклада отдельных модальностей использовался метод SHAP (SHapley Additive exPlanations), основанный на концепции значений Шепли из теории игр. В отличие от весов внимания, SHAP оценивает вклад признаков в формирование итогового предсказания. Подробное описание метода приведено в Приложении Ж.

Средние значения вкладов текстовой и визуальной модальностей для тестовой выборки представлены в таблице 21.

Таблица 21 – Вклад модальностей в предсказание психологических факторов по результатам SHAP-анализа

| Фактор 16PF             | Текстовая модальность | Визуальная модальность | Интерпретация                             |
|-------------------------|-----------------------|------------------------|-------------------------------------------|
| М (Мечтательность)      | 65 %                  | 35 %                   | Преобладает вклад текстовых признаков     |
| I (Чувствительность)    | 48 %                  | 52 %                   | Сопоставимый вклад обеих модальностей     |
| Н (Социальная смелость) | 42 %                  | 58 %                   | Преобладает вклад визуальных признаков    |
| Q1 (Радикализм)         | 55 %                  | 45 %                   | Вклад модальностей близок                 |
| Среднее по 26 факторам  | 54 %                  | 46 %                   | Обе модальности вносят сопоставимый вклад |

Результаты SHAP-анализа показывают, что в среднем текстовая модальность вносит несколько больший вклад в формирование предсказаний (54% против 46%). Вместе с тем для отдельных факторов наблюдается существенное перераспределение значимости между модальностями. Так, для факторов Н и I более информативными оказываются визуальные признаки, тогда как для факторов М и Q1 большую роль играют текстовые характеристики цифрового профиля.

Таким образом, анализ вкладов модальностей подтверждает, что текстовые и визуальные признаки содержат различную, но взаимодополняющую информацию о личности пользователя.

Результаты SHAP-анализа показывают, что текстовая модальность вносит больший вклад в прогнозирование факторов, связанных с когнитивными характеристиками личности (например, факторов М и Q1), тогда как визуальная модальность оказывается более информативной для факторов, отражающих социально-эмоциональные особенности пользователя (например, факторов Н и I). Полученные результаты показывают, что использование только одной модальности приводит к потере части информации, значимой для прогнозирования личностных характеристик.

Механизм межмодального внимания формирует соответствия между текстовыми и визуальными признаками цифрового профиля. Высокие значения весов внимания наблюдаются в случаях, когда текстовые и визуальные компоненты отражают сходное смысловое содержание. В таких ситуациях обе модальности совместно участвуют в формировании предсказания, повышая его устойчивость.

При отсутствии выраженных семантических соответствий между текстовыми и визуальными признаками механизм адаптивного объединения перераспределяет вклад модальностей в пользу более информативного источника данных. В результате

снижается чувствительность модели к неполноте и неоднородности цифрового профиля.

Карты внимания используются для визуального анализа работы модели и позволяют выявлять признаки, оказывающие наибольшее влияние на формирование предсказания. Это упрощает интерпретацию результатов и анализ причин формирования предсказаний.

Рассмотренные примеры иллюстрируют три характерных режима работы мультимодальной модели.

В первом случае наблюдается согласованность текстовых и визуальных признаков, что приводит к взаимному усилению информационного сигнала и повышению точности прогнозирования. Во втором случае одна модальность компенсирует недостаток информации в другой, сохраняя качество предсказания при неполноте исходных данных. В третьем случае механизм адаптивного объединения признаков разрешает противоречие между модальностями путём перераспределения их вкладов в соответствии с информативностью для рассматриваемой задачи.

Анализ весов внимания и SHAP-значений показывает, что модель учитывает не только отдельные признаки, но и взаимосвязи между различными источниками информации. Это даёт возможность проследить вклад текстовых и визуальных компонентов в формирование итогового предсказания и выполнить дополнительную проверку его обоснованности.

Таким образом, предложенная архитектура сочетает повышение точности прогнозирования с возможностью анализа вклада отдельных модальностей и признаков в итоговый результат.

## 5.5 Экспериментальное сравнение GAT с базовыми моделями

Для оценки эффективности предложенной архитектуры были рассмотрены следующие базовые модели.

1. MLP (Flat Features) – многослойный перцептрон, обучаемый на векторных признаках без использования графовой структуры. Входной вектор формируется путём объединения текстовых, визуальных, табличных и демографических признаков пользователя:

$$x_u^{flat} = [x_{text}^{(u)}, x_{visual}^{(u)}, x_{behavior}^{(u)}, age, gender, \dots] \in \mathbb{R}^{2900}.$$

В качестве базовой модели используется многослойный перцептрон с входным слоем размерности 2900, тремя скрытыми слоями размерности 512, 256 и 128 нейронов

и выходным слоем размерности 26, соответствующим прогнозируемым психологическим факторам.

2. GCN (Graph Convolutional Network) – графовая сверточная сеть, использующая структуру графа знаний, но не выполняющая дифференцированное взвешивание соседних узлов посредством механизма внимания. Все соседние узлы участвуют в агрегировании информации с одинаковым весом [13].

3. Hybrid Fusion – мультимодальная модель, рассмотренная в параграфе 5.4, объединяющая текстовые, визуальные признаки и данные цифровой активности без использования графа знаний и графовых нейронных сетей. Данная модель представляет лучший результат среди моделей мультимодальной интеграции и используется в качестве основной неграфовой базовой модели ( $MAE = 3,1$ ;  $Pearson\ r = 0,72$ ).

4. GAT (предлагаемая модель) – разработанная графовая нейронная архитектура, включающая три слоя Graph Attention Network, механизм онтологической регуляризации и процедуру итеративного обновления весов рёбер графа знаний. Модель выполняет предсказание психологических факторов посредством распространения информации по графу с использованием механизма внимания.

Для проведения экспериментов использовались данные 1722 пользователей, для которых доступны результаты психометрического тестирования по модели 16PF.

Разделение выборки выполнялось в соотношении 70 % / 15 % / 15 %:

- обучающая выборка – 1206 пользователей;
- валидационная выборка – 258 пользователей;
- тестовая выборка – 258 пользователей.

Разбиение выполнялось со стратификацией по полу и возрасту.

Для графовых моделей (GCN и GAT) обучение осуществлялось на полном графе знаний, содержащем 19 690 пользователей. Использовалась полуконтролируемая трансдуктивная постановка: распространение информации выполнялось по всему графу, тогда как функция потерь вычислялась только для узлов пользователей, имеющих психометрические метки (1722 пользователя).

В результате при обучении модели используются как психометрически размеченные данные, так и структурные связи между всеми пользователями графа.

Для оценки качества моделей использовались следующие метрики:

1. MAE (Mean Absolute Error) – средняя абсолютная ошибка по всем факторам и пользователям:

$$MAE = \frac{1}{N \cdot 26} \sum_{u,k} |\hat{p}_{u,k} - p_{u,k}|,$$

где  $\hat{p}_{u,k}$  – предсказанное значение  $k$ -го фактора для пользователя  $u$ ,  $p_{u,k}$  – истинное значение.

2. Pearson  $r$  – средняя корреляция Пирсона между предсказанными и истинными значениями факторов.

3. F1-score (macro) вычисляется после дискретизации факторов на три уровня: низкий, средний и высокий – по 33-му и 67-му перцентилям.

4.  $R^2$  (коэффициент детерминации) – доля объяснённой дисперсии между предсказанными и истинными значениями факторов.

Результаты сравнения моделей на тестовой выборке ( $N = 258$ ) по указанным метрикам приведены в таблице 22.

Таблица 22 – Сравнение моделей по метрикам качества на тестовой выборке

| Модель                    | MAE ↓         | Pearson $r$ ↑   | F1-score ↑ | $R^2$ ↑ | Параметры |
|---------------------------|---------------|-----------------|------------|---------|-----------|
| MLP (векторные признаки)  | $3,9 \pm 0,5$ | $0,58 \pm 0,06$ | 0,69       | 0,34    | 2,8 млн   |
| GCN (3 слоя)              | $3,5 \pm 0,4$ | $0,66 \pm 0,05$ | 0,74       | 0,44    | 3,2 млн   |
| Hybrid Fusion             | $3,1 \pm 0,3$ | $0,72 \pm 0,04$ | 0,81       | 0,52    | 158 млн   |
| GAT (предлагаемая модель) | $2,7 \pm 0,3$ | $0,78 \pm 0,04$ | 0,85       | 0,61    | 3,5 млн   |

По сравнению с базовыми моделями предлагаемая архитектура GAT демонстрирует устойчивое улучшение качества предсказания.

По сравнению с моделью MLP, использующей только векторные признаки без графовой структуры, наблюдается снижение MAE на 31%, увеличение корреляции Pearson  $r$  на 34% и рост F1-score на 23%.

По сравнению с GCN, использующей граф, но не учитывающей механизм внимания, MAE уменьшается на 23%, коэффициент Пирсона  $r$  увеличивается на 18%, а F1-score – на 15 %.

По сравнению с моделью Hybrid Fusion (лучший результат среди моделей мультимодальной интеграции) достигается снижение MAE на 13%, увеличение Pearson  $r$  на 8% и рост F1-score на 5%.

В предлагаемом подходе обучению подлежит только графовый модуль (GAT, 3,5 млн параметров), тогда как мультимодальные энкодеры используются в режиме фиксированных параметров. Это сокращает число обучаемых параметров и

вычислительные затраты на обучение по сравнению со сквозной моделью Hybrid Fusion, обучаемой целиком (158 млн параметров), при одновременном улучшении качества прогнозирования по всем рассматриваемым метрикам. Полная вычислительная стоимость вывода в обоих случаях сопоставима, поскольку включает предобученные энкодеры; выигрыш относится именно к этапу обучения.

Для проверки статистической значимости различий между предлагаемой моделью и базовыми моделями был выполнен парный t-критерий Стьюдента по значениям метрик для 26 факторов личности, вычисленным на тестовой выборке (258 пользователей).

Для сравнения моделей GAT и Hybrid Fusion был выполнен парный t-критерий Стьюдента. Полученное значение

$$t(25) = 4,82, \quad p < 0,0001$$

свидетельствует о статистически значимом улучшении качества прогнозирования предлагаемой модели по сравнению с архитектурой Hybrid Fusion.

Для более детального анализа результатов рассмотрены ошибки предсказания по отдельным факторам 16PF. В таблице 23 представлены факторы, для которых наблюдается наибольшее улучшение по сравнению с моделью Hybrid Fusion, а также факторы, характеризующиеся наибольшей ошибкой прогнозирования.

Таблица 23 – Результаты прогнозирования отдельных факторов 16PF

| Фактор                         | GAT MAE | Hybrid Fusion MAE | $\Delta$ MAE (%) | Интерпретация                                   |
|--------------------------------|---------|-------------------|------------------|-------------------------------------------------|
| Фактор М (Абстрактность)       | 2,1     | 2,9               | -28%             | литературные никнеймы, книги в аватарах         |
| Фактор Н (Социальная смелость) | 2,3     | 3,1               | -26%             | размер сети, романтические фото                 |
| Фактор G (Нормативность)       | 2,4     | 3,0               | -20%             | утренняя активность, данные цифровой активности |
| Фактор I (Чувствительность)    | 2,5     | 3,2               | -22%             | зоонимы, эмоции на фото                         |
| Фактор А (Общительность)       | 2,6     | 3,1               | -16%             | позитивная тональность, высокая активность      |
| Фактор В (Интеллект)           | 3,8     | 3,5               | +9%              | слабая связь признаков                          |
| Фактор О (Тревожность)         | 3,6     | 3,4               | +6%              | слабая проявленность в цифровом следе           |
| Фактор Q2 (Самодостаточность)  | 3,4     | 3,3               | +3%              | трудно наблюдаемое поведение                    |

Анализ результатов показывает, что наибольший прирост качества достигается для факторов, связанных с особенностями социального поведения, чувствительности и абстрактного мышления. Для ряда факторов, отражающих внутренние когнитивные и эмоциональные процессы, улучшение оказывается менее выраженным. Вероятной



причиной является ограниченная представленность соответствующих характеристик в цифровом следе пользователей, что снижает информативность наблюдаемых признаков.

Модель GAT обеспечивает меньшую ошибку MAE по пяти факторам из восьми (М, Н, G, I, A) со снижением до 28%, тогда как для трёх слабо проявленных в цифровом следе факторов (В, О, Q2) преимущество не достигается.

Следует отметить, что улучшение качества прогнозирования наблюдается преимущественно для факторов, имеющих выраженные табличные, текстовые или визуальные проявления. Гипотеза о том, что использование графа знаний наиболее эффективно для характеристик личности, представленных в нескольких модальностях цифрового следа одновременно, подтверждается.

Для оценки вклада отдельных компонентов архитектуры был проведён абляционный анализ. В ходе эксперимента из полной модели GAT последовательно исключались отдельные компоненты, после чего измерялось изменение качества прогнозирования. Результаты представлены в таблице 24.

Таблица 24 – Абляционный анализ компонентов модели GAT

| Модель (удалённый компонент)            | MAE | $\Delta$ MAE | Pearson $r$ | $\Delta r$ |
|-----------------------------------------|-----|--------------|-------------|------------|
| GAT (полная модель)                     | 2,7 | –            | 0,78        | –          |
| Без многоголового внимания (1 head)     | 3,0 | +11 %        | 0,73        | –6 %       |
| Без онтологической регуляризации        | 2,9 | +7 %         | 0,75        | –4 %       |
| Без итеративного обновления весов рёбер | 3,1 | +15 %        | 0,72        | –8 %       |
| Без третьего слоя GAT                   | 2,9 | +7 %         | 0,76        | –3 %       |
| Без графа знаний (плоские признаки)     | 3,9 | +44 %        | 0,58        | –26 %      |

Результаты абляционного анализа показывают, что наибольший вклад в качество прогнозирования вносит механизм итеративного обновления весов рёбер. Его исключение приводит к увеличению значения MAE на 15 % и снижению коэффициента корреляции Пирсона на 8 %. Полученный результат указывает на важность адаптации весов связей графа знаний в процессе обучения модели.

Значимый вклад также вносит механизм многоголового внимания. При использовании одной головы внимания значение MAE возрастает на 11%, что указывает на эффективность одновременного анализа различных типов связей и признаков в графе.

Наиболее существенное ухудшение качества наблюдается при отказе от графовой структуры и переходе к модели на плоских признаках. В этом случае значение MAE увеличивается на 44 %, а коэффициент корреляции Пирсона уменьшается на 26%. Полученные результаты указывают на то, что учёт структуры

связей между объектами и признаками в графе знаний вносит существенный вклад в качество прогнозирования.

Дополнительно была выполнена оценка динамики обучения модели по значениям функции потерь на обучающей и валидационной выборках. Графики изменения функции потерь представлены на рисунке 20.

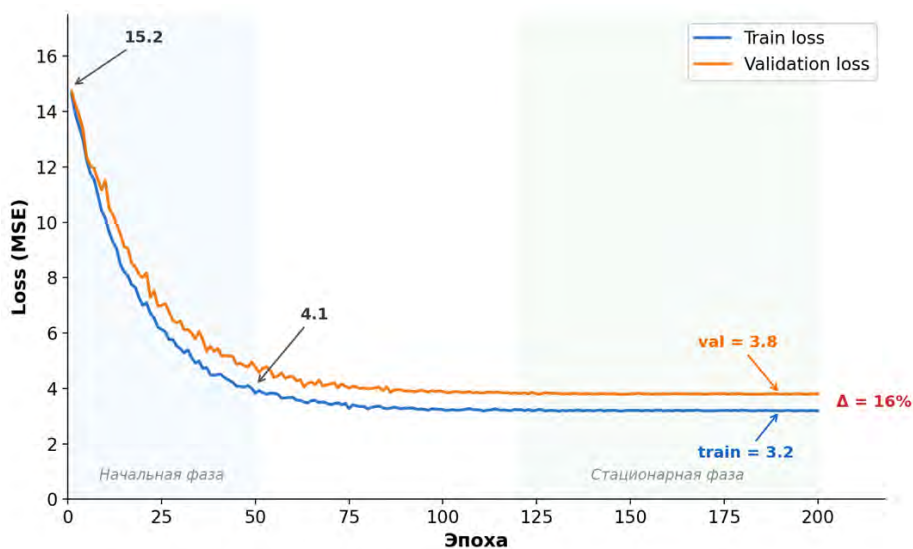


Рисунок 20 – Кривые обучения модели GAT

Наиболее интенсивное снижение функции потерь наблюдается в первые 50 эпох обучения. После этого скорость улучшения постепенно уменьшается, а начиная приблизительно со 120-й эпохи значения функции потерь на валидационной выборке стабилизируются. Низкая вариативность функции потерь на завершающих эпохах обучения свидетельствует о достижении сходимости модели.

Признаков существенного переобучения не выявлено. Значение функции потерь на обучающей выборке составляет 3,2, а на валидационной – 3,8, что соответствует разнице около 16%. Небольшой разрыв между значениями функции потерь на обучающей и валидационной выборках указывает на хорошую способность модели к обобщению.

Помимо качества прогнозирования, важной характеристикой является вычислительная эффективность моделей. В таблице 25 приведено сравнение рассматриваемых подходов по точности и числу обучаемых параметров.

Таблица 25 – Сравнение моделей по точности и числу параметров

| Модель        | Параметры | MAE |
|---------------|-----------|-----|
| MLP           | 2,8 млн   | 3,9 |
| GCN           | 3,2 млн   | 3,5 |
| Hybrid Fusion | 158 млн   | 3,1 |
| GAT           | 3,5 млн   | 2,7 |

Результаты показывают, что модель GAT достигает наименьшего значения MAE при сопоставимом числе параметров с другими графовыми моделями и существенно меньшей размерности по сравнению с Hybrid Fusion.

Модель GAT содержит 3,5 млн обучаемых параметров, тогда как Hybrid Fusion использует 158 млн параметров. Сочетание более высокой точности и меньшего числа параметров делает модель перспективной для практического применения.

Таким образом, разработанная архитектура Graph Attention Network реализует нейросимвольный подход к прогнозированию личностных характеристик на основе графа знаний. Модель использует механизм распространения информации по графу с вниманием и онтологическую регуляризацию, учитывающую психометрические взаимосвязи между факторами личности.

Экспериментальная проверка на выборке из 1722 пользователей показала улучшение качества прогнозирования по сравнению с базовыми моделями. Значение MAE снизилось до 2,7 против 3,1–3,9 у альтернативных подходов, а коэффициент корреляции Пирсона достиг 0,78.

Абляционный анализ показал существенный вклад механизма адаптивного обновления весов рёбер графа знаний. Увеличение MAE на 15 % при отключении данного механизма показывает, что адаптация весов связей графа знаний является важным компонентом предлагаемой архитектуры.

## 5.6 Валидация метрики GPI и примеры интерпретации

### 5.6.1 Примеры интерпретации предсказаний

Для иллюстрации работы предложенного метода интерпретации рассмотрим примеры объяснения предсказаний модели GAT на основе цепочек вывода, извлекаемых из графа знаний.

#### Пример 1. Пользователь 127 («PhilosopherStone»)

Для пользователя 127 модель предсказывает значение фактора M (Абстрактность), равное 8,9 из 10.

Результаты представлены в таблице 26.

Таблица 26 – Топ-5 путей объяснения для пользователя 127

| Ранг | Путь                                                                                                        | Вес пути                              | GPI |
|------|-------------------------------------------------------------------------------------------------------------|---------------------------------------|-----|
| 1    | User_127 → Feature_NicknameType_Literary → PsychConstruct_Factor_M                                          | $0,92 \times 0,38 = 0,35$             | 1,0 |
| 2    | User_127 → Feature_AvatarType_SingleMaleWithBooks → Feature_NicknameType_Literary → PsychConstruct_Factor_M | $0,87 \times 0,72 \times 0,38 = 0,24$ | 1,0 |
| 3    | User_127 → Feature_NicknameType_Literary → PsychConstruct_Factor_Q1 → PsychConstruct_Factor_M               | $0,92 \times 0,36 \times 0,44 = 0,15$ | 1,0 |
| 4    | User_127 → Feature_TemporalCluster_MorningActive → PsychConstruct_Factor_G → PsychConstruct_Factor_M        | $1,00 \times 0,51 \times 0,28 = 0,14$ | 1,0 |
| 5    | User_127 → Feature_StatusSentiment_Neutral → PsychConstruct_Factor_C → PsychConstruct_Factor_M              | $0,78 \times 0,12 \times 0,19 = 0,02$ | 1,0 |

Для объяснения предсказания извлекаются пять наиболее значимых путей от узла User\_127 к узлу Factor\_M\_Abstraction с использованием функции `extract_attention_paths` (Приложение А, листинг А.8).

Для наглядности те же объясняющие пути могут быть представлены содержательными утверждениями в терминах предметной области:

- поскольку никнейм пользователя относится к литературному типу, прогнозируется повышенное значение фактора М (мечтательность, развитое воображение) – вес пути 0,35;
- поскольку на аватаре присутствует одиночное фото с книгами, а никнейм относится к литературному типу, прогноз фактора М усиливается – вес пути 0,24;
- поскольку пользователь проявляет преимущественно утреннюю активность, связанную с высокой нормативностью поведения (фактор G), вывод косвенно поддерживает фактор М – вес пути 0,14.

Подобные формулировки делают вывод модели понятным специалисту и допускают экспертную проверку.

Анализ ранжированных путей показывает, что основной вклад в предсказание фактора М вносят признаки, связанные с семантикой никнейма и визуальными характеристиками профиля. Именно данные элементы графа знаний формируют наиболее значимые цепочки вывода, выявленные моделью.

Наибольший вес имеет прямой путь через признак литературного типа никнейма. Дополнительный вклад вносят пути, проходящие через визуальные признаки профиля и связанные психологические конструкты. При этом все выявленные пути характеризуются максимальным значением метрики  $GPI = 1,0$ , что свидетельствует о наличии непрерывной объясняющей цепочки от исходных признаков до прогнозируемой характеристики.

Таким образом, предсказание фактора М (Абстрактность) формируется преимущественно на основе взаимосвязанных текстовых и визуальных признаков, представленных в графе знаний пользователя. Тем самым итоговое значение фактора сопоставляется с конкретными элементами цифрового профиля, участвующими в формировании предсказания.

На основе извлечённых путей вывода система формирует текстовое объяснение предсказания на естественном языке.

1. Никнейм «PhilosopherStone» был отнесён к категории литературных референсов с вероятностью 0,92. В графе знаний данный признак связан с фактором М (Абстрактность) весом 0,38 ( $p < 0,0001$ ). Анализ обучающей выборки выявил положительную корреляцию между данным типом никнеймов и фактором М ( $r = 0,38$ ).

2. Визуальный анализ аватара выявил признаки, связанные с категорией «книги», с вероятностью 0,87. Данный признак участвует во втором по значимости пути вывода и вносит дополнительный вклад в формирование предсказания фактора М.

3. Признак NicknameType\_Literary также связан с фактором Q1 (Открытость к новому), который, в свою очередь, коррелирует с фактором М ( $r = 0,44$ ). Это формирует дополнительную цепочку подтверждения внутри графа знаний.

4. Кластер временной активности MorningActive связан с фактором G (Нормативность), который имеет умеренную связь с фактором М ( $r = 0,28$ ). Соответствующий путь характеризуется меньшим весом, однако также участвует в формировании итогового предсказания.

Структура объясняющего подграфа для пользователя User\_127 представлена на рисунке 21.

В представленном субграфе отображены наиболее значимые узлы и связи, участвующие в формировании предсказания фактора М (Абстрактность). Узел пользователя связан с признаками цифрового профиля, которые через промежуточные психологические конструкторы формируют объясняющие цепочки вывода к прогнозируемой характеристике.

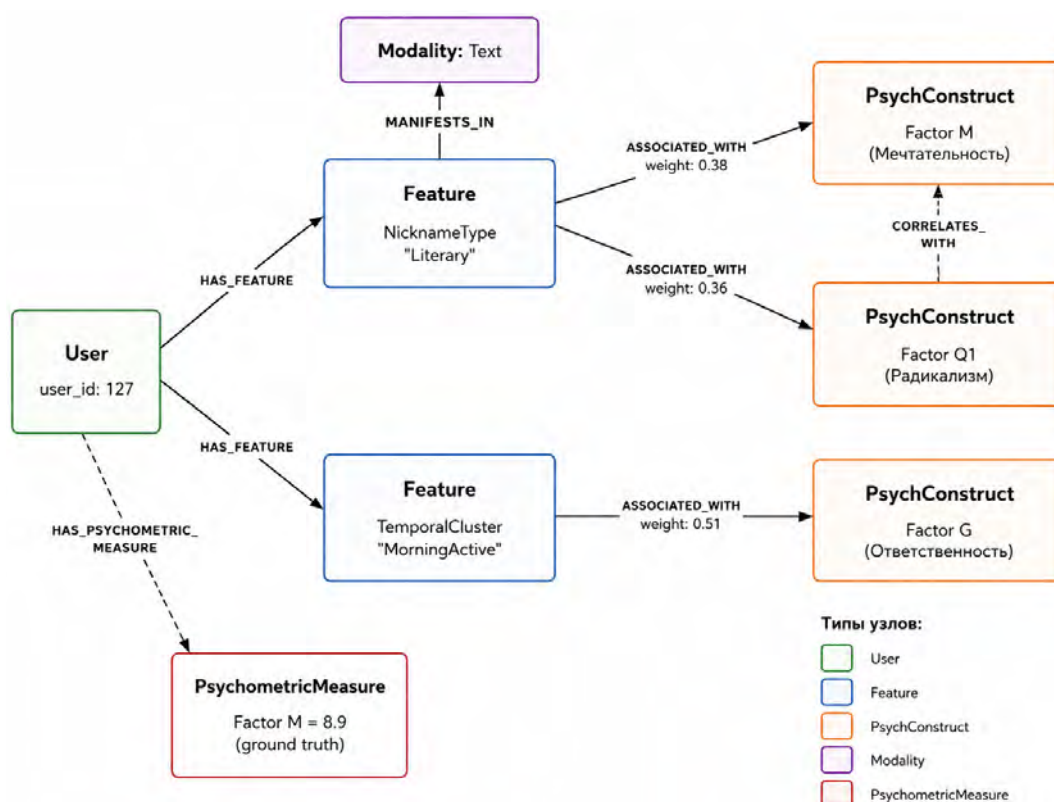


Рисунок 21 – Пример субграфа знаний для пользователя User\_127

Для данного примера значение метрики GPI составляет 1,0, что соответствует наличию непрерывных объясняющих цепочек между исходными признаками и

прогнозируемым фактором. Все выделенные пути проходят через семантически значимые связи графа знаний и участвуют в формировании итогового предсказания.

Полученное объяснение связывает итоговое значение фактора с конкретными признаками и связями графа знаний, использованными моделью при формировании предсказания.

#### Пример 2. Пользователь 239 («user847392»)

Для пользователя 239 модель предсказывает значение фактора Н (Социальная смелость), равное 6,8 из 10. Для объяснения предсказания были извлечены наиболее значимые пути вывода от узла User\_239 к узлу Factor\_H\_SocialBoldness. Результаты представлены в таблице 27.

Таблица 27 – Основные пути объяснения для пользователя 239

| Ранг | Путь                                                                                              | Вес пути                                   | Объяснение                                                   |
|------|---------------------------------------------------------------------------------------------------|--------------------------------------------|--------------------------------------------------------------|
| 1    | User_239 →<br>Feature_AvatarType_GroupPhoto →<br>PsychConstruct_Factor_H                          | $0,92 \times 0,72 =$<br>0,66               | Основной путь через признак группового изображения           |
| 2    | User_239 →<br>Feature_TemporalCluster_EveningActive →<br>PsychConstruct_Factor_H                  | $1,00 \times 0,18 =$<br>0,18               | Путь через характеристики временной активности               |
| 3    | User_239 → Feature_BehaviorType_Neutral<br>→ PsychConstruct_Factor_N →<br>PsychConstruct_Factor_H | $0,82 \times 0,26 \times$<br>$0,35 = 0,07$ | Косвенный путь через промежуточный психологический конструкт |

Анализ ранжированных путей показывает, что основной вклад в предсказание фактора Н (Социальная смелость) вносит признак AvatarType\_GroupPhoto, связанный с визуальными характеристиками цифрового профиля пользователя. Данный путь обладает наибольшим весом (0,66) и формирует основную часть объясняющего подграфа.

Дополнительный вклад вносят признаки временной активности пользователя и путь, проходящий через промежуточный психологический конструкт Factor\_N. Их влияние оказывается менее выраженным, однако также учитывается моделью при формировании итогового предсказания.

Для данного примера значение метрики GPI составляет 0,91, что соответствует высокой степени интерпретируемости результата. Большинство значимых компонентов предсказания может быть сопоставлено с конкретными признаками и связями графа знаний.

Полученное объяснение демонстрирует, каким образом визуальные признаки, характеристики активности пользователя и промежуточные психологические конструкты совместно участвуют в формировании итоговой оценки фактора Н.

### Пример 3. Пользователь 305 («SeriousEngineer»)

Для пользователя 305 модель предсказывает значение фактора Q1 (Открытость к новому), равное 5,2 из 10.

Для объяснения предсказания были извлечены наиболее значимые пути вывода от узла User\_305 к узлу Factor\_Q1\_Openness. Результаты представлены в таблице 28.

Таблица 28 – Основные пути объяснения для пользователя 305

| Ранг | Путь                                                                                             | Вес пути                                | Объяснение                                       |
|------|--------------------------------------------------------------------------------------------------|-----------------------------------------|--------------------------------------------------|
| 1    | User_305 → Feature_NicknameType_RealName → PsychConstruct_Factor_Q1                              | $0,88 \times 0,05 = 0,04$               | Путь через признак использования реального имени |
| 2    | User_305 → Feature_StatusSentiment_Negative → PsychConstruct_Factor_O → PsychConstruct_Factor_Q1 | $0,74 \times 0,18 \times -0,12 = -0,02$ | Косвенный путь через фактор O                    |
| 3    | User_305 → Feature_BehaviorType_Neutral → PsychConstruct_Factor_N → PsychConstruct_Factor_Q1     | $0,81 \times 0,26 \times -0,12 = -0,03$ | Косвенный путь через фактор N                    |

В отличие от предыдущих примеров, все выделенные пути характеризуются низкими значениями весов, что указывает на сравнительно слабую связь между наблюдаемыми признаками цифрового профиля и прогнозируемой характеристикой.

Для данного примера значение метрики GPI составляет 0,61, что соответствует сравнительно низкому уровню интерпретируемости по сравнению с предыдущими примерами.

Анализ объясняющего подграфа показывает отсутствие выраженных путей вывода: максимальный вес пути составляет всего 0,04. Выявленные признаки формируют слабые и частично разнонаправленные связи с прогнозируемой характеристикой, вследствие чего итоговое предсказание опирается на ограниченный объём интерпретируемой информации.

Полученное значение фактора Q1 = 5,2 близко к среднему значению по выборке, что соответствует ситуации высокой неопределённости модели относительно данной характеристики.

Рассмотренный пример демонстрирует одно из ограничений предложенного подхода. Для отдельных психологических факторов цифровой след пользователя может не содержать достаточного количества устойчивых табличных, текстовых или визуальных индикаторов. В подобных случаях качество интерпретации снижается, что отражается уменьшением значения метрики GPI и отсутствием выраженных объясняющих путей в графе знаний.

### 5.6.2 Валидация метрики интерпретируемости GPI

Для количественной оценки интерпретируемости предсказаний использовалась предложенная метрика GPI (Graph Path Interpretability). Валидация метрики выполнялась на тестовой выборке из 258 пользователей.

Анализ извлечённых объясняющих подграфов показал следующие характеристики:

- средняя длина объясняющего пути составляет 2,3 вершины, что соответствует преобладанию цепочек вида User → Feature → PsychConstruct;
- доля связей с весом внимания более 0,5 составляет 68 %, что указывает на концентрацию модели на наиболее информативных элементах графа;
- для 82 % пользователей (212 из 258) значение GPI превышает 0,7, что свидетельствует о наличии выраженных объясняющих цепочек для большинства предсказаний.

Результаты сравнения различных моделей представлены в таблице 29.

Таблица 29 – Характеристики интерпретируемости различных моделей

| Модель             | GPI ↑ | Средняя длина пути | Доля весов внимания > 0,5 | Согласие экспертов, %  |
|--------------------|-------|--------------------|---------------------------|------------------------|
| GAT (предлагаемая) | 0,82  | 2,3                | 68 %                      | 87 % (224 из 258)      |
| GCN                | 0,54  | 2,8                | 42 %                      | 61 % (157 из 258)      |
| MLP                | 0,00  | –                  | –                         | 28 % (72 из 258, SHAP) |

Полученные результаты показывают, что модель GAT обеспечивает наиболее высокий уровень интерпретируемости среди рассматриваемых подходов. Значение GPI для данной модели составляет 0,82, тогда как для GCN оно снижается до 0,54 вследствие отсутствия механизма внимания, а для модели MLP построение объясняющих путей в графе знаний невозможно.

Распределение значений метрики GPI по пользователям имеет следующий вид:

- высокая интерпретируемость ( $GPI \geq 0,8$ ) – 67% пользователей (173 случая);
- средняя интерпретируемость ( $0,6 \leq GPI < 0,8$ ) – 25% пользователей (65 случаев);
- низкая интерпретируемость ( $GPI < 0,6$ ) – 8% пользователей (20 случаев).

Таким образом, для большинства пользователей модель формирует объяснения, основанные на устойчивых цепочках связей в графе знаний. Пример случая с низким значением GPI рассмотрен в подпараграфе 5.5.1 (Пример 3).

Для дополнительной проверки корректности сформированных объяснений была проведена экспертная оценка. В качестве экспертов выступили три психолога со



стажем практической работы более 10 лет и опытом использования методики 16PF Кэттелла.

Эксперты оценивали автоматически сформированные объяснения по пятибалльной шкале. Критерии оценки приведены в таблице 30.

Таблица 30 – Шкала экспертной оценки интерпретируемости

| Оценка | Интерпретация                                                     |
|--------|-------------------------------------------------------------------|
| 1      | Объяснение неверное или не имеет содержательной интерпретации     |
| 3      | Объяснение частично соответствует психометрической интерпретации  |
| 5      | Объяснение полностью соответствует психометрической интерпретации |

Средние оценки интерпретируемости составили:

- GAT – 4,2 балла из 5 (87% объяснений получили оценку не ниже 4 баллов);
- GCN – 3,1 балла из 5 (61% объяснений получили оценку не ниже 4 баллов);
- MLP (с использованием SHAP-анализа) – 2,4 балла из 5 (28% объяснений получили оценку не ниже 4 баллов).

Для оценки согласованности метрики GPI с экспертным мнением был выполнен корреляционный анализ между значениями GPI и средними экспертными оценками. Полученное значение коэффициента корреляции Пирсона составило:

$$r = 0,76, p < 0,001.$$

Полученный результат указывает на сильную положительную связь между значениями GPI и экспертной оценкой качества объяснений.

Таким образом, предложенная метрика GPI согласуется с экспертными представлениями об интерпретируемости результатов и может использоваться для количественной оценки качества объяснений, формируемых графовой моделью.

Проведённая валидация показала, что модель GAT характеризуется наиболее высоким уровнем интерпретируемости среди рассмотренных подходов. Значение GPI для данной модели составляет 0,82, тогда как для GCN оно равно 0,54, а для модели MLP – 0,00. Наличие сильной корреляции между GPI и экспертными оценками подтверждает применимость метрики для анализа интерпретируемости предсказаний в графовых нейронных моделях.

### 5.6.3 Визуализация весов внимания

Для анализа процесса формирования предсказаний использовалась визуализация весов внимания, отражающих вклад отдельных признаков и связей графа знаний в результат работы модели.

На рисунке 22 представлена тепловая карта весов внимания для пользователя 127.

Наибольшие значения весов внимания соответствуют следующим связям:

- NicknameType\_Literary → Factor M (0,38);
- AvatarType\_GroupPhoto → Factor H (0,72);
- TemporalCluster\_MorningActive → Factor G (0,51).

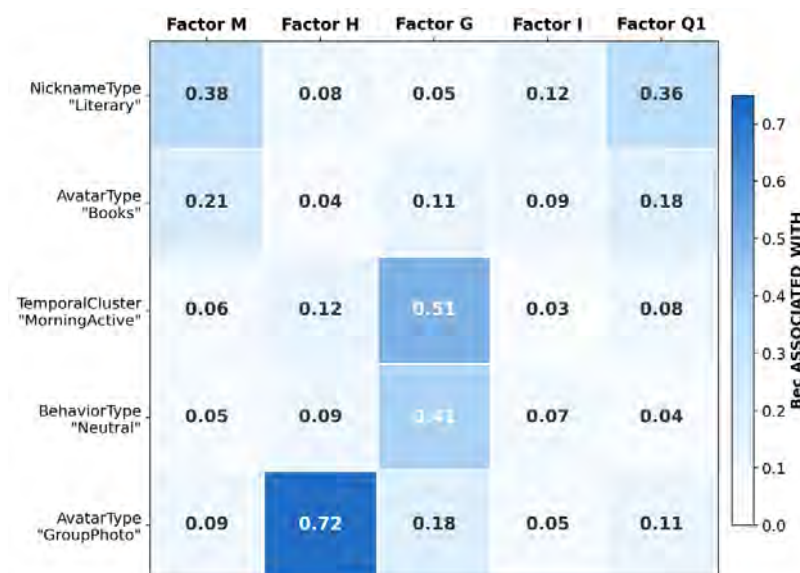


Рисунок 22 – Тепловая карта весов внимания признаков для пользователя 127

Полученные результаты согласуются с ранее рассмотренными примерами интерпретации и подтверждают, что механизм внимания концентрируется на наиболее информативных элементах графа знаний.

Анализ весов внимания показывает, что вклад признаков в формирование предсказания является неоднородным. Наибольшие значения весов соответствуют ограниченному числу связей графа знаний, которые формируют основную часть объясняющих путей. Это согласуется с результатами анализа интерпретации предсказаний, приведёнными ранее.

Использование тепловых карт и графовых представлений облегчает анализ предсказаний и позволяет сопоставлять итоговые значения психологических факторов с признаками цифрового профиля и связями графа знаний, участвующими в формировании результата.

Таким образом, интерпретация предсказаний модели осуществляется на основе анализа весов внимания и объясняющих путей графа знаний. Полученные результаты показывают, что модель GAT обеспечивает высокий уровень интерпретируемости ( $GPI = 0,82$ ) и формирует объяснения, согласующиеся как со структурой графа знаний, так и с экспертными оценками психологов.

## 5.7 Апробация системы на прикладной задаче

### 5.7.1 Экспериментальный сценарий: задача предварительного отбора профилей

В качестве прикладного сценария апробации предложенного подхода рассмотрена задача предварительного анализа цифровых профилей пользователей, соответствующая типовой задаче первичного отбора кандидатов. Данный сценарий позволяет оценить практическую применимость разработанной нейросимвольной архитектуры в условиях обработки множества объектов и принятия решений на основе результатов автоматизированного анализа.

Предложенный подход к решению данной задачи представлен автором в работе [55\*].

Для проведения эксперимента сформирована выборка из 50 цифровых профилей пользователей социальной сети. Для каждого профиля выполнен полный цикл обработки, включающий сбор данных, извлечение мультимодальных признаков, построение графа знаний и прогнозирование личностных характеристик с использованием модели Graph Attention Network.

Традиционно подобные задачи решаются посредством ручного анализа профилей, что требует значительных временных затрат со стороны экспертов. В связи с этим рассматриваемый сценарий является показательным для оценки возможностей автоматизации анализа цифрового следа.

#### Протокол эксперимента

1. Сформирована выборка из 50 пользователей, представленных ссылками на цифровые профили.

2. Для каждого объекта выполнен полный цикл обработки данных. Общее время обработки выборки составило около 20 минут при использовании параллельного режима выполнения. Основная часть времени была связана со сбором данных через API социальной сети. Среднее время последующего анализа одного профиля (извлечение признаков, построение графа знаний и выполнение предсказания) не превышало 4 секунд.

3. Для демонстрации возможности использования результатов анализа были заданы пороговые значения по ряду факторов личности:

- $V \geq 7$  – интеллект;
- $M \geq 7$  – абстрактность мышления;
- $Q1 \geq 6$  – открытость к новому;

- $Q3 \geq 6$  – самоконтроль;
- $Q2 \geq 5$  – самостоятельность.

В результате применения указанных критериев было отобрано 18 профилей, что составляет 36% исходной выборки.

Для оценки качества полученных результатов выполнено сопоставление прогнозов модели с результатами психометрического тестирования по методике 16PF Кэттелла. Валидационная выборка включала 28 пользователей: 18 профилей, удовлетворяющих условиям отбора, и 10 случайно выбранных профилей из числа не прошедших фильтрацию. Результаты сопоставления представлены в таблице 31.

Таблица 31 – Результаты валидации модели на основе сопоставления с результатами психометрического тестирования

| Метрика                                                  | Значение                   |
|----------------------------------------------------------|----------------------------|
| Коэффициент корреляции Пирсона (усреднённый по факторам) | $r = 0,68$ ( $p < 0,001$ ) |
| Средняя абсолютная ошибка (MAE)                          | 3,1 стена                  |
| Согласие по бинарной классификации (порог $\geq 7$ )     | 78 % ( $\kappa = 0,61$ )   |
| Ложноположительные срабатывания (False Positive)         | 3 из 18 (17 %)             |
| Ложноотрицательные срабатывания (False Negative)         | 2 из 10 (20 %)             |

Полученные результаты свидетельствуют о статистически значимой согласованности прогнозов модели с результатами психометрического тестирования. Значение коэффициента корреляции Пирсона  $r = 0,68$  и уровень согласия  $\kappa = 0,61$  указывают на способность модели выявлять пользователей, удовлетворяющих заданным критериям отбора.

В ходе эксперимента также была выполнена оценка временных затрат на обработку данных. При использовании автоматизированного анализа обработка выборки из 50 профилей была выполнена примерно за 20 минут, тогда как аналогичная процедура при ручном анализе требует существенно больших трудозатрат со стороны экспертов.

Полученные результаты демонстрируют возможность использования предложенного подхода для предварительного анализа и отбора цифровых профилей. Автоматизация обработки позволяет существенно сократить время анализа при сохранении согласованности результатов с психометрическими измерениями.

### 5.7.2 Анализ ошибок

В рамках валидационного эксперимента выполнен анализ случаев, в которых наблюдалось снижение качества прогнозирования. Рассмотренные ошибки могут быть разделены на несколько характерных групп.

### 1. Ограниченная доступность данных

Наиболее распространённой причиной снижения качества является недостаточная доступность данных цифрового профиля вследствие ограничений приватности. В подобных случаях объём доступной информации оказывается недостаточным для формирования полноценного признакового описания пользователя.

Данный тип ошибок связан с неполнотой входных данных и отражает ограничения применимости метода, а не особенности используемой модели. В подобных ситуациях система формирует предупреждение о сниженной достоверности результатов либо сообщает о невозможности выполнения анализа.

### 2. Недостаточность цифрового следа

Для профилей с минимальным количеством публикаций, отсутствием текстовых данных или визуального контента наблюдается снижение качества прогнозирования. При этом модель корректно отражает повышенную неопределённость результата посредством уменьшения показателя уверенности (среднее значение confidence  $\approx 0,42$  против 0,78 для полной выборки).

### 3. Искажения цифрового профиля

В отдельных случаях выявлены расхождения между результатами модели и психометрическими измерениями, обусловленные особенностями цифровой самопрезентации пользователей. Подобные ситуации возникают при наличии в профиле признаков, ассоциированных с определёнными личностными характеристиками, но не отражающих реальные особенности пользователя.

Поскольку модель анализирует наблюдаемый цифровой след, подобные искажения могут приводить к завышению или занижению отдельных факторов личности.

### 4. Ограничения обучающей выборки

Зафиксирован случай повышенной ошибки прогнозирования для пользователя, относящегося к культурной среде, недостаточно представленной в обучающей выборке. Данный результат указывает на чувствительность модели к особенностям цифрового поведения пользователей и подтверждает необходимость дальнейшей кросс-культурной валидации.

Проведённый анализ показывает, что основные источники ошибок связаны преимущественно с ограничениями входных данных и особенностями цифрового поведения пользователей. При этом в большинстве случаев модель корректно отражает

уровень неопределённости прогнозирования посредством снижения показателя уверенности.

### 5.7.3 Сравнение с альтернативными подходами

Для оценки полученных результатов выполнено сопоставление предложенного подхода с распространёнными методами анализа личностных характеристик по цифровому следу, представленными в научной литературе.

В качестве альтернативных подходов рассмотрены:

- методы, основанные на лингвистических признаках (LIWC);
- нейросетевые модели обработки последовательностей (CNN-LSTM);
- ансамбли трансформерных моделей.

Сравнение выполнялось по трём критериям: качество прогнозирования, вычислительная эффективность и интерпретируемость результатов.

Результаты сопоставления приведены в таблице 32.

Таблица 32 – Сравнение предложенного подхода с альтернативными методами

| Метод                                  | Pearson $r$ | Время обработки профиля | Интерпретируемость   | Ограничения                            |
|----------------------------------------|-------------|-------------------------|----------------------|----------------------------------------|
| Предлагаемый подход (граф знаний, GAT) | 0,78        | ~3–4 с                  | Высокая (GPI = 0,82) | Необходимость построения графа знаний  |
| LIWC                                   | 0,43        | ~1 с                    | Средняя              | Используются только текстовые данные   |
| CNN-LSTM                               | 0,58        | ~2 с                    | Низкая               | Ограниченные возможности интерпретации |
| Ансамбль трансформерных моделей        | 0,65        | ~8 с                    | Низкая               | Высокие вычислительные затраты         |

Сопоставление показывает, что предложенный подход обеспечивает наиболее высокое качество прогнозирования среди рассмотренных методов. При этом время обработки одного профиля остаётся приемлемым для практического применения.

Важным отличием разработанной архитектуры является наличие встроенного механизма интерпретации, основанного на графе знаний и механизме внимания. В отличие от большинства нейросетевых подходов, объяснение предсказания формируется в виде цепочек вывода, связывающих признаки цифрового профиля с прогнозируемыми характеристиками личности.

Методы, основанные исключительно на анализе текста, учитывают лишь часть цифрового следа пользователя и не используют визуальные и табличные данные. Современные нейросетевые модели обеспечивают более высокую точность

прогнозирования, однако, как правило, не предоставляют явных механизмов объяснения результатов.

Таким образом, предложенный подход сочетает высокое качество прогнозирования, умеренные вычислительные требования и возможность интерпретации получаемых результатов. Основным ограничением является необходимость построения и поддержки графа знаний, однако данное ограничение компенсируется возможностью получения объяснимых предсказаний и анализа причин их формирования.

## 5.8 Абляция онтологических ограничений

Для количественной оценки влияния онтологических ограничений на точность прогнозирования проведён абляционный эксперимент, изолирующий эффект информирования нейронной сети от эффектов мультимодального слияния. Эксперимент выполнен на упрощённом примере с одной (табличной) модальностью: использованы данные анкетирования 438 пользователей социальной сети «ВКонтакте» – 108 бинарных интерпретируемых признаков профиля (one-hot-кодирование 34 социально-демографических и поведенческих атрибутов); целевыми характеристиками служат 16 первичных факторов личности по Кэттеллу (стены 1–10). Выборка разбита в соотношении 70/15/15 (обучение/валидация/тест), результаты усреднены по трём запускам с различными начальными приближениями.

Во всех вариантах используется одна и та же графовая сеть внимания на графе «объект → признаки → характеристики» (архитектура и число параметров идентичны); варьируются только два механизма информирования нейронной сети. Топологический механизм: рёбра ассоциируются с включаются либо только для статистически значимых связей «признак – характеристика» (поправка Бенджамини–Хохберга,  $q = 0,05$ ; отобрано 46 рёбер из 1728 возможных), либо для всех пар (полный двудольный граф). Оптимизационный механизм: коэффициент  $\lambda$  при корреляционном члене  $L_{\text{corr}}$  функции потерь  $L = L_{\text{MSE}} + \lambda \cdot L_{\text{corr}}$  варьируется от 0 до 3. Вариант «полный граф,  $\lambda = 0$ » соответствует сети того же класса без онтологических ограничений.

Результаты приведены в таблице 33 и на рисунке 23. Точность оценивается корнем из среднеквадратичной ошибки на тестовой выборке (RMSE, в стенах); сохранение предметных взаимосвязей – средневзвешенным отклонением корреляционной структуры прогнозов от известной корреляционной структуры

характеристик ( $\Delta\rho$ , меньше – лучше); интерпретируемость – долей массы внимания модели, приходящейся на статистически значимые объясняющие пути ( $GPI_w$ , Определение 4).

Таблица 33 – Абляция онтологических ограничений: точность и интерпретируемость (438 профилей, 16 факторов Кэттелла, среднее по трём запускам)

| Вариант                                          | RMSE, стены ↓     | $R^2$ ↑ | $\Delta\rho$ ↓ | $GPI_w$ ↑ |
|--------------------------------------------------|-------------------|---------|----------------|-----------|
| Сеть без онтологии (полный граф, $\lambda = 0$ ) | $2,190 \pm 0,011$ | 0,049   | 0,468          | 0,033     |
| Онтологические рёбра, $\lambda = 0$              | $2,187 \pm 0,001$ | 0,047   | 0,195          | 1,000     |
| Онтологические рёбра, $\lambda = 0,03$           | $2,191 \pm 0,003$ | 0,042   | 0,161          | 1,000     |
| Онтологические рёбра, $\lambda = 0,1$            | $2,208 \pm 0,026$ | 0,030   | 0,154          | 1,000     |
| Онтологические рёбра, $\lambda = 0,3$            | $2,234 \pm 0,013$ | 0,008   | 0,159          | 1,000     |
| Онтологические рёбра, $\lambda = 1$              | $2,241 \pm 0,013$ | 0,001   | 0,163          | 1,000     |
| Онтологические рёбра, $\lambda = 3$              | $2,250 \pm 0,001$ | -0,005  | 0,232          | 1,000     |

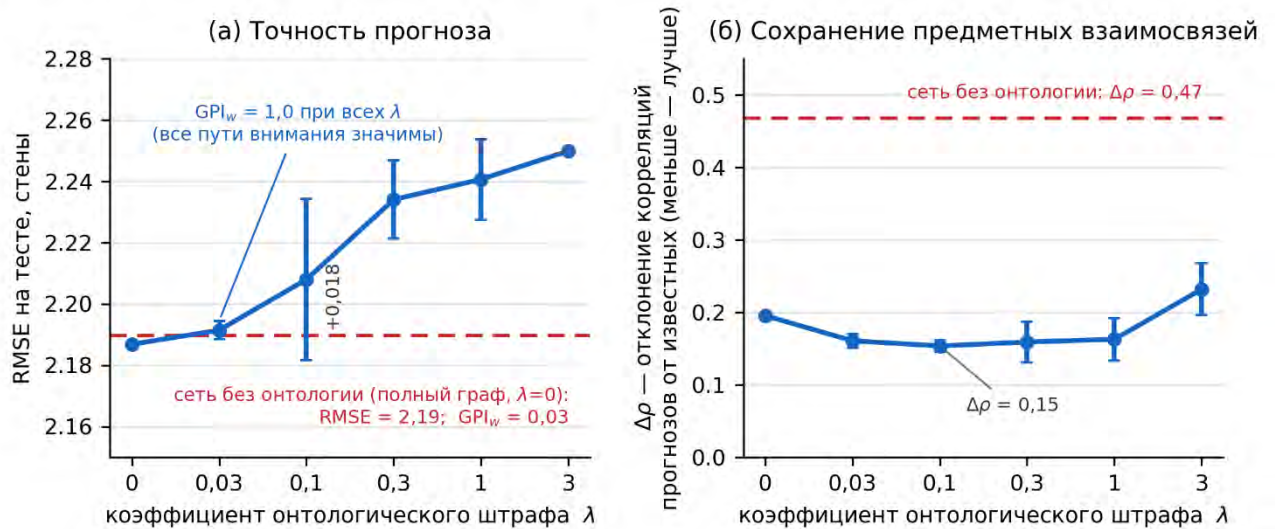


Рисунок 23 – Абляция онтологических ограничений

Топологический механизм информирования не снижает точность: ограничение графа только статистически значимыми рёбрами (46 из 1728) даёт RMSE 2,187 против 2,190 у полного графа – различие в пределах разброса запусков; отбор рёбер действует как регуляризация. При этом доля внимания, приходящаяся на значимые объясняющие пути, возрастает с 3,3% до 100%: интерпретируемость каждого решения гарантируется конструкцией модели.

Оптимизационный механизм ( $\lambda > 0$ ) создает управляемый компромисс: при  $\lambda = 0,03 \dots 0,1$  плата составляет 0,2–0,8 % RMSE, при этом отклонение корреляционной структуры прогнозов от известных предметных взаимосвязей сокращается дополнительно относительно варианта с онтологическими рёбрами при  $\lambda = 0$  ( $\Delta\rho$ : 0,195  $\rightarrow$  0,154; относительно полного графа без онтологии  $\Delta\rho$  снижается с 0,468 втрое уже самим топологическим ограничением). Дальнейшее увеличение  $\lambda$  избыточно: точность деградирует без роста согласованности (минимум  $\Delta\rho$  достигается при  $\lambda \approx 0,1$ ).



Таким образом, нейронная сеть того же класса без онтологических ограничений не даёт статистически значимого выигрыша в точности (различие не превышает 0,8% RMSE при  $\lambda \leq 0,1$ ), однако утрачивает согласованность прогнозов с предметными взаимосвязями ( $\Delta p$  хуже в три раза) и возможность построения значимых объясняющих путей ( $GPI\_w = 0,033$  против 1,0). Онтологические ограничения в рассматриваемой задаче являются не разменом точности на интерпретируемость, а приобретением встроенной интерпретируемости при сохранении точности.

Различие между коэффициентом детерминации  $R^2 = 0,61$  для полной модели (таблица 22) и  $R^2 \approx 0,05$  в абляционном эксперименте (таблица 33) обусловлено различием постановок, а не качеством модели. Абляционный эксперимент намеренно сужен для изоляции вклада онтологических ограничений: он выполнен на одной табличной модальности, 16 первичных факторах Кэттелла и уменьшенной выборке (438 профилей), поэтому низкий  $R^2$  в нём практически одинаков для вариантов с онтологией и без неё (0,047 против 0,049) и отражает предел одномодального предсказания, а не эффект онтологии. Лучший результат  $R^2 = 0,61$  получен полной мультимодальной моделью (текстовая, визуальная и табличная модальности) на 26 целевых характеристиках, включающих 6 интегральных факторов с более высокой предсказуемостью, в трансдуктивной постановке на полном графе из 1722 размеченных пользователей. Этим же объясняется превышение метааналитического диапазона  $R^2 = 0,29\text{--}0,40$ , установленного преимущественно для одномодальных текстовых моделей прогнозирования личностных характеристик.

### 5.9 Кросс-доменная проверка переносимости метода

Для проверки доменной инвариантности предложенных онтологического паттерна, метода и метрики GPI за пределами исходной задачи психологического профилирования выполнен перенос на три независимые предметные области: дерматоскопическую диагностику новообразований кожи, интерпретируемую классификацию токсичных текстов и классификацию эмоций. Во всех случаях инвариантное ядро метода (классы концептов, предикаты, правила реализации онтологии в граф знаний и подключения графовой сети внимания) применялось без изменения структуры; изменялись предметное наполнение онтологии, состав признаков и целевых характеристик. Полные постановки и результаты приведены в приложениях К и Л.

Дерматоскопическая диагностика. На наборе HAM10000 (10 015 дерматоскопических изображений, семь диагностических классов) слой признаков паттерна инстанцирован интерпретируемыми дескрипторами правила ABCD, а граф знаний «признак – диагноз» построен по статистически значимым точечно-бисериальным ассоциациям (поправка Бенджамини–Хохберга,  $q < 0,05$ ). При равной

точности рассуждающих модулей ( $\approx 0,67$ ) метрика GPI отделила графовое рассуждение с концептами от «чёрного ящика»: взвешенная согласованность путей GPI\_w составила 0,42 для GAT и 0,56 для GCN против 0,00 для MLP при покрытии GPI\_cov  $\approx 1,0$ ; тест верности удалением концептов подтвердил, что выделенные пути действительно определяют решения. Переносимость подтверждена внешней валидацией на независимом наборе ISIC 2018 (1 511 изображений из других источников), где воспроизведены как показатели классификатора, так и значимые рёбра графа знаний.

Классификация токсичных комментариев. На наборе HateXplain (20 148 комментариев) слой признаков инстанцирован интерпретируемыми категориями токсичности, а граф знаний «признак  $\rightarrow$  токсичность» построен по тем же правилам. Принципиальная особенность набора – эталонные обоснования решений, размеченные аннотаторами на уровне токенов, что позволило объективно проверить совпадение формируемых объяснений с признаками, на которые опирается человек. Дополнительно смена языка (русскоязычный психологический домен – англоязычная токсичность) подтвердила языковую независимость метода.

Классификация эмоций. На русскоязычном корпусе CEDR (Corpus for Emotion Detection in Russian; 7 528 сообщений обучающей и 1 882 сообщения тестовой выборки из LiveJournal и новостных лент) проверена работа полного онтологического паттерна в многоцелевой задаче с несколькими взаимосвязанными целевыми характеристиками. В отличие от предыдущих экспериментов, здесь задействованы все три предиката паттерна: имеет\_признак, ассоциируется\_с и коррелирует\_с, последний из которых описывает статистические взаимосвязи между эмоциями. Метрика GPI вновь отделила графовое рассуждение с концептами от «чёрного ящика»: взвешенная согласованность путей GPI\_w составила 0,96 для GAT против 0,18 для GCN и 0,00 для MLP при полном покрытии GPI\_cov = 1,0. Эксперимент подтвердил работоспособность полного онтологического паттерна со всеми тремя типами семантических связей и языковую независимость метода.

Результаты подтверждают, что доменно-инвариантными являются именно онтологический паттерн, нейросимвольный метод и метрика интерпретируемости GPI: их перенос между предметными областями, модальностями и языками не требует изменения структуры метода, при этом изменяются предметное наполнение онтологии, состав признаков и целевых характеристик. Обобщаемость предложенного подхода подтверждена экспериментально на трех дополнительных независимых доменах и внешнем наборе данных.

#### 5.10 Выводы по главе 5

В данной главе выполнена экспериментальная проверка разработанных методов и моделей.

1. Определена система метрик оценки качества прогнозирования, сохранения психометрической структуры и интерпретируемости результатов, а также методика статистической валидации, включающая проверку значимости различий и перекрёстную проверку.

2. Психометрическая валидация выявила статистически значимые связи между признаками цифрового следа и прогнозируемыми характеристиками личности, что подтверждает информативность используемых признаков.

3. Результаты мультимодальной интеграции показали преимущество совместного использования текстовых, визуальных и табличных данных по сравнению с одномодальными подходами. Анализ вкладов признаков с использованием SHAP подтвердил комплементарность различных модальностей.

4. Экспериментальное сравнение показало преимущество графовой нейронной сети GAT над базовыми моделями по показателям точности прогнозирования при сохранении интерпретируемости результатов. Тем самым результаты подтверждают применимость метода в системах поддержки принятия решений: интерпретируемый прогноз с объясняющими путями и оценкой GPI даёт лицу, принимающему решение, обоснование вывода, недоступное моделям типа «чёрный ящик».

5. Валидация метрики GPI показала её согласованность с экспертными оценками и продемонстрировала возможность количественной оценки интерпретируемости предсказаний на основе графа знаний.

6. Апробация системы на задаче предварительного анализа цифровых профилей продемонстрировала согласованность результатов модели с психометрическими измерениями и возможность практического применения разработанного подхода.

7. Совокупность проведённых экспериментов подтверждает работоспособность предложенной нейросимвольной архитектуры и её применимость для интерпретируемого анализа мультимодальных данных. Экспериментальная проверка переносимости разработанного подхода на задачу дерматоскопической диагностики новообразований кожи приведена в Приложении К.

8. Кросс-доменная проверка подтвердила переносимость метода: онтологический паттерн, нейросимвольная модель и метрика GPI применены без изменения структуры к дерматоскопической диагностике (HAM10000 с внешней валидацией на ISIC 2018) и интерпретируемой классификации токсичных текстов (HateXplain), где объяснения объективно согласованы с человеческими обоснованиями; тем самым доменная инвариантность паттерна, метода и метрики подтверждена экспериментально.

## ЗАКЛЮЧЕНИЕ

В диссертационной работе решена научная проблема разработки методов, моделей и архитектур интерпретируемого анализа типизированной мультимодальной информации на основе нейросимвольной интеграции графов знаний и графовых нейронных сетей. Предложенные методы апробированы на задаче формирования психологических профилей пользователей социальной сети «ВКонтакте». Их доменно-инвариантный характер и переносимость подтверждены экспериментальным применением в задачах интерпретируемой дерматоскопической диагностики, классификации токсичных комментариев и распознавания эмоций, результаты которых представлены в приложениях К и Л.

Основные результаты диссертационной работы заключаются в следующем.

1. Разработана гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации, интегрирующая нейросетевую модель с интерпретируемым графом знаний, построенным на основе онтологического паттерна «объект анализа – признак – целевая характеристика». Типы рёбер графа задаются семантическими предикатами предметной онтологии, их веса инициализируются статистически значимыми связями и уточняются в процессе обучения. Ограничение графа 46 статистически значимыми рёбрами из 1728 не изменяет ошибку предсказания в пределах разброса запусков (RMSE 2,187 против 2,190 для полного графа), при этом доля суммарного веса внимания, приходящаяся на статистически значимые объясняющие пути, возрастает с 0,033 до 1,0. Дополнительное онтологическое ограничение функции потерь снижает отклонение корреляционной структуры прогнозов от известных предметных взаимосвязей с 0,195 до 0,154 при увеличении RMSE на 0,2–0,8 %.

2. Разработан нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации со встроенным механизмом интерпретируемости: объяснение формируется непосредственно в процессе вывода в виде семантических путей графа знаний, энкодеры модальностей используются с фиксированными параметрами, обучению подлежит только графовый модуль. Получены MAE = 2,7, Pearson  $r$  = 0,78 и F1 = 0,85 при уменьшении числа обучаемых параметров со 158 млн до 3,5 млн относительно сквозной мультимодальной архитектуры.

3. Предложена метрика GPI (Graph Path Interpretability) для количественной оценки интерпретируемости результатов анализа по семантическим путям графа

знаний; верность объяснений проверяется методом удаления концептов. Для модели GAT получено  $GPI = 0,82$ ; для 82 % пользователей  $GPI$  превышает 0,7; корреляция метрики с экспертными оценками интерпретируемости составила  $r = 0,76$  при  $p < 0,001$ .

4. Разработана технология обработки и анализа типизированной мультимодальной информации с выделенным доменно-инвариантным ядром – онтологическим паттерном, структурой графа знаний, механизмом нейросимвольного вывода и метрикой  $GPI$ ; при переносе между предметными областями сохраняется инвариантное ядро метода, а изменяются предметное наполнение онтологии, состав признаков и целевых характеристик. Переносимость подтверждена в трёх независимых предметных областях: дерматоскопической диагностике новообразований кожи, интерпретируемой классификации токсичных комментариев и распознавании эмоций – при неизменном инвариантном ядре.

5. Разработана информационная система интерпретируемого анализа типизированной мультимодальной информации, реализующая полный цикл обработки данных от извлечения мультимодальных признаков и формирования графа знаний до нейросимвольного вывода и построения объясняющих путей. Экземплярный граф знаний содержит 63 161 узел и 777 365 рёбер; базовое время обработки одного профиля составляет 3-4 с, при кэшировании – около 1 с, пакетная обработка превышает 200 профилей в минуту; снижение MAE относительно базовых моделей составляет 13-31 % при  $p < 0,0001$ .

Кроме того, выполнена экспериментальная апробация разработанных модели, метода, метрики, технологии и программного комплекса на наборах данных, сформированных на основе более 31 тыс. пользовательских профилей, подтвердившая работоспособность подхода и его переносимость на предметные области дерматоскопической диагностики, классификации токсичных комментариев и распознавания эмоций.

Практическая значимость работы заключается в возможности использования разработанных нейросимвольных методов, технологии и программного комплекса в интеллектуальных информационных системах, предназначенных для анализа типизированной мультимодальной информации и содержательной интерпретации получаемых результатов.

Область применения результатов включает:

- интеллектуальный анализ мультимодальных данных в системах поддержки принятия решений;

- анализ цифровой активности пользователей социальных сетей и мультимедийных платформ;
- образовательная аналитика, мониторинг образовательных результатов и персонализация образовательных траекторий;
- HR-аналитика, оценка компетенций, профессиональный скрининг и поддержка кадровых решений;
- рекомендательные, экспертные и аналитические системы, использующие методы объяснимого искусственного интеллекта;
- медицинские, биометрические и другие интеллектуальные системы, в которых требуется интерпретация и обоснование принимаемых решений.

Внедрение результатов работы осуществлено в:

- ГБУ ДПО «Региональный социопсихологический центр» (г. Самара);
- ФГБУ «ЦЭНКИ» – Космодром «Восточный» (г. Циолковский);
- ПАО «Сбербанк России» (г. Самара);
- ООО «Профиль» (г. Тольятти);
- ООО «ФОТОН-БИО» (г. Черноголовка);
- АНО «Иннокуб» (г. Краснодар);
- ООО НПП «Интеллект Софт» (г. Самара);
- ООО «Международный образовательный центр «ОПТИМА СТАДИ»» (г. Самара);
- филиал РТРС Саратовский ОРТПЦ (г. Саратов);
- ООО «Самара-Сигнал» (г. Самара);
- учебный процесс Самарского университета (курсы «Системы искусственного интеллекта», «Нейронные сети», «ИТ-практикум») (г. Самара).

Перспективными направлениями дальнейшего развития предложенного подхода являются:

1. Развитие методов анализа динамических мультимодальных данных с учётом временной изменчивости объектов и адаптации графов знаний к изменениям предметной области.
2. Расширение состава анализируемых модальностей за счёт аудио-, видео- и данных о социальных взаимодействиях при сохранении единого онтологического представления и принципов нейросимвольной интеграции.

3. Совершенствование механизмов формирования интерпретируемых объяснений, включая использование больших языковых моделей для генерации объяснений на естественном языке и диалогового взаимодействия с пользователем.

4. Развитие методов конфиденциальной обработки мультимодальных данных на основе федеративного обучения и распределённых вычислений без централизации персональных данных.

В диссертационной работе разработан метод интерпретируемого анализа типизированной мультимодальной информации на основе гибридной нейросимвольной модели, объединяющей нейросетевое представление данных, предметную онтологию и граф знаний. Предложенный подход формализует совместное использование статистических закономерностей и предметных знаний и включает встроенное формирование интерпретируемых объяснений результатов анализа.

Полученные результаты вносят вклад в развитие методов объяснимого искусственного интеллекта, нейросимвольных систем и анализа мультимодальной информации и могут использоваться при создании интеллектуальных информационных систем, в которых требуется совместная обработка разнородных данных и содержательная интерпретация получаемых результатов.

## СПИСОК УСЛОВНЫХ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

|                                                        |                                                                                                                                                           |
|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| $o$                                                    | объект анализа                                                                                                                                            |
| $M(o) = \langle T, V, D \rangle$                       | мультимодальное представление объекта: текстовая ( $T$ ), визуальная ( $V$ ) и табличная (числовая) ( $D$ ) модальности; $D$ – данные цифровой активности |
| $c(o) = (c_1, \dots, c_d)$                             | интерпретируемое признаковое описание объекта; $c_i$ - значение $i$ -го признака                                                                          |
| $e_{\text{текст}}, e_{\text{визуал}}, e_{\text{табл}}$ | векторы признаков текстовой ( $\in \mathbb{R}^{768}$ ), визуальной ( $\in \mathbb{R}^{2048}$ ) и табличной модальностей                                   |
| $h_m$                                                  | представление модальности $m$ в общем латентном пространстве ( $\in \mathbb{R}^{512}$ )                                                                   |
| $W_m, b_m$                                             | матрица и вектор проекции модальности $m$ в общее пространство                                                                                            |
| $\lambda$                                              | обучаемый коэффициент адаптивного слияния модальностей, $\lambda \in [0, 1]$                                                                              |
| $G = (N, E)$                                           | граф знаний: множества узлов $N$ и рёбер $E$                                                                                                              |
| $F = \{f_i\}$                                          | узлы признаков графа знаний                                                                                                                               |
| $P = \{p_k\}$                                          | узлы целевых характеристик                                                                                                                                |
| $r$                                                    | точечно-бисериальный коэффициент корреляции (вес ребра ассоциируется <sub>c</sub> )                                                                       |
| $\rho$                                                 | коэффициент корреляции целевых характеристик (вес ребра коррелирует <sub>c</sub> )                                                                        |
| $\pi_{ik}$                                             | объясняющий путь «объект $\rightarrow$ признак $f_i \rightarrow$ характеристика $p_k$ »                                                                   |
| $w(\pi)$                                               | вес (значимость) объясняющего пути                                                                                                                        |
| $\tau$                                                 | порог присутствия признака                                                                                                                                |
| $\alpha_{ij}$                                          | коэффициент внимания графовой сети между узлами $i$ и $j$ (механизм GAT)                                                                                  |
| $\theta$                                               | обучаемые параметры графовой нейронной сети                                                                                                               |
| $\alpha$                                               | уровень значимости статистического критерия (в задачах отбора рёбер графа)                                                                                |
| $\alpha' = \alpha/m$                                   | уровень значимости с поправкой Бонферрони на $m$ проверок                                                                                                 |
| $q$                                                    | значение, скорректированное по процедуре Бенджамини– Хохберга (FDR)                                                                                       |
| $U, K$                                                 | число объектов и число целевых характеристик                                                                                                              |
| $u, k$                                                 | индексы объекта и целевой характеристики                                                                                                                  |
| $P_{uk}, \hat{P}_{uk}$                                 | истинное и прогнозируемое значение характеристики $k$ для объекта $u$                                                                                     |
| $p$                                                    | достигаемый уровень значимости (p-значение)                                                                                                               |
| GPI <sub>cov</sub> , GPI <sub>w</sub>                  | покрытие значимыми путями и взвешенная согласованность путей                                                                                              |
| СППР                                                   | система поддержки принятия решений                                                                                                                        |



|                |                                                                                 |
|----------------|---------------------------------------------------------------------------------|
| GNN            | графовая нейронная сеть (Graph Neural Network)                                  |
| GAT            | графовая сеть внимания (Graph Attention Network)                                |
| GCN            | графовая свёрточная сеть (Graph Convolutional Network)                          |
| MLP            | многослойный перцептрон (Multilayer Perceptron)                                 |
| GPI            | метрика интерпретируемости путей (Graph Path Interpretability)                  |
| RoBERTa        | трансформерная модель кодирования текста                                        |
| ResNet         | остаточная свёрточная нейронная сеть                                            |
| XGBoost        | градиентный бустинг над деревьями решений                                       |
| FWER           | групповая вероятность ошибки (Family-Wise Error Rate)                           |
| FDR            | ожидаемая доля ложных открытий (False Discovery Rate)                           |
| MAE, RMSE, MSE | средняя абсолютная, среднеквадратичная и квадратичная ошибки                    |
| R <sup>2</sup> | коэффициент детерминации                                                        |
| ETL            | извлечение, преобразование и загрузка данных (Extract–Transform–Load)           |
| API            | программный интерфейс приложения                                                |
| 16PF           | 16-факторный личностный опросник Кэттелла                                       |
| ЛПР            | лицо, принимающее решение                                                       |
| HCC            | нейросимвольная система                                                         |
| SHAP           | аддитивные объяснения Шепли (SHapley Additive exPlanations)                     |
| RDF            | модель описания ресурсов (Resource Description Framework)                       |
| OWL            | язык веб-онтологий (Web Ontology Language)                                      |
| REST API       | программный интерфейс на принципах передачи состояния представления             |
| ABCD           | правило дерматоскопии (Asymmetry, Border, Color, Diameter)                      |
| AUC            | площадь под ROC-кривой (Area Under the Curve)                                   |
| F1             | F-мера (гармоническое среднее точности и полноты)                               |
| ISIC           | архив дерматоскопических изображений (International Skin Imaging Collaboration) |

## СПИСОК ЛИТЕРАТУРЫ

1. Reinsel, D. The Digitization of the World: From Edge to Core / D. Reinsel, J. Gantz, J. Rydning // IDC White Paper. – November 2018. – 28 p.
2. Xu, P. Multimodal Learning with Transformers: A Survey / P. Xu, X. Zhu, D. A. Clifton // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, № 10. – P. 12113–12132. – DOI: 10.1109/TPAMI.2023.3275156.
3. Liang, P. P. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions / P. P. Liang, A. Zadeh, L.-P. Morency // ACM Computing Surveys. – 2024. – Vol. 56, № 10. – P. 1–42. – DOI: 10.1145/3656580.
4. Baltrusaitis, T. Multimodal Machine Learning: A Survey and Taxonomy / T. Baltrusaitis, C. Ahuja, L.-P. Morency // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41, № 2. – P. 423–443. – DOI: 10.1109/TPAMI.2018.2798607.
5. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
6. Liu, Y. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Y. Liu, M. Ott, N. Goyal [et al.] // arXiv. – 2019. – arXiv:1907.11692.
7. He, K. Deep Residual Learning for Image Recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 770–778.
8. Dosovitskiy, A. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov [et al.] // International Conference on Learning Representations (ICLR). – 2021.
9. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead / C. Rudin // Nature Machine Intelligence. – 2019. – Vol. 1. – P. 206–215. – DOI: 10.1038/s42256-019-0048-x.
10. Doshi-Velez, F. Towards a Rigorous Science of Interpretable Machine Learning / F. Doshi-Velez, B. Kim // arXiv. – 2017. – arXiv:1702.08608.
11. Garcez, A. d'A. Neurosymbolic AI: The 3rd Wave / A. d'A. Garcez, L. C. Lamb // Artificial Intelligence Review. – 2023. – Vol. 56, № 11. – P. 12387–12406. – DOI: 10.1007/s10462-023-10448-w.
12. De Raedt, L. From Statistical Relational to Neuro-Symbolic Artificial Intelligence / L. De Raedt, S. Dumančić, R. Manhaeve, G. Marra // Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. – 2020. – P. 4943–4950. – DOI: 10.24963/ijcai.2020/688.

13. Kipf, T. N. Semi-Supervised Classification with Graph Convolutional Networks / T. N. Kipf, M. Welling // International Conference on Learning Representations (ICLR). – 2017.
14. Veličković, P. Graph Attention Networks / P. Veličković [et al.] // International Conference on Learning Representations (ICLR). – 2018.
15. Ribeiro, M. T. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 1135–1144. – DOI: 10.1145/2939672.2939778.
16. Lundberg, S. M. A Unified Approach to Interpreting Model Predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765–4774.
17. Ngiam, J. Multimodal Deep Learning / J. Ngiam, A. Khosla, M. Kim [et al.] // Proceedings of the 28th International Conference on Machine Learning (ICML). – 2011. – P. 689–696.
18. Zadeh, A. Tensor Fusion Network for Multimodal Sentiment Analysis / A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. – Copenhagen : Association for Computational Linguistics, 2017. – P. 1103–1114. – DOI: 10.18653/v1/D17-1115.
19. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. – Minneapolis : Association for Computational Linguistics, 2019. – P. 4171–4186. – DOI: 10.18653/v1/N19-1423.
20. Kuratov, Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / Y. Kuratov, M. Arkhipov // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. – Moscow, 2019.
21. Tan, M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. V. Le // Proceedings of the 36th International Conference on Machine Learning. – 2019. – Vol. 97. – P. 6105–6114.
22. Sowa, J. F. Knowledge Representation: Logical, Philosophical, and Computational Foundations / J. F. Sowa. – Pacific Grove, CA : Brooks/Cole, 2000. – XIV, 594 p.
23. Bordes, A. Translating Embeddings for Modeling Multi-Relational Data / A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko // Advances in Neural Information Processing Systems. – 2013. – P. 2787–2795.

24. Yang, B. Embedding Entities and Relations for Learning and Inference in Knowledge Bases / B. Yang, S. W.-t. Yih, X. He, J. Gao, L. Deng // International Conference on Learning Representations (ICLR). – 2015.
25. Trouillon, T. Complex Embeddings for Simple Link Prediction / T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, G. Bouchard // Proceedings of the 33rd International Conference on Machine Learning. – 2016. – Vol. 48. – P. 2071–2080.
26. Wang, Q. Knowledge Graph Embedding: A Survey of Approaches and Applications / Q. Wang, Z. Mao, B. Wang, L. Guo // IEEE Transactions on Knowledge and Data Engineering. – 2017. – Vol. 29, № 12. – P. 2724–2743. – DOI: 10.1109/TKDE.2017.2754499.
27. Hamilton, W. L. Inductive Representation Learning on Large Graphs / W. L. Hamilton, Z. Ying, J. Leskovec // Advances in Neural Information Processing Systems. – 2017. – P. 1024–1034.
28. Schlichtkrull, M. Modeling Relational Data with Graph Convolutional Networks / M. Schlichtkrull, T. N. Kipf, P. Bloem [et al.] // The Semantic Web: ESWC 2018. – Cham : Springer, 2018. – Vol. 10843. – P. 593–607. – DOI: 10.1007/978-3-319-93417-4\_38.
29. Selvaraju, R. R. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization / R. R. Selvaraju, M. Cogswell, A. Das [et al.] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). – 2017. – P. 618–626. – DOI: 10.1109/ICCV.2017.74.
30. Arrieta, A. B. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI / A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser [et al.] // Information Fusion. – 2020. – Vol. 58. – P. 82–115. – DOI: 10.1016/j.inffus.2019.12.012.
31. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable / C. Molnar. – 2nd ed. – 2022. – 317 p. – ISBN 9798411463330.
32. Marcus, G. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence / G. Marcus // arXiv. – 2020. – arXiv:2002.06177.
- 33\*. Даниленко, А. Н. Автоматизированный психодиагностический комплекс в системе профориентационной деятельности вуза : монография / А. Н. Даниленко, В. В. Комаков, С. Г. Ихсанова. – Красноярск : Научно-инновационный центр, 2012. – 25 с.
- 34\*. Даниленко, А. Н. Многофакторная диагностика личности студента вуза как инструмент повышения качества образовательного мониторинга: опыт создания программного комплекса : монография / А. Н. Даниленко. – Москва : Издательство МАИ, 2013. – 16 с.

- 35\*. **Даниленко, А. Н.** Интеллектуальная поддержка принятия решений при оценивании человеческого фактора в сфере профессиональной деятельности : монография / А. Н. Даниленко. – Самара : Издательство СНЦ РАН, 2013. – 126 с.
- 36\*. **Жданова, А. Н.** Исследование эффективности различных архитектур рекуррентных нейронных сетей для решения задачи анализа тональности русскоязычных комментариев пользователей социальных сетей / А. Н. Жданова, А. В. Куприянов, А. А. Голова, А. С. Булгаков, Д. С. Баканов // Автометрия. – 2023. – Т. 59, № 4. – С. 29–38. – DOI: 10.15372/AUT20230404.
37. Оре, О. Теория графов / О. Оре ; пер. с англ. И. Н. Врублевской. – 2-е изд., стер. – Москва : Наука, 1980. – 336 с.
38. Харари, Ф. Теория графов / Ф. Харари ; пер. с англ. В. П. Козырева. – Москва : Мир, 1973. – 300 с.
39. Socher, R. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. – Seattle : Association for Computational Linguistics, 2013. – P. 1631–1642.
40. Хайкин, С. Нейронные сети: полный курс / С. Хайкин ; пер. с англ. – 2-е изд. – Москва : Издательский дом «Вильямс», 2006. – 1104 с.
41. Cover, T. M. Elements of Information Theory / T. M. Cover, J. A. Thomas. – 2nd ed. – Hoboken, NJ : John Wiley & Sons, 2006. – 784 p. – DOI: 10.1002/047174882X.
42. Клещев, А. С. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология» / А. С. Клещев, И. Л. Артемьева // Научно-техническая информация. Серия 2. Информационные процессы и системы. – 2001. – № 2. – С. 20–27.
43. Гаврилова, Т. А. Базы знаний интеллектуальных систем : учебник / Т. А. Гаврилова, В. Ф. Хорошевский. – Санкт-Петербург : Питер, 2000. – 384 с. – ISBN 5-272-00071-4.
44. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets / J. Demšar // Journal of Machine Learning Research. – 2006. – Vol. 7. – P. 1–30.
45. Poria, S. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion / S. Poria, E. Cambria, R. Bajpai, A. Hussain // Information Fusion. – 2017. – Vol. 37. – P. 98–125. – DOI: 10.1016/j.inffus.2017.02.003.
46. Gandomi, A. Beyond the Hype: Big Data Concepts, Methods, and Analytics / A. Gandomi, M. Haider // International Journal of Information Management. – 2015. – Vol. 35, № 2. – P. 137–144. – DOI: 10.1016/j.ijinfomgt.2014.10.007.
47. Hashem, I. A. T. The Rise of “Big Data” on Cloud Computing: Review and Open Research Issues / I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U.

- Khan // Information Systems. – 2015. – Vol. 47. – P. 98–115. – DOI: 10.1016/j.is.2014.07.006.
48. Lahat, D. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects / D. Lahat, T. Adalı, C. Jutten // Proceedings of the IEEE. – 2015. – Vol. 103, № 9. – P. 1449–1477. – DOI: 10.1109/JPROC.2015.2460697.
  49. Ramachandram, D. Deep Multimodal Learning: A Survey on Recent Advances and Trends / D. Ramachandram, G. W. Taylor // IEEE Signal Processing Magazine. – 2017. – Vol. 34, № 6. – P. 96–108.
  50. Gao, J. A Survey on Deep Learning for Multimodal Data Fusion / J. Gao, P. Li, Z. Chen, J. Zhang // Neural Computation. – 2020. – Vol. 32, № 5. – P. 829–864. – DOI: 10.1162/neco\_a\_01273.
  51. Conneau, A. Unsupervised Cross-lingual Representation Learning at Scale / A. Conneau, K. Khandelwal, N. Goyal [et al.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Online : Association for Computational Linguistics, 2020. – P. 8440–8451. – DOI: 10.18653/v1/2020.acl-main.747.
  52. Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean // Advances in Neural Information Processing Systems. – 2013. – Vol. 26. – P. 3111–3119.
  53. Pennington, J. GloVe: Global Vectors for Word Representation / J. Pennington, R. Socher, C. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – Doha : Association for Computational Linguistics, 2014. – P. 1532–1543. – DOI: 10.3115/v1/D14-1162.
  54. Pan, S. J. A Survey on Transfer Learning / S. J. Pan, Q. Yang // IEEE Transactions on Knowledge and Data Engineering. – 2010. – Vol. 22, № 10. – P. 1345–1359. – DOI: 10.1109/TKDE.2009.191.
  - 55\*. **Zhdanova, A. N.** Neuro-Symbolic Integration of Graph Attention Networks and Knowledge Graphs for Interpretable Personality Prediction from Digital Footprints / A. N. Zhdanova // Optical Memory and Neural Networks (Information Optics). – 2026. – Vol. 35, № 3. – P. 440–447.
  56. Socher, R. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. – Seattle : Association for Computational Linguistics, 2013. – P. 1631–1642.
  57. Rogers, A. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian / A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, A. Gribov // Proceedings of the 27th International Conference on Computational Linguistics. – Santa Fe, NM : Association for Computational Linguistics, 2018. – P. 755–763.

58. Szeliski, R. Computer Vision: Algorithms and Applications / R. Szeliski. – 2nd ed. – Cham : Springer, 2022. – XXII, 925 p. – DOI: 10.1007/978-3-030-34372-9.
59. Taigman, Y. DeepFace: Closing the Gap to Human-Level Performance in Face Verification / Y. Taigman, M. Yang, M. A. Ranzato, L. Wolf // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2014. – P. 1701–1708. – DOI: 10.1109/CVPR.2014.220.
60. Schroff, F. FaceNet: A Unified Embedding for Face Recognition and Clustering / F. Schroff, D. Kalenichenko, J. Philbin // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2015. – P. 815–823. – DOI: 10.1109/CVPR.2015.7298682.
61. Deng, J. ArcFace: Additive Angular Margin Loss for Deep Face Recognition / J. Deng, J. Guo, N. Xue, S. Zafeiriou // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2019. – P. 4690–4699. – DOI: 10.1109/CVPR.2019.00482.
62. Zhang, K. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks / K. Zhang, Z. Zhang, Z. Li, Y. Qiao // IEEE Signal Processing Letters. – 2016. – Vol. 23, № 10. – P. 1499–1503. – DOI: 10.1109/LSP.2016.2603342.
63. Deng, J. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild / J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 5202–5211. – DOI: 10.1109/CVPR42600.2020.00525.
- 64\*. **Жданова, А. Н.** Анализ семантического разрыва при классификации изображений по функциональному назначению / А. Н. Жданова, А. В. Куприянов // Компьютерная оптика. – 2026. – Т. 50, № 5. – 2079. – DOI: 10.18287/COJ2079.
65. Ismail Fawaz, H. Deep Learning for Time Series Classification: A Review / H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller // Data Mining and Knowledge Discovery. – 2019. – Vol. 33, № 4. – P. 917–963. – DOI: 10.1007/s10618-019-00619-1.
66. Hochreiter, S. Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – Vol. 9, № 8. – P. 1735–1780.
67. Chen, T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794. – DOI: 10.1145/2939672.2939785.
68. Ke, G. LightGBM: A Highly Efficient Gradient Boosting Decision Tree / G. Ke, Q. Meng, T. Finley [et al.] // Advances in Neural Information Processing Systems. – 2017. – P. 3146–3154.

69. Li, S. Multimodal Alignment and Fusion: A Survey / S. Li, H. Tang // International Journal of Computer Vision. – 2026. – Vol. 134. – Art. 103. – DOI: 10.1007/s11263-025-02667-1.
70. Kosinski, M. Private Traits and Attributes Are Predictable from Digital Records of Human Behavior / M. Kosinski, D. Stillwell, T. Graepel // Proceedings of the National Academy of Sciences of the United States of America. – 2013. – Vol. 110, № 15. – P. 5802–5805. – DOI: 10.1073/pnas.1218772110.
71. Kosinski, M. Computational Psychometrics / M. Kosinski, D. Stillwell, T. Graepel // The Cambridge Handbook of Personality Psychology. – 2nd ed. – Cambridge : Cambridge University Press, 2020. – P. 713–728.
72. Azucar, D. Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis / D. Azucar, D. Marengo, M. Settanni // Personality and Individual Differences. – 2018. – Vol. 124. – P. 150–159. – DOI: 10.1016/j.paid.2017.12.018.
- 73\*. **Zhdanova, A. N.** A Graph Path Interpretability Metric for Explainable Classification of Dermoscopic Skin Lesions / A. N. Zhdanova // Journal of Biomedical Photonics & Engineering. – 2026. – Vol. 12, № 3. – Art. 030304.
74. Schwartz, H. A. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach / H. A. Schwartz, J. C. Eichstaedt, M. L. Kern [et al.] // PLOS ONE. – 2013. – Vol. 8, № 9. – e73791. – DOI: 10.1371/journal.pone.0073791.
75. Park, G. Automatic Personality Assessment through Social Media Language / G. Park, H. A. Schwartz, J. C. Eichstaedt [et al.] // Journal of Personality and Social Psychology. – 2015. – Vol. 108, № 6. – P. 934–952. – DOI: 10.1037/pspp0000020.
76. Загорулько, Ю. А. Подход к построению порталов научных знаний / Ю. А. Загорулько, О. И. Боровикова // Автометрия. – 2008. – Т. 44, № 1. – С. 100–110.
77. Farnadi, G. Computational Personality Recognition in Social Media / G. Farnadi, G. Sitaraman, S. Sushmita [et al.] // User Modeling and User-Adapted Interaction. – 2016. – Vol. 26, № 2–3. – P. 109–142. – DOI: 10.1007/s11257-016-9171-0.
78. Liu, L. Analyzing Personality through Social Media Profile Picture Choice / L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. E. Moghaddam, L. H. Ungar // Proceedings of the 10th International AAAI Conference on Web and Social Media. – 2016. – P. 211–220.
79. Segalin, C. The Pictures We Like Are Our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits / C. Segalin, A. Perina, M. Cristani, A. Vinciarelli // IEEE Transactions on Affective Computing. – 2017. – Vol. 8, № 2. – P. 268–285. – DOI: 10.1109/TAFFC.2016.2516994.
80. Tufekci, Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls / Z. Tufekci // Proceedings of the International AAAI



- Conference on Web and Social Media. – 2014. – Vol. 8, № 1. – P. 505–514. – DOI: 10.1609/icwsm.v8i1.14517.
81. Lahat, D. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects / D. Lahat, T. Adalı, C. Jutten // *Proceedings of the IEEE*. – 2015. – Vol. 103, № 9. – P. 1449–1477. – DOI: 10.1109/JPROC.2015.2460697.
  82. Stein, B. E. *The Merging of the Senses* / B. E. Stein, M. A. Meredith. – Cambridge, MA : The MIT Press, 1993. – 232 p.
  83. Ernst, M. O. Merging the Senses into a Robust Percept / M. O. Ernst, H. H. Bühlhoff // *Trends in Cognitive Sciences*. – 2004. – Vol. 8, № 4. – P. 162–169. – DOI: 10.1016/j.tics.2004.02.002.
  84. Рубинштейн, С. Л. *Основы общей психологии* / С. Л. Рубинштейн. – Санкт-Петербург : Питер, 2002. – 720 с.
  85. Ананьев, Б. Г. *Человек как предмет познания* / Б. Г. Ананьев. – 3-е изд. – Санкт-Петербург : Питер, 2001. – 288 с.
  86. Леонтьев, А. Н. *Деятельность. Сознание. Личность : учебное пособие* / А. Н. Леонтьев. – 2-е изд., стер. – Москва : Смысл : Академия, 2005. – 352 с.
  87. Выготский, Л. С. *Мышление и речь* / Л. С. Выготский. – 5-е изд., испр. – Москва : Лабиринт, 1999. – 352 с.
  88. Соколов, Е. Н. *Очерки по психофизиологии сознания* / Е. Н. Соколов. – Москва : МГУ, 2010. – 255 с.
  89. *Психофизиология : учебник для вузов* / под ред. Ю. И. Александрова. – 4-е изд., перераб. и доп. – Санкт-Петербург : Питер, 2014. – 464 с.
  90. Величковский, Б. М. *Когнитивная наука: основы психологии познания : в 2 т. Т. 1* / Б. М. Величковский. – Москва : Смысл : Академия, 2006. – 448 с.
  91. Ушаков, Д. В. *Психология интеллекта и одаренности* / Д. В. Ушаков. – Москва : Институт психологии РАН, 2011. – 464 с.
  92. Петровский, А. В. *Психология : учебник* / А. В. Петровский, М. Г. Ярошевский. – 2-е изд., стер. – Москва : Академия, 2001. – 512 с.
  93. Бодалев, А. А. *Психология общения : избранные психологические труды* / А. А. Бодалев. – 2-е изд. – Москва : МПСИ ; Воронеж : НПО «МОДЭК», 2002. – 256 с.
  94. Леонтьев, Д. А. *Психология смысла: природа, строение и динамика смысловой реальности* / Д. А. Леонтьев. – 2-е изд., испр. – Москва : Смысл, 2003. – 487 с.
  95. Cattell, R. B. *Handbook for the Sixteen Personality Factor Questionnaire (16 PF): In Clinical, Educational, Industrial, and Research Psychology, for Use with All Forms of the Test* / R. B. Cattell, H. W. Eber, M. M. Tatsuoaka. – Champaign, IL : Institute for Personality and Ability Testing, 1970. – XXV, 388 p.

96. Cattell, R. B. The Scientific Analysis of Personality / R. B. Cattell. – Baltimore : Penguin Books, 1965. – 399 p.
97. Капустина, А. Н. Многофакторная личностная методика Р. Кеттелла / А. Н. Капустина. – Санкт-Петербург : Речь, 2001. – 112 с.
98. Шмелев, А. Г. Психодиагностика личностных черт / А. Г. Шмелев. – Санкт-Петербург : Речь, 2002. – 480 с.
99. Tausczik, Y. R. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods / Y. R. Tausczik, J. W. Pennebaker // Journal of Language and Social Psychology. – 2010. – Vol. 29, № 1. – P. 24–54. – DOI: 10.1177/0261927X09351676.
100. Clark, K. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators / K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning // International Conference on Learning Representations (ICLR). – 2020.
101. Conneau, A. Cross-lingual Language Model Pretraining / A. Conneau, G. Lample // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 7057–7067.
102. Snoek, C. G. M. Early versus Late Fusion in Semantic Video Analysis / C. G. M. Snoek, M. Worring, A. W. M. Smeulders // Proceedings of the 13th ACM International Conference on Multimedia. – 2005. – P. 399–402. – DOI: 10.1145/1101149.1101236.
103. Bengio, Y. Decision Trees Do Not Generalize to New Variations / Y. Bengio, O. Delalleau, C. Simard // Computational Intelligence. – 2010. – Vol. 26, № 4. – P. 449–467. – DOI: 10.1111/j.1467-8640.2010.00366.x.
104. Wang, W. What Makes Training Multi-Modal Classification Networks Hard? / W. Wang, D. Tran, M. Feiszli // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2020. – P. 12695–12705.
105. Atrey, P. K. Multimodal Fusion for Multimedia Analysis: A Survey / P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli // Multimedia Systems. – 2010. – Vol. 16, № 6. – P. 345–379. – DOI: 10.1007/s00530-010-0182-0.
106. Lu, J. Hierarchical Question-Image Co-Attention for Visual Question Answering / J. Lu, J. Yang, D. Batra, D. Parikh // Advances in Neural Information Processing Systems. – 2016. – Vol. 29. – P. 289–297.
107. Tsai, Y.-H. H. Multimodal Transformer for Unaligned Multimodal Language Sequences / Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – Florence : Association for Computational Linguistics, 2019. – P. 6558–6569. – DOI: 10.18653/v1/P19-1656.
108. Lu, J. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks / J. Lu, D. Batra, D. Parikh, S. Lee // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

109. Tan, H. LXMERT: Learning Cross-Modality Encoder Representations from Transformers / H. Tan, M. Bansal // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – Hong Kong : Association for Computational Linguistics, 2019. – P. 5100–5111. – DOI: 10.18653/v1/D19-1514.
- 110\*. Коварцев, А. Н. Извлечение нечётких правил из числовых данных и структуризация базы знаний для задачи классификации / А. Н. Коварцев, А. Н. Даниленко // Перспективные информационные технологии для авиации и космоса (ПИТ-2010) : материалы международной конференции с элементами научной школы для молодёжи. – 2010. – С. 24–26.
- 111\*. Даниленко, А. Н. Использование нейронечётких сетей для решения задачи классификации / А. Н. Даниленко // Вестник Совета молодых учёных и специалистов. – 2012. – № 1 (1). – С. 64–77.
- 112\*. Даниленко, А. Н. Корреляционный анализ данных с помощью карт Кохонена для решения задачи профессионального подбора кадров / А. Н. Даниленко // В мире научных открытий. – 2010. – № 4 (10), ч. 11. – С. 35–37.
113. Arevalo Ovalle, J. E. Gated Multimodal Units for Information Fusion / J. E. Arevalo Ovalle, T. Solorio, M. Montes-y-Gómez, F. A. González // International Conference on Learning Representations (ICLR), Workshop Track. – 2017.
114. Ektefaie, Y. Multimodal Learning with Graphs / Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, M. Zitnik // Nature Machine Intelligence. – 2023. – Vol. 5. – P. 340–350. – DOI: 10.1038/s42256-023-00624-6.
115. Shi, C. A Survey of Heterogeneous Information Network Analysis / C. Shi, Y. Li, J. Zhang, Y. Sun, P. S. Yu // IEEE Transactions on Knowledge and Data Engineering. – 2017. – Vol. 29, № 1. – P. 17–37. – DOI: 10.1109/TKDE.2016.2598561.
116. Gilpin, L. H. Explaining Explanations: An Overview of Interpretability of Machine Learning / L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, L. Kagal // 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). – 2018. – P. 80–89. – DOI: 10.1109/DSAA.2018.00018.
117. Nagrani, A. Attention Bottlenecks for Multimodal Fusion / A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 14200–14213.
118. Radford, A. Learning Transferable Visual Models From Natural Language Supervision / A. Radford, J. W. Kim, C. Hallacy [et al.] // Proceedings of the 38th International Conference on Machine Learning. – 2021. – Vol. 139. – P. 8748–8763.
119. Shavrina, T. RuGPT-3: Russian GPT-3 Language Models / T. Shavrina [et al.] // Dialogue. – 2021.

120. Gemini Team. Gemini: A Family of Highly Capable Multimodal Models / Gemini Team // arXiv. – 2023. – arXiv:2312.11805.
121. Girdhar, R. ImageBind: One Embedding Space To Bind Them All / R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, I. Misra // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2023. – P. 15180–15190. – DOI: 10.1109/CVPR52729.2023.01457.
122. Michel, P. Are Sixteen Heads Really Better than One? / P. Michel, O. Levy, G. Neubig // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.
123. Hogan, A. Knowledge Graphs / A. Hogan, E. Blomqvist, M. Cochez [et al.] // ACM Computing Surveys. – 2021. – Vol. 54, № 4. – Art. 71. – DOI: 10.1145/3447772.
124. Quillian, M. R. Semantic Memory / M. R. Quillian // Semantic Information Processing / ed. by M. Minsky. – Cambridge, MA : MIT Press, 1968. – P. 227–270.
125. Minsky, M. A Framework for Representing Knowledge / M. Minsky // The Psychology of Computer Vision / ed. by P. H. Winston. – New York : McGraw-Hill, 1975. – P. 211–277.
126. Singhal, A. Introducing the Knowledge Graph: Things, Not Strings / A. Singhal // Official Google Blog. – 2012. – May 16.
127. Auer, S. DBpedia: A Nucleus for a Web of Open Data / S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives // The Semantic Web. – Berlin ; Heidelberg : Springer, 2007. – Vol. 4825. – P. 722–735. – DOI: 10.1007/978-3-540-76298-0\_52.
128. Speer, R. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge / R. Speer, J. Chin, C. Havasi // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. – 2017. – Vol. 31, № 1. – P. 4444–4451. – DOI: 10.1609/aaai.v31i1.11164.
129. Navigli, R. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network / R. Navigli, S. P. Ponzetto // Artificial Intelligence. – 2012. – Vol. 193. – P. 217–250. – DOI: 10.1016/j.artint.2012.07.001.
130. Nickel, M. A Review of Relational Machine Learning for Knowledge Graphs / M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich // Proceedings of the IEEE. – 2016. – Vol. 104, № 1. – P. 11–33. – DOI: 10.1109/JPROC.2015.2483592.
131. Wang, Z. Knowledge Graph Embedding by Translating on Hyperplanes / Z. Wang, J. Zhang, J. Feng, Z. Chen // Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. – 2014. – Vol. 28, № 1. – P. 1112–1119. – DOI: 10.1609/aaai.v28i1.8870.
132. Sun, Z. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space / Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang // International Conference on Learning Representations (ICLR). – 2019.

133. Ji, S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications / S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu // IEEE Transactions on Neural Networks and Learning Systems. – 2022. – Vol. 33, № 2. – P. 494–514. – DOI: 10.1109/TNNLS.2021.3070843.
134. Wu, Z. A Comprehensive Survey on Graph Neural Networks / Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu // IEEE Transactions on Neural Networks and Learning Systems. – 2021. – Vol. 32, № 1. – P. 4–24. – DOI: 10.1109/TNNLS.2020.2978386.
135. Zhou, J. Graph Neural Networks: A Review of Methods and Applications / J. Zhou, G. Cui, S. Hu [et al.] // AI Open. – 2020. – Vol. 1. – P. 57–81. – DOI: 10.1016/j.aiopen.2021.01.001.
136. Gilmer, J. Neural Message Passing for Quantum Chemistry / J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl // Proceedings of the 34th International Conference on Machine Learning. – 2017. – Vol. 70. – P. 1263–1272.
137. Hu, Z. Heterogeneous Graph Transformer / Z. Hu, Y. Dong, K. Wang, Y. Sun // Proceedings of The Web Conference 2020. – 2020. – P. 2704–2710. – DOI: 10.1145/3366423.3380027.
138. Schlichtkrull, M. Modeling Relational Data with Graph Convolutional Networks / M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling // The Semantic Web: ESWC 2018. – Cham : Springer, 2018. – Vol. 10843. – P. 593–607. – DOI: 10.1007/978-3-319-93417-4\_38.
139. Busbridge, D. Relational Graph Attention Networks / D. Busbridge, D. Sherburn, P. Cavallo, N. Y. Hammerla // arXiv. – 2019. – arXiv:1904.05811.
140. Wang, X. Heterogeneous Graph Attention Network / X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P. S. Yu // Proceedings of the 2019 World Wide Web Conference. – 2019. – P. 2022–2032. – DOI: 10.1145/3308558.3313562.
141. Park, J. Hyperbolic Heterogeneous Graph Transformer / J. Park, S. Han, H. Lee, W.-Y. Shin, S. Lim // arXiv. – 2026. – arXiv:2601.08251.
142. Li, X. ScaleGNN: Towards Scalable Graph Neural Networks via Adaptive High-order Neighboring Feature Fusion / X. Li, J. Qi, H. Liu [et al.] // arXiv. – 2025. – arXiv:2504.15920.
143. Jiang, L. A Survey of Heterogeneous Graph Neural Networks for Cybersecurity Anomaly Detection / L. Jiang, R. Ryan, Q. Li, N. Ferdosian // arXiv. – 2025. – arXiv:2510.26307.
144. Zhao, W. Multimodal Perception-Driven Decision-Making for Human-Robot Interaction: A Survey / W. Zhao, K. Gangaraju, F. Yuan // Frontiers in Robotics and AI. – 2025. – Vol. 12. – Art. 1604472. – DOI: 10.3389/frobt.2025.1604472.
145. Kostopoulos, G. A Survey of Multimodal Learning Analytics: Data, Methods, Systems, and Responsible Deployment / G. Kostopoulos, S. Kotsiantis, T. Panagiotakopoulos, A.

- Kameas // *Future Internet*. – 2026. – Vol. 18, № 3. – Art. 115. – DOI: 10.3390/fi18030115.
146. Wang, T. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions / T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, H. T. Shen // *Proceedings of the IEEE*. – 2024. – Vol. 112, № 11. – P. 1716–1754. – DOI: 10.1109/JPROC.2024.3525147.
  147. Zhang, K. Composed Multi-modal Retrieval: A Survey of Approaches and Applications / K. Zhang, J. Li, Z. Li [et al.] // *arXiv*. – 2025. – arXiv:2503.01334.
  148. Zheng, C. Multimodal Relation Extraction with Efficient Graph Alignment / C. Zheng, J. Feng, Z. Fu, Y. Cai, Q. Li, T. Wang // *Proceedings of the 29th ACM International Conference on Multimedia*. – 2021. – P. 5298–5306. – DOI: 10.1145/3474085.3476968.
  149. Bachiri, K. MM-HGNN: Multimodal Representation Learning Heterogeneous Graph Neural Network / K. Bachiri, A. Yahyaouy, M. Malek, N. Rogovschi // *International Journal of Computational Intelligence Systems*. – 2025. – Vol. 18. – Art. 178. – DOI: 10.1007/s44196-025-00820-9.
  150. Yan, J. Learning Contextualized Knowledge Structures for Commonsense Reasoning / J. Yan, M. Raman, A. Chan [et al.] // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. – Online : Association for Computational Linguistics, 2021. – P. 4038–4051. – DOI: 10.18653/v1/2021.findings-acl.354.
  151. Shirian, A. Heterogeneous Graph Learning for Acoustic Event Classification / A. Shirian, M. Ahmadian, K. Somandepalli, T. Guha // *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. – 2023. – P. 1–5. – DOI: 10.1109/ICASSP49357.2023.10095073.
  152. Xu, F. Hybrid Hypergraph Networks for Multimodal Sequence Data Classification / F. Xu, H. Wang, Y. Huang, D. Zhang, Z. Fan // *arXiv*. – 2025. – arXiv:2508.00926. – DOI: 10.48550/arXiv.2508.00926.
  153. Ammar, K. WeightedHGE: Weighted Heterogeneous Graph Embedding / K. Ammar, W. Inoubli, S. Zghal, E. Mephu Nguifo // *Procedia Computer Science*. – 2025. – Vol. 270. – P. 2949–2958. – DOI: 10.1016/j.procs.2025.09.419.
  154. Serengil, S. I. LightFace: A Hybrid Deep Face Recognition Framework / S. I. Serengil, A. Ozpinar // *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. – 2020. – P. 23–27. – DOI: 10.1109/ASYU50717.2020.9259802.
  155. Ji, Y. A Comprehensive Survey on Self-Interpretable Neural Networks / Y. Ji, Y. Sun, Y. Zhang [et al.] // *Proceedings of the IEEE*. – 2025. – Vol. 113, № 8. – P. 783–813. – DOI: 10.1109/JPROC.2025.3635153.
  156. Lipton, Z. C. The Mythos of Model Interpretability / Z. C. Lipton // *Queue*. – 2018. – Vol. 16, № 3. – P. 31–57. – DOI: 10.1145/3236386.3241340.

157. Goodman, B. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation” / B. Goodman, S. Flaxman // *AI Magazine*. – 2017. – Vol. 38, № 3. – P. 50–57. – DOI: 10.1609/aimag.v38i3.2741.
158. Guidotti, R. A Survey of Methods for Explaining Black Box Models / R. Guidotti [et al.] // *ACM Computing Surveys*. – 2019. – Vol. 51, № 5. – Art. 93.
159. Wang, Y. Gradient based Feature Attribution in Explainable AI: A Technical Review / Y. Wang, T. Zhang, X. Guo, Z. Shen // *arXiv*. – 2024. – arXiv:2403.10415.
160. Ancona, M. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks / M. Ancona, E. Ceolini, C. Öztireli, M. Gross // *International Conference on Learning Representations (ICLR)*. – 2018.
161. Kim, B. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) / B. Kim, M. Wattenberg, J. Gilmer [et al.] // *Proceedings of the 35th International Conference on Machine Learning*. – 2018. – Vol. 80. – P. 2668–2677.
162. Sundararajan, M. Axiomatic Attribution for Deep Networks / M. Sundararajan, A. Taly, Q. Yan // *Proceedings of the 34th International Conference on Machine Learning*. – 2017. – Vol. 70. – P. 3319–3328.
163. Wachter, S. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR / S. Wachter, B. Mittelstadt, C. Russell // *Harvard Journal of Law & Technology*. – 2017. – Vol. 31, № 2. – P. 841–887.
164. Rai, D. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models / D. Rai, Y. Zhou, S. Feng, A. Saparov, Z. Yao // *arXiv*. – 2024. – arXiv:2407.02646.
165. Vijayakumar, S. Interpretability in Activation Space Analysis of Transformers: A Focused Survey / S. Vijayakumar // *CEUR Workshop Proceedings*. – 2022. – Vol. 3318.
166. Nauta, M. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI / M. Nauta, J. Trienes, S. Pathak [et al.] // *ACM Computing Surveys*. – 2023. – Vol. 55, № 13s. – Art. 295. – P. 1–42. – DOI: 10.1145/3583558.
167. Adebayo, J. Sanity Checks for Saliency Maps / J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim // *Advances in Neural Information Processing Systems*. – 2018. – Vol. 31.
168. Hooker, S. A Benchmark for Interpretability Methods in Deep Neural Networks / S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim // *Advances in Neural Information Processing Systems*. – 2019. – Vol. 32. – P. 9737–9748.
169. Alvarez-Melis, D. On the Robustness of Interpretability Methods / D. Alvarez-Melis, T. S. Jaakkola // *ICML Workshop on Human Interpretability in Machine Learning*. – 2018.

170. Slack, D. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods / D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju // Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society. – 2020. – P. 180–186. – DOI: 10.1145/3375627.3375830.
171. Ying, R. GNNExplainer: Generating Explanations for Graph Neural Networks / R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 9240–9251.
172. Luo, D. Parameterized Explainer for Graph Neural Network / D. Luo, W. Cheng, D. Xu [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 19620–19631.
173. Schlichtkrull, M. S. Interpreting Graph Neural Networks for NLP with Differentiable Edge Masking / M. S. Schlichtkrull, N. De Cao, I. Titov // International Conference on Learning Representations (ICLR). – 2021.
174. Raj, K. Neurosymbolic Methods for Explainable Graph Neural Networks: A Survey / K. Raj, A. Mileo // CEUR Workshop Proceedings. – 2025. – Vol. 4125. – P. 11–20.
175. DeLong, L. N. Neurosymbolic AI for Reasoning Over Knowledge Graphs: A Survey / L. N. DeLong, R. Fernández Mir, J. D. Fleuriot // IEEE Transactions on Neural Networks and Learning Systems. – 2025. – Vol. 36, № 5. – P. 7822–7842. – DOI: 10.1109/TNNLS.2024.3420218.
176. Rudin, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges / C. Rudin, C. Chen, Z. Chen [et al.] // Statistics Surveys. – 2022. – Vol. 16. – P. 1–85. – DOI: 10.1214/21-SS133.
177. Lou, Y. Accurate Intelligible Models with Pairwise Interactions / Y. Lou [et al.] // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2013. – P. 623–631.
178. Wiegreffe, S. Attention is not not Explanation / S. Wiegreffe, Y. Pinter // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. – Hong Kong : Association for Computational Linguistics, 2019. – P. 11–20. – DOI: 10.18653/v1/D19-1002.
179. Koh, P. W. Concept Bottleneck Models / P. W. Koh, T. Nguyen, Y. S. Tang [et al.] // Proceedings of the 37th International Conference on Machine Learning. – 2020. – Vol. 119. – P. 5338–5348.
180. Sarker, M. K. Neuro-Symbolic Artificial Intelligence: Current Trends / M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler // AI Communications. – 2021. – Vol. 34, № 3. – P. 197–209. – DOI: 10.3233/AIC-210084.



181. Jain, S. Attention is not Explanation / S. Jain, B. C. Wallace // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Minneapolis : Association for Computational Linguistics, 2019. – P. 3543–3556. – DOI: 10.18653/v1/N19-1357.
182. Колесников, А. В. Технология разработки гибридных интеллектуальных систем : дис. ... д-ра техн. наук : 05.13.01 / А. В. Колесников. – Санкт-Петербург, 2002. – 435 с.
183. Комарцова, Л. Г. Исследование нейросетевых и гибридных методов и технологий в интеллектуальных системах поддержки принятия решений : дис. ... д-ра техн. наук : 05.13.11 / Л. Г. Комарцова. – Калуга, 2003. – 438 с.
184. Шалфеева, Е. А. Методы, модели и технология обеспечения жизнеспособности интеллектуальных систем с декларативными базами знаний : дис. ... д-ра техн. наук / Е. А. Шалфеева. – Владивосток : ИАПУ ДВО РАН, 2021. – 217 с.
185. Северин, Н. Н. Векторные модели графов в задачах машинного обучения на структурных данных : дис. ... канд. компьютерных наук / Н. Н. Северин. – Москва : НИУ ВШЭ, 2025.
186. Сенюшкин, Д. С. Мультимодальное машинное зрение : дис. ... канд. наук / Д. С. Сенюшкин. – Москва : НИУ ВШЭ, 2025. – 99 с.
187. Васильев, Ю. А. Исследование и разработка методов машинного обучения анализа выживаемости : дис. ... канд. наук / Ю. А. Васильев. – Москва : МГУ имени М. В. Ломоносова, 2024. – 142 с.
188. Zheng, K. CI-GNN: A Granger Causality-Inspired Graph Neural Network for Interpretable Brain Network-Based Psychiatric Diagnosis / K. Zheng, S. Yu, B. Chen // Neural Networks. – 2024. – Vol. 172. – Art. 106147.
189. Sengupta, P. FireGNN: Neuro-Symbolic Graph Neural Networks with Trainable Fuzzy Rules for Interpretable Medical Image Classification / P. Sengupta, I. Rekik // arXiv. – 2025. – arXiv:2509.10510.
190. Raj, K. A Neuro-Symbolic Approach to Enhance Interpretability of Graph Neural Network through the Integration of External Knowledge / K. Raj // Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM). – 2023.
191. Liu, L. Neural-Symbolic Reasoning over Knowledge Graphs: A Survey from a Query Perspective / L. Liu, Z. Wang, H. Tong // ACM SIGKDD Explorations Newsletter. – 2025. – Vol. 27, № 1. – P. 124–136.
192. Chang, H. Path-Based Explanation for Knowledge Graph Completion / H. Chang, J. Ye, A. Lopez Avila, J. Du, J. Li // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. – 2024. – P. 231–242.

193. Zhang, S. PaGE-Link: Path-based Graph Neural Network Explanation for Heterogeneous Link Prediction / S. Zhang, J. Zhang, X. Song [et al.] // Proceedings of the ACM Web Conference 2023. – 2023. – P. 3784–3793.
194. Das, R. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases Using Reinforcement Learning / R. Das, S. Dhuliawala, M. Zaheer, A. McCallum // International Conference on Learning Representations (ICLR). – 2018.
195. Guarino, N. What Is an Ontology? / N. Guarino, D. Oberle, S. Staab // Handbook on Ontologies. – 2nd ed. – Berlin : Springer, 2009. – P. 1–17.
196. Воронцов, К. В. Математические методы обучения по прецедентам (теория машинного обучения) : курс лекций / К. В. Воронцов. – 2011. – 141 с. – URL: <http://www.machinelearning.ru/wiki/images/6/6d/voron-ml-1.pdf>.
197. Diestel, R. Graph Theory / R. Diestel. – 5th ed. – Berlin : Springer, 2017. – 428 p.
198. Rossi, A. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis / A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo // ACM Transactions on Knowledge Discovery from Data. – 2021. – Vol. 15, № 2. – Art. 14. – P. 1–49. – DOI: 10.1145/3424672.
199. Zhao, F. Deep Multimodal Data Fusion / F. Zhao, C. Zhang, B. Geng // ACM Computing Surveys. – 2024. – Vol. 56, № 9. – Art. 216. – P. 1–36. – DOI: 10.1145/3649447.
200. Kärkkäinen, K. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age / K. Kärkkäinen, J. Joo // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). – 2021. – P. 1548–1558.
201. Studer, R. Knowledge Engineering: Principles and Methods / R. Studer, V. R. Benjamins, D. Fensel // Data & Knowledge Engineering. – 1998. – Vol. 25, № 1–2. – P. 161–197.
202. Klyne, G. Resource Description Framework (RDF): Concepts and Abstract Syntax / G. Klyne, J. J. Carroll // W3C Recommendation. – 2004. – URL: <https://www.w3.org/TR/rdf-concepts/>.
203. Hitzler, P. OWL 2 Web Ontology Language Primer / P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, S. Rudolph // W3C Recommendation. – 2012. – URL: <https://www.w3.org/TR/owl2-primer/>.
- 204\*. **Жданова, А. Н.** Доменно-независимый онтологический паттерн проектирования интерпретируемых нейросимвольных систем / А. Н. Жданова // Онтология проектирования. – 2026. – Т. 16, № 3 (61). – С. 491–501.
205. Watts, D. J. Collective Dynamics of “Small-World” Networks / D. J. Watts, S. H. Strogatz // Nature. – 1998. – Vol. 393, № 6684. – P. 440–442.
206. Ekman, P. Constants Across Cultures in the Face and Emotion / P. Ekman, W. V. Friesen // Journal of Personality and Social Psychology. – 1971. – Vol. 17, № 2. – P. 124–129.

207. Ekman, P. An Argument for Basic Emotions / P. Ekman // *Cognition & Emotion*. – 1992. – Vol. 6, № 3–4. – P. 169–200. – DOI: 10.1080/02699939208411068.
208. Roberts, B. W. The Rank-Order Consistency of Personality Traits from Childhood to Old Age: A Quantitative Review of Longitudinal Studies / B. W. Roberts, W. F. DelVecchio // *Psychological Bulletin*. – 2000. – Vol. 126, № 1. – P. 3–25. – DOI: 10.1037/0033-2909.126.1.3.
209. Youyou, W. Computer-Based Personality Judgments Are More Accurate Than Those Made by Humans / W. Youyou, M. Kosinski, D. Stillwell // *Proceedings of the National Academy of Sciences of the United States of America*. – 2015. – Vol. 112, № 4. – P. 1036–1040.
210. Goffman, E. *The Presentation of Self in Everyday Life* / E. Goffman. – New York : Doubleday, 1959. – 259 p.
- 211\*. **Жданова, А. Н.** Онтологический подход к предсказанию психологических характеристик пользователей социальных сетей на основе графов знаний / А. Н. Жданова // *Искусственный интеллект и принятие решений*. – 2025. – № 3. – С. 40–59. – DOI: 10.14357/20718594250304.
- 212\*. **Жданова, А. Н.** Двухуровневая стратегия обучения трансформерных моделей для классификации коротких текстов / А. Н. Жданова // *Информационные технологии*. – 2026. – Т. 32, № 10. – С. 519–524.
- 213\*. **Даниленко, А. Н.** Интеллектуальная поддержка принятия решений при оценивании человеческого фактора во время катастроф / А. Н. Даниленко // *Труды МАИ*. – 2013. – № 71. – С. 1–6.
214. Robinson, I. *Graph Databases* / I. Robinson, J. Webber, E. Eifrem. – 2nd ed. – Sebastopol, CA : O'Reilly Media, 2015. – 236 p.
- 215\*. **Даниленко, А. Н.** Решение задачи классификации при принятии управленческих решений в условиях нечёткости исходных данных с использованием гибридного нейронечёткого классификатора / А. Н. Даниленко, О. П. Солдатов // *Известия Самарского научного центра РАН*. – 2014. – Т. 16, № 4 (2). – С. 350–358.
216. Mikolov, T. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corrado, J. Dean // *arXiv*. – 2013. – arXiv:1301.3781.
217. Wolf, T. Transformers: State-of-the-Art Natural Language Processing / T. Wolf, L. Debut, V. Sanh [et al.] // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. – Online : Association for Computational Linguistics, 2020. – P. 38–45. – DOI: 10.18653/v1/2020.emnlp-demos.6.
218. Hendrycks, D. Gaussian Error Linear Units (GELUs) / D. Hendrycks, K. Gimpel // *arXiv*. – 2016. – arXiv:1606.08415.

219. Loshchilov, I. Decoupled Weight Decay Regularization / I. Loshchilov, F. Hutter // International Conference on Learning Representations (ICLR). – 2019.
220. Rothe, R. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks / R. Rothe, R. Timofte, L. Van Gool // International Journal of Computer Vision. – 2018. – Vol. 126. – P. 144–157. – DOI: 10.1007/s11263-016-0940-3.
221. Deng, J. ImageNet: A Large-Scale Hierarchical Image Database / J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2009. – P. 248–255. – DOI: 10.1109/CVPR.2009.5206848.
222. Goodfellow, I. J. Challenges in Representation Learning: A Report on Three Machine Learning Contests / I. J. Goodfellow, D. Erhan, P. L. Carrier [et al.] // Neural Networks. – 2015. – Vol. 64. – P. 59–63. – DOI: 10.1016/j.neunet.2014.09.005.
223. Landis, J. R. The Measurement of Observer Agreement for Categorical Data / J. R. Landis, G. G. Koch // Biometrics. – 1977. – Vol. 33, № 1. – P. 159–174. – DOI: 10.2307/2529310.
224. Bishop, C. M. Pattern Recognition and Machine Learning / C. M. Bishop. – New York : Springer, 2006. – 738 p.
225. Szegedy, C. Rethinking the Inception Architecture for Computer Vision / C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – P. 2818–2826. – DOI: 10.1109/CVPR.2016.308.
226. Simonyan, K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations (ICLR). – 2015.
227. Krizhevsky, A. ImageNet Classification with Deep Convolutional Neural Networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Advances in Neural Information Processing Systems. – 2012. – Vol. 25. – P. 1097–1105.
228. Rothe, R. DEX: Deep EXpectation of Apparent Age from a Single Image / R. Rothe, R. Timofte, L. Van Gool // Proceedings of the IEEE International Conference on Computer Vision Workshops. – 2015. – P. 252–257. – DOI: 10.1109/ICCVW.2015.41.
229. Loukachevitch, N. RuThes Linguistic Ontology vs. Russian Wordnets / N. Loukachevitch, B. Dobrov // Proceedings of the Seventh Global WordNet Conference. – Tartu : University of Tartu Press, 2014. – P. 154–162. – DOI: 10.18653/v1/W14-0121.
230. Jocher, G. YOLOv5 by Ultralytics [Computer software] / G. Jocher. – 2020. – DOI: 10.5281/zenodo.3908559.

231. Lin, T.-Y. Microsoft COCO: Common Objects in Context / T.-Y. Lin, M. Maire, S. Belongie [et al.] // Computer Vision – ECCV 2014. – Cham : Springer, 2014. – Vol. 8693. – P. 740–755. – DOI: 10.1007/978-3-319-10602-1\_48.
232. React: The Library for Web and Native User Interfaces [Электронный ресурс]. – URL: <https://react.dev> (дата обращения: 06.07.2026).
233. Bostock, M. D3: Data-Driven Documents / M. Bostock, V. Ogievetsky, J. Heer // IEEE Transactions on Visualization and Computer Graphics. – 2011. – Vol. 17, № 12. – P. 2301–2309. – DOI: 10.1109/TVCG.2011.185.
234. FastAPI: Fast (High-Performance) Web Framework for Building APIs with Python [Электронный ресурс]. – URL: <https://fastapi.tiangolo.com> (дата обращения: 06.07.2026).
235. Neo4j: Graph Database and Analytics [Электронный ресурс]. – URL: <https://neo4j.com> (дата обращения: 06.07.2026).
236. PlantUML: Open-Source Tool for Creating UML Diagrams [Электронный ресурс]. – URL: <https://plantuml.com> (дата обращения: 06.07.2026).
237. Mathew, B. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection / B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee // Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Vol. 35, № 17. – P. 14867–14875. – DOI: 10.1609/aaai.v35i17.17745.
238. Russian Language Toxic Comments: набор размеченных комментариев с ресурсов 2ch.hk и pikabu.ru [Электронный ресурс] // Kaggle. – 2019. – URL: <https://www.kaggle.com/datasets/blackmoon/russian-language-toxic-comments> (дата обращения: 15.07.2026).
239. Sboev, A. Data-Driven Model for Emotion Detection in Russian Texts / A. Sboev, A. Naumov, R. Rybka // Procedia Computer Science. – 2021. – Vol. 190. – P. 637–642. – DOI: 10.1016/j.procs.2021.06.075.
240. DeYoung, J. ERASER: A Benchmark to Evaluate Rationalized NLP Models / J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – Online : Association for Computational Linguistics, 2020. – P. 4443–4458. – DOI: 10.18653/v1/2020.acl-main.408.
241. Chen, X. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion / X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, H. Chen // Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York : ACM, 2022. – P. 904–915. – DOI: 10.1145/3477495.3531992.

242. Zhao, Y. MoSE: Modality Split and Ensemble for Multimodal Knowledge Graph Completion / Y. Zhao, X. Cai, Y. Wu, H. Zhang, Y. Zhang, G. Zhao, N. Jiang // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. – Abu Dhabi : Association for Computational Linguistics, 2022. – P. 10527–10536. – DOI: 10.18653/v1/2022.emnlp-main.719.
243. Mavromatis, C. GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning / C. Mavromatis, G. Karypis // arXiv. – 2024. – arXiv:2405.20139.
- 244\*. **Zhdanova, A. N.** Using Data Mining Algorithms to Solve the Problem of Predicting Personal Characteristics of a Person Based on the Analysis of Open Data from Social Networks / A. N. Zhdanova, A. V. Kupriyanov, A. A. Khoroshilova // 2021 International Conference on Information Technology and Nanotechnology (ITNT). – IEEE, 2021. – DOI: 10.1109/ITNT52450.2021.9649117.
- 245\*. **Даниленко, А. Н.** Автоматизированная информационная система психологического тестирования и анализа профессиональной пригодности «АИСТ» : свидетельство о государственной регистрации программы для ЭВМ № 2009612163 / А. Н. Даниленко, В. В. Комаков, О. П. Солдатова. – Зарег. 27.04.2009.
- 246\*. **Даниленко, А. Н.** Автоматизированная информационная система многофакторной диагностики профессионально значимых качеств личности на основе интерактивного перечня требований к профессии : свидетельство о государственной регистрации программы для ЭВМ № 2012610996 / А. Н. Даниленко, В. В. Комаков, С. Г. Ихсанова. – Зарег. 24.01.2012.
- 247\*. **Жданова, А. Н.** Программный модуль формирования мультимодального цифрового профиля на основе анализа цифровой активности : свидетельство о государственной регистрации программы для ЭВМ № 2026687279 / А. Н. Жданова. – Зарег. 07.09.2026.
- 248\*. **Жданова, А. Н.** Программный модуль графового вывода и интерпретируемого анализа на основе графа знаний и графовой нейронной сети : свидетельство о государственной регистрации программы для ЭВМ № 2026669561 / А. Н. Жданова. – Зарег. 09.07.2026.
- 249\*. Мудрова, К. Р. Исследование взаимосвязи онлайн-активности пользователя социальной сети с его поведенческими реакциями / К. Р. Мудрова, **А. Н. Жданова**, А. В. Куприянов // Информационные процессы. – 2024. – Т. 24, № 2. – С. 149–162.
- 250\*. **Даниленко, А. Н.** Поиск и оптимизация функции аппроксимации при анализе данных испытаний узлов ГТД / А. Н. Даниленко, Д. А. Попова-Коварцева // Известия Самарского научного центра РАН. – 2019. – Т. 21, № 1. – С. 150–155.
- 251\*. Коварцев, А. Н. Технология визуального моделирования параллельных алгоритмов. Язык PGRAPH / А. Н. Коварцев, В. В. Жидченко, Д. А. Попова-Коварцева, **А. Н.**

- Даниленко // Известия Самарского научного центра РАН. – 2019. – Т. 21, № 1. – С. 145–149.
- 252\*. Даниленко, А. Н. Разработка многопользовательского web-интерфейса системы как инструмента эффективного использования математических пакетов для решения инженерных задач / А. Н. Даниленко, И. А. Жданов // Вестник Самарского университета. Аэрокосмическая техника, технологии и машиностроение. – 2018. – Т. 17, № 2. – С. 183–190.
- 253\*. Danilenko, A. N. Developing Methods and Algorithms for a Decision-Making Intellectual Support in Personnel Management Systems / A. N. Danilenko // CEUR Workshop Proceedings. – 2015. – Vol. 1490. – P. 381–388. – DOI: 10.18287/1613-0073-2015-1490-381-388.
- 254\*. Danilenko, A. N. Intelligent Psycho-Diagnostic Forecasting System / A. N. Danilenko // CEUR Workshop Proceedings. – 2016. – Vol. 1638. – P. 527–535. – DOI: 10.18287/1613-0073-2016-1638-527-535.
- 255\*. Danilenko, A. N. Neural Network Prediction Model of the Pilots' Errors / A. N. Danilenko // CEUR Workshop Proceedings. – 2017. – Vol. 1904. – P. 296–299. – DOI: 10.18287/1613-0073-2017-1904-296-299.
- 256\*. Zhdanova, A. N. Morphological Text Analysis Using Neural Networks / A. N. Zhdanova, A. V. Kupriyanov, D. S. Sherenkov // 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT). – IEEE, 2022. – P. 1–4. – DOI: 10.1109/ITNT55410.2022.9848738.

## ПРИЛОЖЕНИЕ А Описание программной реализации нейросимвольной системы анализа мультимодальных данных

Настоящее приложение содержит полный исходный код программной реализации системы, разработанной в рамках диссертационного исследования. Приведены все основные модули: сбор данных, предобработка, извлечение мультимодальных признаков, построение графа знаний, графовый вывод на основе GAT, генерация объяснений, кэширование, управление моделями и REST API.

Код написан на Python 3.11+ с использованием библиотек: PyTorch 2.0.1, Transformers 4.33.0, DGL 1.1.2, Neo4j 5.11.0, vk\_api 11.9.9, XGBoost 1.7.6, FastAPI 0.103.0, Redis 7.2.0, ultralytics 8.0.147.

В таблице А.1 представлены модули системы и их назначение.

Таблица А.1 – Модули системы и их назначение

| Модуль                     | Вход                      | Выход                                         | Назначение                                                 |
|----------------------------|---------------------------|-----------------------------------------------|------------------------------------------------------------|
| А.1 Сбор данных            | ID пользователя           | Профиль, посты, фото, друзья, группы          | Сбор данных из VK API с учётом приватности                 |
| А.2 Предобработка          | Сырые данные              | Очищенные текст, изображения, метаданные      | Нормализация данных, фиксация пропусков                    |
| А.3 Текстовые признаки     | Никнейм, статус           | Категории и эмбединги                         | Классификация текстов (6 и 2 класса) через RoBERTa, RuBERT |
| А.4 Визуальные признаки    | Аватар                    | Возраст, пол, тип аватара, эмбединг           | Извлечение признаков через ResNet50/VGG16                  |
| А.5 Поведенческие признаки | Активность (посты, время) | Кластеры, XGBoost-оценки                      | Анализ временных паттернов и вовлечённости                 |
| А.6 Граф знаний            | Мультимодальные признаки  | Подграф User→Feature→PsychConstruct           | Построение семантического графа в Neo4j                    |
| А.7 Графовый вывод         | Граф и признаки узлов     | Прогноз по 26 факторам, веса внимания         | Предсказание через GAT с механизмом внимания               |
| А.8 Генерация объяснений   | Прогноз, веса внимания    | SHAP-вклады, пути в графе, текстовое описание | Интерпретация результата через графовые пути               |
| А.9 Кэширование            | Эмбединги                 | Кэшированные векторы                          | Хранение промежуточных представлений в Redis               |
| А.10 Управление моделями   | —                         | Загруженные модели                            | Предзагрузка моделей при старте                            |
| А.11 API                   | Запрос на анализ          | JSON с прогнозом и объяснением                | REST-интерфейс для вызова всего конвейера                  |

На рисунке А.1 представлена обобщённая архитектура системы, отражающая последовательность обработки данных от сбора до выдачи интерпретируемых прогнозов.



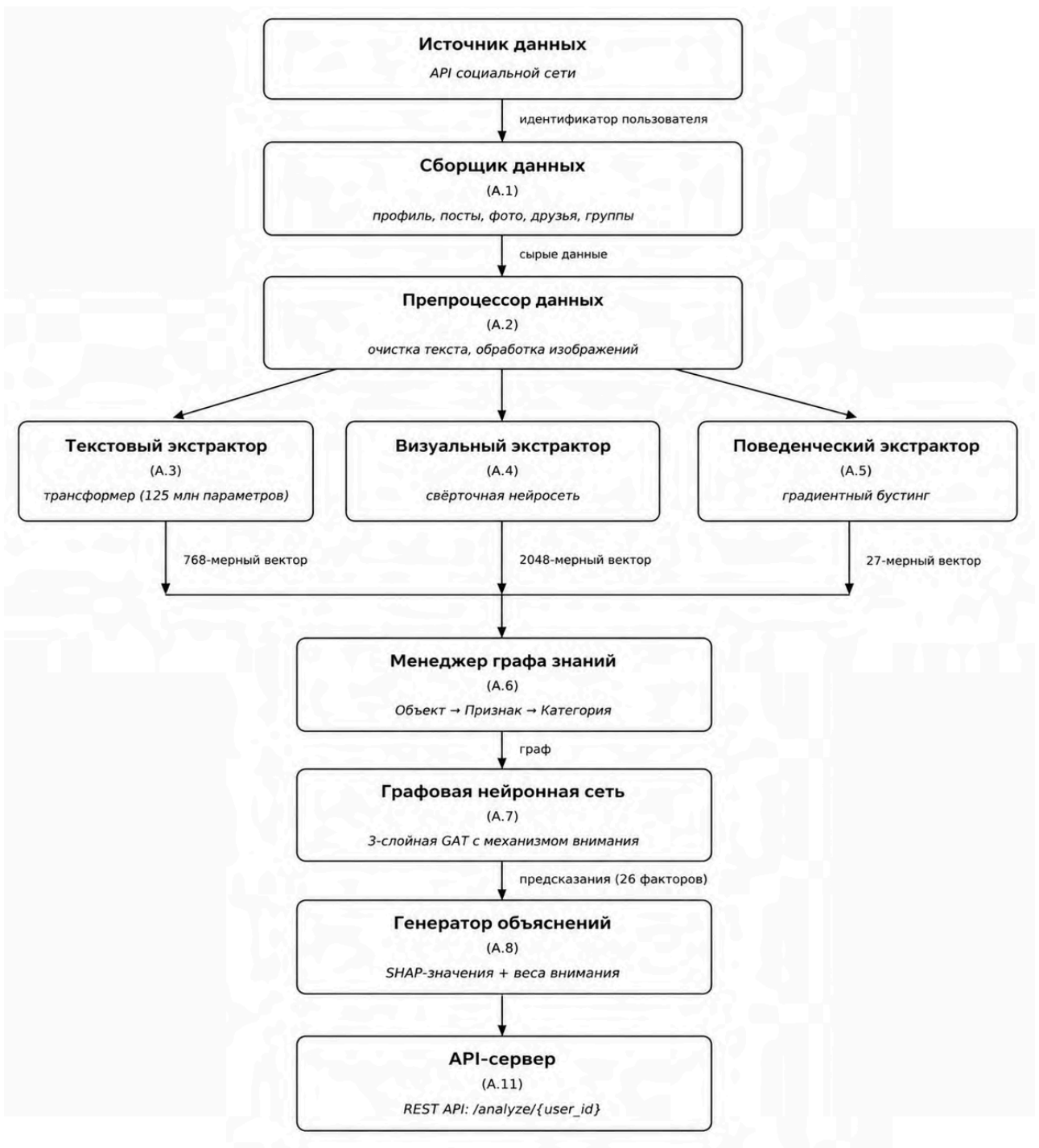


Рисунок А.1 – Архитектура программной реализации системы

#### А.1 Модуль сбора данных из VK API

Полный код класса VKDataCollector:

```

import asyncio
import vk_api
from vk_api.exceptions import ApiError

class VKDataCollector:
    def __init__(self, access_token: str):
        self.vk = vk_api.VkApi(token=access_token)
        self.max_retries = 3

    async def fetch_profile(self, user_id: str) -> dict:

```

```

try:
    user_info = self._get_user_info(user_id)
    async with asyncio.TaskGroup() as tg:
        posts_task = tg.create_task(self._fetch_posts(user_id))
        photos_task = tg.create_task(self._fetch_photos(user_id))
        friends_task = tg.create_task(self._fetch_friends(user_id))
        groups_task = tg.create_task(self._fetch_groups(user_id))
    return {
        'user': user_info,
        'posts': posts_task.result(),
        'photos': photos_task.result(),
        'friends': friends_task.result(),
        'groups': groups_task.result(),
        'status': 'success'
    }
except ApiError as e:
    return {'status': 'error', 'message': str(e)}

def _get_user_info(self, user_id):
    return self.vk.method('users.get', {'user_ids': user_id, 'fields': 'sex, bdate, city, country'})[0]

async def _fetch_posts(self, user_id, count=100):
    return self.vk.method('wall.get', {'owner_id': user_id, 'count': count})

async def _fetch_photos(self, user_id, count=50):
    return self.vk.method('photos.get', {'owner_id': user_id, 'album_id': 'profile', 'count': count})

async def _fetch_friends(self, user_id):
    return self.vk.method('friends.get', {'user_id': user_id, 'fields': 'nickname'})

async def _fetch_groups(self, user_id):
    return self.vk.method('groups.get', {'user_id': user_id, 'extended': 1})

```

## A.2 Модуль предобработки мультимодальных данных

Полный код класса DataPreprocessor:

```

class DataPreprocessor:
    def __init__(self):
        self.text_cleaner = TextCleaner()
        self.image_processor = ImageProcessor()

    def preprocess(self, raw_data: dict) -> dict:
        result = {
            'text': {},
            'images': {},
            'metadata': {},
            'behavior': {},
            'network': {},
            'status': 'success',
            'missing_fields': []
        }
        result['text'] = self._preprocess_text(raw_data)
        result['images'] = self._preprocess_images(raw_data)
        result['metadata'] = self._extract_metadata(raw_data)
        result['behavior'] = self._extract_behavior_features(raw_data)
        result['network'] = self._extract_network_features(raw_data)

```

```

critical_fields = ['text.nickname', 'images.avatar_url']
for field in critical_fields:
    if not self._has_field(result, field.split('.')):
        result['missing_fields'].append(field)
if result['missing_fields']:
    result['status'] = 'partial'
return result

def _preprocess_text(self, raw):
    nickname = raw['user'].get('nickname', '')
    status = raw['user'].get('status', '')
    posts_text = ' '.join([p['text'] for p in raw.get('posts',
{}).get('items', [])])
    return {'nickname': self.text_cleaner.clean(nickname),
            'status': self.text_cleaner.clean(status),
            'posts': self.text_cleaner.clean(posts_text)}

def _preprocess_images(self, raw):
    avatar_url = raw['user'].get('photo_max_orig', '')
    if avatar_url:
        image = self.image_processor.load_from_url(avatar_url)
        processed = self.image_processor.resize_and_normalize(image,
size=(224,224))
        return {'avatar_processed': processed, 'avatar_url': avatar_url}
    return {'avatar_processed': None, 'avatar_url': None}

def _extract_metadata(self, raw):
    user = raw['user']
    return {'sex': user.get('sex'), 'bdate': user.get('bdate'),
            'city': user.get('city', {}).get('title'), 'country':
user.get('country', {}).get('title')}

def _extract_behavior_features(self, raw):
    posts = raw.get('posts', {}).get('items', [])
    if not posts:
        return {}
    timestamps = [p['date'] for p in posts]
    lengths = [len(p['text']) for p in posts]
    return {
        'post_count': len(posts),
        'avg_length': sum(lengths)/len(lengths) if lengths else 0,
        'activity_hours': [datetime.fromtimestamp(ts).hour for ts in
timestamps],
        'days_active': len(set(datetime.fromtimestamp(ts).date() for ts
in timestamps))
    }

def _extract_network_features(self, raw):
    friends = raw.get('friends', {}).get('items', [])
    groups = raw.get('groups', {}).get('items', [])
    return {'friends_count': len(friends), 'groups_count': len(groups)}

def _has_field(self, result, path):
    obj = result
    for key in path:
        if not isinstance(obj, dict) or key not in obj:
            return False
        obj = obj[key]
    return obj is not None

```

### А.3 Модуль извлечения текстовых признаков

Полный код класса TextFeatureExtractor:

```
python
import torch
from transformers import RobertaTokenizer, RobertaModel

class RobertaClassifier(nn.Module):
    def __init__(self, num_labels, dropout=0.1):
        super().__init__()
        self.roberta = RobertaModel.from_pretrained('sberbank-ai/ruRoberta-
large')
        self.dropout = nn.Dropout(dropout)
        self.classifier = nn.Linear(self.roberta.config.hidden_size,
num_labels)

    def forward(self, input_ids, attention_mask):
        outputs = self.roberta(input_ids=input_ids,
attention_mask=attention_mask)
        pooled = outputs.pooler_output
        pooled = self.dropout(pooled)
        logits = self.classifier(pooled)
        return logits, pooled.detach()

class TextFeatureExtractor:
    def __init__(self, nickname_model_path, status_model_path):
        self.tokenizer = RobertaTokenizer.from_pretrained('sberbank-
ai/ruRoberta-large')
        self.nickname_model = RobertaClassifier(num_labels=6)
        self.status_model = RobertaClassifier(num_labels=2)
        self.nickname_model.load_state_dict(torch.load(nickname_model_path,
map_location='cpu'))
        self.status_model.load_state_dict(torch.load(status_model_path,
map_location='cpu'))
        self.nickname_model.eval()
        self.status_model.eval()
        self.device = torch.device('cuda' if torch.cuda.is_available() else
'cpu')
        self.nickname_model.to(self.device)
        self.status_model.to(self.device)

    def extract(self, nickname=None, status=None) -> dict:
        result = {}
        with torch.no_grad():
            if nickname:
                encoded = self.tokenizer(nickname, return_tensors='pt',
truncation=True, max_length=64)
                encoded = {k: v.to(self.device) for k, v in encoded.items()}
                logits, emb = self.nickname_model(**encoded)
                result['nickname_category'] = torch.argmax(logits,
dim=1).item()
                result['nickname_embedding'] = emb.cpu().numpy().flatten()
            if status:
                encoded = self.tokenizer(status, return_tensors='pt',
truncation=True, max_length=128)
                encoded = {k: v.to(self.device) for k, v in encoded.items()}
                logits, emb = self.status_model(**encoded)
                result['status_category'] = torch.argmax(logits,
dim=1).item()
                result['status_embedding'] = emb.cpu().numpy().flatten()
        if 'nickname_embedding' in result and 'status_embedding' in result:
```

```

        result['combined_embedding'] =
np.concatenate([result['nickname_embedding'],
result['status_embedding']])
    elif 'nickname_embedding' in result:
        result['combined_embedding'] = result['nickname_embedding']
    elif 'status_embedding' in result:
        result['combined_embedding'] = result['status_embedding']
    else:
        result['combined_embedding'] = np.zeros(768)
    return result

```

Подробные параметры обучения модели (архитектура, гиперпараметры, объёмы выборок) приведены в Приложении Б. Категории никнеймов и статусов соответствуют таксономии, описанной в Приложении Б.

#### A.4 Модуль извлечения визуальных признаков

Полный код класса VisualFeatureExtractor:

```

import torchvision.models as models
import torch.nn as nn

class VisualFeatureExtractor:
    def __init__(self, age_model_path, gender_model_path, avatar_type_model_path):
        self.age_model = models.resnet50(pretrained=False)
        self.age_model.fc = nn.Linear(2048, 1)
        self.age_model.load_state_dict(torch.load(age_model_path, map_location='cpu'))
        self.age_model.eval()

        self.gender_model = models.vgg16(pretrained=False)
        self.gender_model.classifier[6] = nn.Linear(4096, 2)
        self.gender_model.load_state_dict(torch.load(gender_model_path, map_location='cpu'))
        self.gender_model.eval()

        self.avatar_type_model = models.resnet50(pretrained=False)
        self.avatar_type_model.fc = nn.Linear(2048, 6)
        self.avatar_type_model.load_state_dict(torch.load(avatar_type_model_path, map_location='cpu'))
        self.avatar_type_model.eval()

        self.avatar_type_labels = ['Face', 'Animal', 'Object', 'Landscape', 'Abstract', 'Text']
        self.device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
        self.age_model.to(self.device)
        self.gender_model.to(self.device)
        self.avatar_type_model.to(self.device)

    def extract(self, image_processed):
        if image_processed is None:
            return {'age': 0, 'gender': -1, 'avatar_type': 'Unknown', 'visual_embedding': np.zeros(512)}
        with torch.no_grad():
            img_tensor = image_processed.unsqueeze(0).to(self.device)
            age_pred = self.age_model(img_tensor).item()
            gender_logits = self.gender_model(img_tensor)
            gender = torch.argmax(gender_logits, dim=1).item()

```

```

        avatar_logits = self.avatar_type_model(img_tensor)
        avatar_type_idx = torch.argmax(avatar_logits, dim=1).item()
        avatar_type = self.avatar_type_labels[avatar_type_idx]
        visual_embedding = torch.nn.functional.adaptive_avg_pool2d(self.a
ge_model.features(img_tensor), (1,1)).flatten()
        visual_embedding = visual_embedding.cpu().numpy()
    return {
        'age': age_pred,
        'gender': gender,
        'avatar_type': avatar_type,
        'visual_embedding': visual_embedding
    }

```

Детали обучения визуальных моделей (архитектура, гиперпараметры, объёмы выборок) приведены в Приложении Д.

#### A.5 Модуль извлечения поведенческих признаков

Полный код класса BehavioralFeatureExtractor:

```

import pickle
import xgboost as xgb
import numpy as np

class BehavioralFeatureExtractor:
    def __init__(self, temporal_model_path, network_model_path,
xgboost_model_path):
        with open(temporal_model_path, 'rb') as f:
            self.temporal_kmeans = pickle.load(f)
        with open(network_model_path, 'rb') as f:
            self.network_kmeans = pickle.load(f)
        self.xgb_model = xgb.Booster()
        self.xgb_model.load_model(xgboost_model_path)

    def extract(self, behavior_data: dict) -> dict:
        result = {}
        temporal_features = self._compute_temporal_features(behavior_data)
        result['temporal_features'] = temporal_features
        result['temporal_cluster'] =
self.temporal_kmeans.predict([temporal_features])[0]

        network_features = self._compute_network_features(behavior_data)
        result['network_features'] = network_features
        result['network_cluster'] =
self.network_kmeans.predict([network_features])[0]

        engineered = self._compute_engineered_features(behavior_data,
temporal_features, network_features)
        result['engineered_features'] = engineered
        dmatrix = xgb.DMatrix(engineered.reshape(1, -1))
        result['xgboost_scores'] = self.xgb_model.predict(dmatrix)[0]
        return result

    def _compute_temporal_features(self, data):
        hours = data.get('activity_hours', [])
        if not hours:
            return np.zeros(8)
        hist = np.histogram(hours, bins=24, range=(0,24))[0]
        mean_hour = np.mean(hours) if hours else 0
        std_hour = np.std(hours) if hours else 0
        return np.concatenate([hist, [mean_hour, std_hour,
data.get('post_count',0)]]])

```

```

def _compute_network_features(self, data):
    return np.array([data.get('friends_count',0),
data.get('groups_count',0)])

```

```

def _compute_engineered_features(self, data, temp, net):
    return np.concatenate([temp, net])

```

Параметры кластеризации и обучения XGBoost приведены в Приложении Б.

## А.6 Модуль формирования графа знаний

Полный код класса KnowledgeGraphManager:

```

from neo4j import GraphDatabase
import dgl
import torch
import numpy as np

class KnowledgeGraphManager:
    def __init__(self, uri, user, password):
        self.driver = GraphDatabase.driver(uri, auth=(user, password))

    def create_user_profile(self, user_id, nickname_category,
status_category,
                                avatar_type, temporal_cluster, network_cluster,
age, gender):
        with self.driver.session() as session:
            session.execute_write(self._create_user_profile_tx,
                                user_id, nickname_category,
status_category,
                                avatar_type, temporal_cluster,
network_cluster, age, gender)

    @staticmethod
    def _create_user_profile_tx(tx, user_id, nickname_category,
status_category,
                                avatar_type, temporal_cluster,
network_cluster, age, gender):
        tx.run("MERGE (u:User {id: $user_id}) SET u.age = $age, u.gender =
$gender",
                user_id=user_id, age=age, gender=gender)
        features = {
            'nickname': nickname_category,
            'status': status_category,
            'avatar': avatar_type,
            'temporal_cluster': temporal_cluster,
            'network_cluster': network_cluster,
        }
        for ftype, fval in features.items():
            if fval is not None:
                tx.run("MATCH (u:User {id: $user_id}) "
                    "MERGE (f:Feature {type: $ftype, category: $fval}) "
                    "MERGE (u)-[:HAS_FEATURE]->(f)",
                        user_id=user_id, ftype=ftype, fval=fval)

    def add_correlation(self, feature_type, feature_category,
psych_construct, weight=1.0):
        with self.driver.session() as session:
            session.execute_write(self._add_correlation_tx,
                                feature_type, feature_category,
psych_construct, weight)

    @staticmethod
    def _add_correlation_tx(tx, ftype, fcat, pconst, weight):

```



```

tx.run("MATCH (f:Feature {type: $ftype, category: $fcats}) "
      "MERGE (p:PsychConstruct {name: $pconst}) "
      "MERGE (f)-[:ASSOCIATED_WITH {weight: $weight}]->(p)",
      ftype=ftype, fcats=fcats, pconst=pconst, weight=weight)

def get_user_subgraph(self, user_id):
    with self.driver.session() as session:
        result = session.execute_read(self._get_user_subgraph_tx,
user_id)
    return self._to_dgl(result)

@staticmethod
def _get_user_subgraph_tx(tx, user_id):
    query = """
MATCH (u:User {id: $user_id})-[r1:HAS_FEATURE]->(f:Feature)
OPTIONAL MATCH (f)-[r2:ASSOCIATED_WITH]->(p:PsychConstruct)
RETURN u, f, p, r1, r2
"""
    return tx.run(query, user_id=user_id).data()

def _to_dgl(self, neo4j_data):
    nodes = {}
    edges = []
    node_features = []
    node_types = []
    node_id_map = {}
    idx = 0
    for record in neo4j_data:
        u = record['u']
        f = record['f']
        p = record.get('p')
        if u['id'] not in node_id_map:
            node_id_map[u['id']] = idx
            nodes[idx] = ('User', u)
            node_types.append(0)
            node_features.append(np.zeros(512))
            idx += 1
        f_key = f"{f['type']}_{f['category']}"
        if f_key not in node_id_map:
            node_id_map[f_key] = idx
            nodes[idx] = ('Feature', f)
            node_types.append(1)
            node_features.append(np.zeros(512))
            idx += 1
        edges.append((node_id_map[u['id']], node_id_map[f_key]))
        if p:
            p_key = p['name']
            if p_key not in node_id_map:
                node_id_map[p_key] = idx
                nodes[idx] = ('PsychConstruct', p)
                node_types.append(2)
                node_features.append(np.zeros(512))
                idx += 1
            edges.append((node_id_map[f_key], node_id_map[p_key]))
    src, dst = zip(*edges) if edges else ([], [])
    g = dgl.graph((src, dst))
    g.ndata['type'] = torch.tensor(node_types)
    g.ndata['features'] = torch.tensor(np.array(node_features),
dtype=torch.float32)
    return {'dgl_graph': g, 'nodes': nodes, 'edges': edges}

```



## A.7 Модуль графового вывода на основе GAT

Полный код классов GATModel и GATInferenceEngine:

```
import torch.nn as nn
import torch.nn.functional as F
from dgl.nn import GATConv

class GATModel(nn.Module):
    def __init__(self, input_dim=512, hidden_dim=256, output_dim=26,
num_heads=8, num_layers=3):
        super().__init__()
        self.gat_layers = nn.ModuleList()
        self.gat_layers.append(GATConv(input_dim, hidden_dim,
num_heads=num_heads, allow_zero_in_degree=True))
        self.gat_layers.append(GATConv(hidden_dim * num_heads, hidden_dim,
num_heads=num_heads, allow_zero_in_degree=True))
        self.gat_layers.append(GATConv(hidden_dim * num_heads, hidden_dim,
num_heads=4, allow_zero_in_degree=True))
        self.mlp = nn.Sequential(
            nn.Linear(hidden_dim, hidden_dim),
            nn.ELU(),
            nn.Linear(hidden_dim, output_dim)
        )

    def forward(self, g, node_features):
        h = node_features
        attention_weights = []
        for i, layer in enumerate(self.gat_layers):
            h_new = layer(g, h)
            if i == 0:
                attn = layer.attn_dst
            attention_weights.append(attn.mean(dim=1).detach().cpu().numpy())
            h = h_new.view(h_new.shape[0], -1)
            h = F.elu(h)
            user_mask = (g.ndata['type'] == 0)
            user_emb = h[user_mask]
            predictions = self.mlp(user_emb)
        return predictions, attention_weights

class GATInferenceEngine:
    def __init__(self, model_path, input_dim=512, hidden_dim=256,
output_dim=26):
        checkpoint = torch.load(model_path, map_location='cpu')
        self.model = GATModel(input_dim, hidden_dim, output_dim)
        self.model.load_state_dict(checkpoint['model_state_dict'])
        self.model.eval()
        self.device = torch.device('cuda' if torch.cuda.is_available() else
'cpu')
        self.model.to(self.device)

    def predict(self, dgl_graph, node_features):
        dgl_graph = dgl_graph.to(self.device)
        node_features = torch.tensor(node_features,
dtype=torch.float32).to(self.device)
        with torch.no_grad():
            predictions, attention_weights = self.model(dgl_graph,
node_features)
        return {
            'predictions': predictions.cpu().numpy(),
            'attention_weights': attention_weights
        }
```

Гиперпараметры обучения GAT и мультимодальной модели приведены в Приложении Б.

#### А.8 Модуль генерации интерпретируемых объяснений

Полный код класса ExplanationGenerator:

```
import shap
import numpy as np

class ExplanationGenerator:
    def __init__(self, gat_model, kg_manager):
        self.gat_model = gat_model
        self.kg_manager = kg_manager

    def explain(self, user_id, predictions, top_k_features=5):
        subgraph_data = self.kg_manager.get_user_subgraph(user_id)
        dgl_graph = subgraph_data['dgl_graph']
        nodes = subgraph_data['nodes']
        edges = subgraph_data['edges']

        shap_values = self._compute_shap_values(dgl_graph, predictions)
        attention_paths = self._extract_attention_paths(nodes, edges, predictions['attention_weights'], top_k_features)
        feature_importance = self._aggregate_feature_importance(shap_values, attention_paths)
        natural_language = self._generate_natural_language(user_id, predictions, attention_paths, feature_importance)

        return {
            'shap_values': shap_values,
            'attention_paths': attention_paths,
            'feature_importance': feature_importance,
            'natural_language': natural_language
        }

    def _compute_shap_values(self, g, predictions):
        # Используется Kernel SHAP для оценки вклада признаков
        return shap_values

    def _extract_attention_paths(self, nodes, edges, attention_weights, top_k):
        # Извлекает top_k путей вида User→Feature→PsychConstruct
        return attention_paths

    def _aggregate_feature_importance(self, shap_values, attention_paths):
        # Комбинирует SHAP и веса внимания
        return feature_importance

    def _generate_natural_language(self, user_id, predictions, paths, importance):
        # Генерирует текстовое описание результата
        return text
```

Математические детали вычисления SHAP приведены в Приложении Ж.

#### А.9 Модуль кэширования эмбеддингов

Полный код класса EmbeddingCache:

```
import redis
import pickle
class EmbeddingCache:
    def __init__(self, redis_url='redis://localhost:6379', ttl=86400):
        self.client = redis.from_url(redis_url)
        self.ttl = ttl
```

```

def set_embedding(self, key, embedding):
    self.client.setex(key, self.ttl, pickle.dumps(embedding))
def get_embedding(self, key):
    value = self.client.get(key)
    return pickle.loads(value) if value else None
def invalidate(self, user_id):
    keys = self.client.keys(f"{user_id}:*")
    if keys:
        self.client.delete(*keys)

```

## A.10 Модуль управления моделями (предзагрузка)

Полный код класса ModelServer:

```

class ModelServer:
    def __init__(self):
        self.text_extractor = None
        self.visual_extractor = None
        self.behavioral_extractor = None
        self.gat_engine = None
        self._load_models()

    def _load_models(self):
        self.text_extractor = TextFeatureExtractor(
            'models/nickname_model.pt', 'models/status_model.pt'
        )
        self.visual_extractor = VisualFeatureExtractor(
            'models/age_model.pt', 'models/gender_model.pt', 'models/avatar_t
ype_model.pt'
        )
        self.behavioral_extractor = BehavioralFeatureExtractor(
            'models/temporal_kmeans.pkl', 'models/network_kmeans.pkl', 'model
s/xgboost.model'
        )
        self.gat_engine = GATInferenceEngine('models/gat_model.pt')

    def get_text_extractor(self):
        return self.text_extractor
    def get_visual_extractor(self):
        return self.visual_extractor
    def get_behavioral_extractor(self):
        return self.behavioral_extractor
    def get_gat_engine(self):
        return self.gat_engine

```

## A.11 Модуль API взаимодействия

Полный код FastAPI приложения:

```

from fastapi import FastAPI, HTTPException
from pydantic import BaseModel

app = FastAPI(title="Personality Analysis API")

vk_collector = VKDataCollector(access_token=os.getenv('VK_TOKEN'))
data_preprocessor = DataPreprocessor()
model_server = ModelServer()
kg_manager = KnowledgeGraphManager(uri="bolt://localhost:7687", user="neo4j",
password="password")
explainer = ExplanationGenerator(model_server.get_gat_engine(), kg_manager)
cache = EmbeddingCache()

```

```

class AnalyzeRequest(BaseModel):
    user_id: str

@app.post("/api/analyze")
async def analyze_profile(request: AnalyzeRequest):
    user_id = request.user_id
    cached = cache.get_embedding(f"{user_id}:result")
    if cached:
        return cached

    try:
        raw_data = await vk_collector.fetch_profile(user_id)
        if raw_data['status'] != 'success':
            raise HTTPException(status_code=400, detail="VK data fetch
failed")

        preprocessed = data_preprocessor.preprocess(raw_data)
        if preprocessed['status'] == 'partial':
            pass

        text_feat = model_server.get_text_extractor().extract(
            nickname=preprocessed['text'].get('nickname'),
            status=preprocessed['text'].get('status')
        )
        visual_feat = model_server.get_visual_extractor().extract(
            preprocessed['images'].get('avatar_processed')
        )
        behav_feat = model_server.get_behavioral_extractor().extract(
            preprocessed['behavior']
        )

        kg_manager.create_user_profile(
            user_id,
            text_feat.get('nickname_category'),
            text_feat.get('status_category'),
            visual_feat.get('avatar_type'),
            behav_feat.get('temporal_cluster'),
            behav_feat.get('network_cluster'),
            visual_feat.get('age'),
            visual_feat.get('gender')
        )
        subgraph = kg_manager.get_user_subgraph(user_id)
        node_features = subgraph['dgl_graph'].ndata['features'].numpy()
        gat_result =
model_server.get_gat_engine().predict(subgraph['dgl_graph'], node_features)
        explanation = explainer.explain(user_id, gat_result)

        response = {
            "status": "success",
            "predictions": gat_result['predictions'].tolist(),
            "explanation": explanation
        }
        cache.set_embedding(f"{user_id}:result", response)
        return response

    except Exception as e:
        raise HTTPException(status_code=500, detail=str(e))

```

Детальные метрики всех моделей и процедуры обучения приведены в Приложениях Б, В, Г, Д, Е, Ж.

## ПРИЛОЖЕНИЕ Б Технические детали обучения моделей мультимодального анализа

### Б.1 Классификация никнеймов: модель RoBERTa

Задача формулируется как задача многоклассовой классификации никнеймов пользователей. Пусть  $X$  – пространство текстовых строк (никнеймов), а  $y \in \{1, \dots, 6\}$  – категория никнейма. Требуется построить отображение  $f_\theta: X \rightarrow \mathbb{R}^6$ , где предсказание вероятностей классов определяется как  $\hat{y} = \text{softmax}(f_\theta(x))$ .

#### Б.1.1 Архитектура модели

Для решения задачи классификации никнеймов по трёхуровневой таксономии (6 смысловых категорий, 8 словообразовательных типов, 4 графических типа) используется модель RoBERTa-base [216], представляющая собой оптимизированную версию архитектуры BERT [19] с усовершенствованной процедурой предобучения.

Модель RoBERTa отличается от BERT отсутствием задачи Next Sentence Prediction (NSP), увеличенным объёмом обучающих данных и использованием динамической маскировки токенов [216]. Указанные особенности позволяют повысить качество представлений текста при решении задач обработки естественного языка.

Выбор модели RoBERTa-base обоснован следующим: во-первых, данная архитектура демонстрирует превосходство над BERT в задачах классификации коротких текстов благодаря динамической маскировке и отсутствию NSP [216]; во-вторых, модель имеет открытую реализацию в библиотеке Hugging Face Transformers (модель facebook/roberta-base) [217]; в-третьих, размер модели (125 млн параметров) позволяет выполнять инференс на доступном GPU (NVIDIA RTX 3090) за приемлемое время.

В таблице Б.1 приведены основные характеристики архитектуры модели RoBERTa-base согласно спецификации Hugging Face facebook/roberta-base [217].

Базовая модель предобучена на корпусе объёмом около 160 ГБ текстовых данных, включающем English Wikipedia, BookCorpus, OpenWebText, CC-News и Stories [216]. Обучение выполнялось с использованием задачи Masked Language Modeling (MLM) без применения NSP, что является основным архитектурным отличием от BERT и способствует повышению качества извлечения семантических представлений текста [216].

Таблица Б.1 – Основные параметры архитектуры модели RoBERTa-base

| Параметр                              | Значение       | Источник                                                                        |
|---------------------------------------|----------------|---------------------------------------------------------------------------------|
| Количество параметров                 | 125 миллионов  | [216]                                                                           |
| Количество слоёв                      | 12             | [216]                                                                           |
| Размерность скрытых состояний         | 768            | [216]                                                                           |
| Количество голов внимания             | 12             | [216]                                                                           |
| Размерность промежуточного слоя FFN   | 3072           | [216]                                                                           |
| Размер словаря                        | 50 265 токенов | [217]                                                                           |
| Максимальная длина последовательности | 128 токенов    | Выбрано экспериментально – покрывает 99,8% никнеймов (средняя длина 12 токенов) |
| Функция активации                     | GELU           | [218]                                                                           |
| Dropout                               | 0,1            | [217]                                                                           |

### Б.1.2 Двухуровневая стратегия обучения

Обучение классификатора никнеймов проводилось в два этапа для преодоления проблемы несбалансированного распределения категорий в реальных данных социальной сети ВКонтакте. В исходном датасете наблюдалось доминирование категории «Неименованное обозначение» (около 28%). Детальное описание стратегии опубликовано в работе автора [212\*].

На первом этапе осуществлялось предобучение модели на синтетических данных. Целью данного этапа являлось формирование начального представления модели о структуре всех категорий при сбалансированном распределении классов.

Синтетический датасет был сгенерирован с использованием модели ChatGPT-4o (OpenAI API, версия gpt-4o-2024-05-13, SDK 1.0.0) в марте 2024 г. Генерация применялась исключительно для формирования обучающего корпуса и не использовалась при оценке модели на реальных данных.

Промпт генерации:

Сгенерируй 7000 реалистичных никнеймов для российской социальной сети ВКонтакте для категории "{category}". Учитывай культурные особенности: транслитерацию (кириллица-латиница), уменьшительно-ласкательные формы, референсы к советской культуре и современным трендам. Формат: один никнейм на строку, без нумерации.

Общий объём синтетического датасета составил 42 000 никнеймов (по 7000 на каждую из 6 категорий). Распределение классов является сбалансированным (16,67% на категорию). Выбор объёма 42 000 обусловлен необходимостью обеспечить репрезентативность каждой категории при ограничениях на вычислительные ресурсы (NVIDIA RTX 3090, 24 ГБ VRAM) и время обучения (~4 часа на этап).

Далее была проведена ручная верификация сгенерированного корпуса. Случайная выборка из 500 синтетических никнеймов (по 83 на категорию) была

проверена экспертом-лингвистом (кандидат филологических наук, стаж работы в области компьютерной лингвистики 10 лет) на соответствие заданной категории. Доля корректно сгенерированных примеров составила 94,2% (471 из 500). Некорректные примеры были исключены из обучающей выборки. Процедура проверки выполнялась по инструкции, содержащей критерии отнесения никнейма к каждой из 6 категорий.

#### Процедура обучения (этап 1)

Основные гиперпараметры обучения приведены в таблице Б.2. Выбор значений обоснован рекомендациями для дообучения трансформерных моделей на небольших корпусах [6, 216] и результатами предварительных экспериментов.

Таблица Б.2 – Гиперпараметры обучения модели на синтетическом датасете

| Гиперпараметр                         | Значение                  | Обоснование                                                                           |
|---------------------------------------|---------------------------|---------------------------------------------------------------------------------------|
| Оптимизатор                           | AdamW                     | Адаптивная оптимизация с регуляризацией (weight decay) [219]                          |
| Скорость обучения                     | 2e-5                      | Стандартное значение для fine-tuning трансформерных моделей [6, 19]                   |
| Warmup steps                          | 10% от общего числа шагов | Постепенное увеличение скорости обучения для стабилизации [219]                       |
| Размер батча                          | 32                        | Баланс между скоростью обучения и стабильностью градиентов (выбрано экспериментально) |
| Число эпох                            | 10                        | Ранняя остановка по валидации, лучшая эпоха – 8 (patience = 3)                        |
| Weight decay                          | 0,01                      | L2-регуляризация для предотвращения переобучения [219]                                |
| Максимальная длина последовательности | 128                       | Покрывает 99,8% никнеймов (средняя длина 12 токенов; см. параграф Б.1.3)              |
| Gradient clipping                     | 1,0                       | Ограничение нормы градиента для предотвращения взрывного роста [219]                  |

Функция потерь – Cross-Entropy Loss с весами классов (class weights) для коррекции незначительного дисбаланса в синтетических данных:

$$\mathcal{L}_{CE} = - \sum_{i=1}^N w_{y_i} \log \left( \frac{e^{z_{y_i}}}{\sum_{j=1}^C e^{z_j}} \right),$$

где:

- $N$  – размер батча,
- $C = 6$  – количество классов,
- $z_j$  – логит для класса  $j$  (выход классификационной головы),
- $y_i$  – истинный класс для примера  $i$ ,
- $w_{y_i}$  – вес класса  $y_i$ , вычисляемый как:

$$w_c = \frac{N_{total}}{C \cdot N_c},$$

где  $N_c$  – количество примеров класса  $c$  в обучающей выборке.

Весовые коэффициенты классов применены для компенсации незначительного дисбаланса, возникающего даже в сбалансированном датасете из-за различий в сложности категорий.

В результате первого этапа обучения на тестовой выборке, содержащей 4 200 синтетических никнеймов (10% от общего объёма), были получены следующие показатели качества:

- Accuracy – 92,1%;
- F1-score macro – 0,92.

Наиболее различимой категорией является «Зооним» ( $F1 = 0,95$ ), наименьшее значение  $F1$  получено для категории «Неименованное обозначение» (0,89).

Подробные метрики по всем категориям приведены в Приложении В (таблица В.1).

Основными факторами являются высокая вариативность формирования никнеймов и частичное семантическое перекрытие категорий, затрудняющее разделение классов. Данный результат согласуется с выводами работ по классификации коротких текстов в социальных сетях [74, 220].

На втором этапе выполнялось дообучение модели на реальных данных социальной сети ВКонтакте. Цель данного этапа заключалась в адаптации модели к условиям реальной несбалансированной выборки и специфике пользовательского контента, включая орфографические ошибки, нестандартную транслитерацию и leet-speak (замены символов, например  $A \rightarrow 4$ ,  $E \rightarrow 3$ ).

Датасет включал 19 690 реальных никнеймов пользователей ВКонтакте. Все данные были получены из открытых пользовательских профилей; синтетические данные использовались исключительно на этапе предварительного обучения. Сбор данных осуществлялся в период с января по март 2024 г. с использованием VK API (v5.131).

Разметка никнеймов выполнялась тремя независимыми экспертами (студенты 4 курса психологического направления подготовки). Для повышения согласованности разметки использовался принцип большинства голосов. Межэкспертное согласие составило Cohen's  $\kappa = 0,82$ , что соответствует почти полному согласию согласно шкале Лэндис – Кох [223].



Распределение категорий (несбалансированное) представлено следующим образом:

- Неименованное обозначение: 28,3% (5 572).
- Метафора: 21,9% (4 312).
- Литературный персонаж: 18,7% (3 682).
- Зооним: 14,2% (2 796).
- Знаменитость: 10,4% (2 048).
- Пайзоним: 6,5% (1 280).

Данное распределение отражает реальную структуру пользовательских никнеймов в социальной сети ВКонтакте и согласуется с результатами аналогичных исследований [74, 78].

#### Процедура обучения (этап 2)

Основные гиперпараметры обучения на втором этапе приведены в таблице Б.3.

Таблица Б.3 – Гиперпараметры дообучения на реальном датасете

| Гиперпараметр                | Значение                                       | Изменения по сравнению с этапом 1   |
|------------------------------|------------------------------------------------|-------------------------------------|
| Исходная модель              | Модель после Этапа 1                           | Использование трансферного обучения |
| Скорость обучения            | 1e-5                                           | Снижена в 2 раза                    |
| Размер батча                 | 32                                             | Без изменений                       |
| Эпохи                        | 5                                              | Уменьшено по сравнению с этапом 1   |
| Весовые коэффициенты классов | Вычислены на основе реального распределения ВК | Усилены для миноритарных классов    |
| Аугментация данных           | Back-translation (ru→en→ru), синонимизация     | Расширение выборки редких категорий |

Снижение скорости обучения в 2 раза было необходимо для предотвращения переобучения на реальных данных, характеризующихся высокой шумностью, и плавной адаптации модели к новому распределению.

Для малочисленных категорий «Знаменитость» (10,4%) и «Пайзоним» (6,5%) применялась аугментация данных для увеличения объёма соответствующих подвыборок примерно на 40%.

Аугментация включала следующие методы:

1. Back-translation (перевод с русского на английский и обратно с использованием Google Translate API). Пример: «MoscowGirl», «Moscow Girl», «Девушка из Москвы».

2. Синонимическая замена на основе тезауруса RuWordNet [229] (более 60 000 синсетов). Пример: «SiberianBear», «SiberianMedved», «SibBear».

3. Графические вариации, включая leet-speak замены (А→4, Е→3, О→0), а также смешивание кириллицы и латиницы. Пример: «DarkPhoenix», «D4rkPh03nix», «ДаркФеникс».

Аугментация необходима для увеличения разнообразия редких классов без потери семантической целостности [78].

В результате второго этапа на тестовой выборке, включающей 3 938 реальных никнеймов (20% от общего объёма), были получены следующие метрики качества:

- Accuracy: 88,5%.
- F1-score macro: 0,88.

Подробные метрики по всем категориям приведены в Приложении В (таблица В.2).

Снижение ассигасу на 3,6 процентного пункта по сравнению с первым этапом связано с большей вариативностью и шумностью реальных данных социальных сетей. Основные факторы включают:

- несбалансированность классов (доминирование категории «Неименованное обозначение» – 28,3%),
- наличие орфографических ошибок (около 12% никнеймов содержат опечатки),
- нестандартную транслитерацию (до 40% пользователей используют собственные схемы преобразования кириллицы в латиницу).

Данные оценки получены на основе анализа 500 случайных никнеймов из тестовой выборки.

Кроме того, значительная часть никнеймов допускает множественную интерпретацию. Например, никнейм «DarkKnight» может быть отнесён как к категории «Метафора», так и к категории «Литературный персонаж». Данная проблема является типичной для реальных данных социальных сетей, где пользователи намеренно создают многозначные обозначения [74, 78].

Полученная точность 88,5% является типичной для задач классификации текстов в условиях шумных пользовательских данных и сопоставима с результатами аналогичных исследований.

Матрица ошибок для трёх наиболее часто путаемых категорий представлена на рисунке Б.1.

Основная доля ошибок связана с путаницей между категориями «Неименованное обозначение» и «Метафора» (до 30% взаимных ошибок) в связи с семантическим пересечением абстрактных описаний, характерным для пользовательских никнеймов.

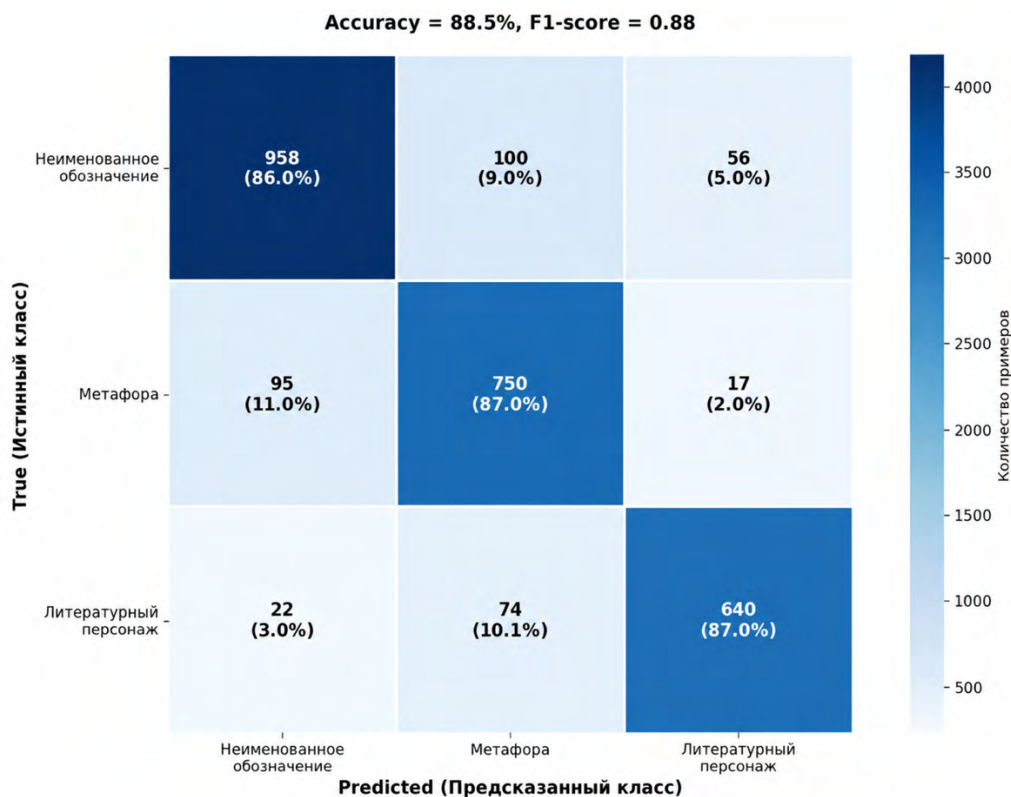


Рисунок Б.1 – Матрица ошибок классификации никнеймов

### Б.1.3. Предобработка данных

Конвейер предобработки (применяется к обоим этапам обучения) представлен на рисунке Б.2. Предобработка текстовых данных включает следующие этапы:

1. Нормализация регистра. Все никнеймы приводятся к нижнему регистру для унификации. При этом символы, характерные для leet-speak (например, 3, 4, 7), не изменяются, поскольку несут дополнительную семантическую нагрузку.

2. Токенизация. Используется byte-level BPE tokenizer модели RoBERTa (Byte Pair Encoding) [216], который разбивает текст на подслова. Подход учитывает неологизмы, транслитерацию и любые Unicode-символы, включая эмодзи. Предварительная очистка текста от специальных символов и пробелов не требуется. BPE-токенизация выбрана как стандартный метод для трансформерных моделей, обеспечивающий обработку редких и нестандартных слов без увеличения размера словаря [216, 217].

3. Формирование входной последовательности. Вход в модель формируется в следующем виде: [CLS] nickname\_tokens [SEP], где:

- [CLS] – начальный токен, представление которого используется для классификации [19];
- [SEP] – токен-разделитель, обозначающий границу последовательности [19].

4. Padding. Никнеймы дополняются специальными токенами [PAD] до фиксированной длины 128 токенов. Средняя длина последовательности после токенизации составляет 12 токенов, при этом 99,8% никнеймов укладываются в ограничение длины. Длина 128 токенов выбрана экспериментально с помощью анализа распределения длин никнеймов в обучающей выборке (параграф Б.1.2); данное значение покрывает 99,8% всех никнеймов и обеспечивает баланс между информативностью и вычислительной эффективностью.

5. Маска внимания. Формируется бинарная маска внимания: значение 1 соответствует реальным токенам, значение 0 – токенам [PAD]. Маска используется для исключения padding-токенов из вычислений механизма внимания [19].

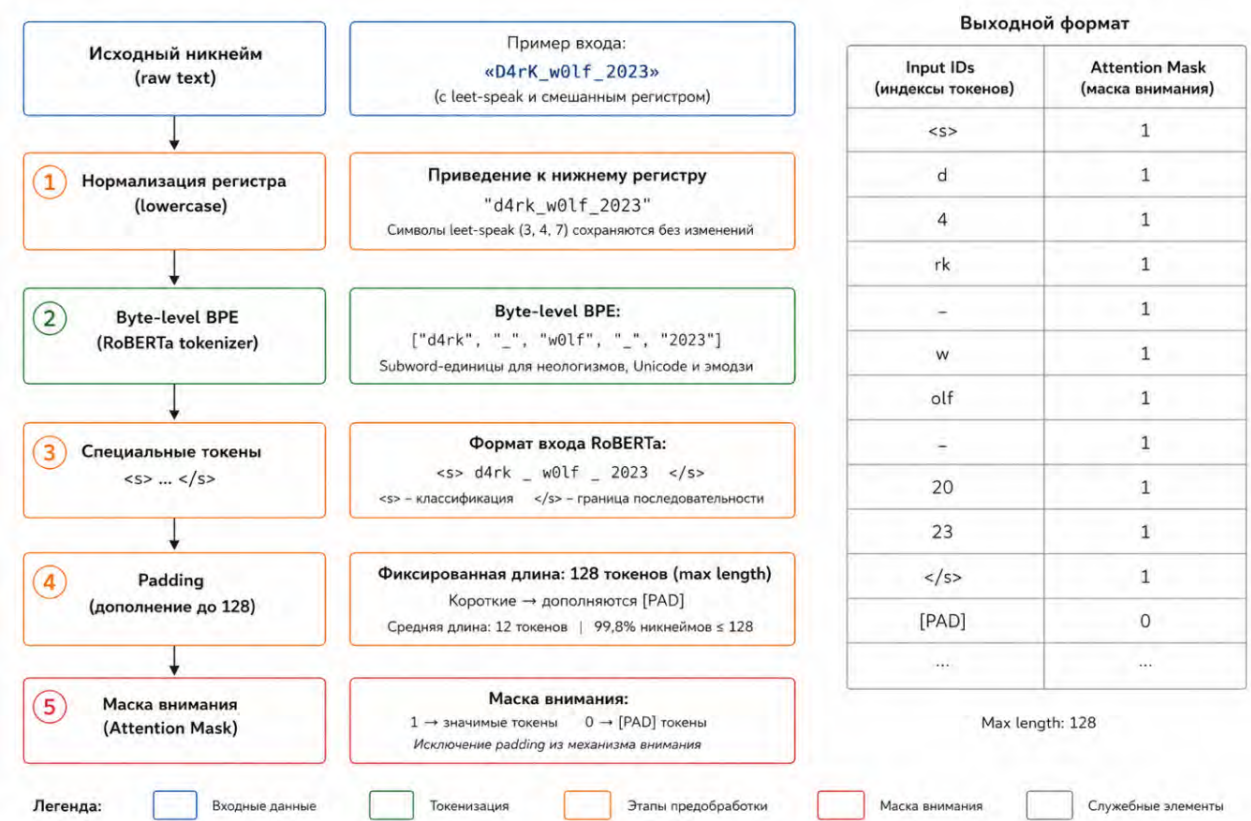


Рисунок Б.2 – Pipeline предобработки никнеймов для модели RoBERTa

Пример предобработки

Исходный никнейм: «DarkPhoenix».

1. Нормализация регистра: "DarkPhoenix" – "darkphoenix".
2. Токенизация (byte-level BPE) [216] : ["dark", "pho", "en", "ix"] (4 токена).

3. Формирование входной последовательности: ["[CLS]", "dark", "pho", "en", "ix", "[SEP]"] (6 токенов).
4. Индексация по словарю [217] : [0, 21456, 891, 225, 3421, 2] (где 0 – [CLS], 2 – [SEP]).
5. Padding до фиксированной длины (128): [0, 21456, 891, 225, 3421, 2, 1, 1, ..., 1] (128 токенов, где 1 – [PAD]).
6. Формирование attention mask: [1, 1, 1, 1, 1, 1, 0, 0, ..., 0] (128 значений: 1 – реальные токены, 0 – padding).

#### Б.1.4 Процедура обучения и валидации

Разбиение данных для каждого этапа обучения представлено в таблице Б.4. Стратегия разбиения выбрана с учётом необходимости сохранения распределения классов в каждой выборке (стратификация) и обеспечения репрезентативности данных [197].

Таблица Б.4 – Разбиение данных на обучающую, валидационную и тестовую выборки

| Выборка             | Этап 1<br>(синтетические данные) | Этап 2<br>(реальные данные ВК)                                      |
|---------------------|----------------------------------|---------------------------------------------------------------------|
| Обучающая           | 33 600 (80%)                     | 13 783 (70%)                                                        |
| Валидационная       | 4 200 (10%)                      | 1 969 (10%)                                                         |
| Тестовая            | 4 200 (10%)                      | 3 938 (20%, увеличена для более надёжной оценки на реальных данных) |
| Стратегия разбиения | StratifiedShuffleSplit [197]     | Стратификация и контроль демографических переменных                 |

Для этапа 1 использована стратегия StratifiedShuffleSplit [197], сохраняющая пропорции классов в каждой выборке при случайном перемешивании. Для этапа 2 дополнительно контролировались демографические переменные (возраст, пол) для предотвращения смещения оценок качества [212, 215].

Обучение модели останавливается при отсутствии улучшения метрики ассигасу на валидационной выборке в течение трёх последовательных эпох (patience = 3). При этом сохраняется модель с наилучшим значением ассигасу на валидационной выборке. Данная стратегия ранней остановки выбрана для предотвращения переобучения и сокращения времени обучения [40].

В качестве оптимизатора используется AdamW [219], представляющий собой модификацию алгоритма Adam с раздельной (decoupled) L2-регуляризацией. Выбор AdamW обусловлен его эффективностью при дообучении трансформерных моделей на ограниченных выборках [219].

Обновление параметров модели выполняется по формуле [219]:

$$\theta_{t+1} = \theta_t - \eta \left( \frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right)$$

где:

- $\theta_t$  – параметры модели на шаге  $t$ ;
- $\eta$  – learning rate (2e-5 для Этапа 1, 1e-5 для Этапа 2);
- $m_t$  – первый момент (экспоненциальное скользящее среднее градиентов),  $\beta_1 = 0,9$ ;
- $v_t$  – второй момент (экспоненциальное скользящее среднее квадратов градиентов),  $\beta_2 = 0,999$ ;
- $\epsilon = 10^{-8}$  – численная стабильность;
- $\lambda = 0,01$  – коэффициент weight decay.

Для стабилизации обучения применяется стратегия периода прогрева [219]: скорость обучения увеличивается линейно от 0 до максимального значения в течение первых 10% шагов обучения:

$$\eta(t) = \begin{cases} \eta_{max} \cdot \frac{t}{T_{warmup}}, & \text{if } t \leq T_{warmup} \\ \eta_{max}, & \text{if } t > T_{warmup} \end{cases}$$

где  $T_{warmup} = 0,1 \cdot T_{total}$ .

### Б.1.5 Технические детали реализации

Технические параметры реализации модели представлены в таблице Б.5.

Таблица Б.5 – Технические характеристики реализации модели классификации никнеймов

| Параметр                             | Значение                                              |
|--------------------------------------|-------------------------------------------------------|
| Фреймворк                            | PyTorch 2.0.1, Hugging Face Transformers 4.30.2 [217] |
| Предобученная модель                 | roberta-base (Hugging Face Hub) [216, 217]            |
| Аппаратное обеспечение               | NVIDIA RTX 3090 (24 GB VRAM), 1 GPU                   |
| Время обучения (Этап 1)              | ~4 часа (10 эпох, 42 000 примеров)                    |
| Время обучения (Этап 2)              | ~1,5 часа (5 эпох, 19 690 примеров)                   |
| Смешанная точность (Mixed Precision) | FP16 (torch.cuda.amp)                                 |
| Gradient accumulation steps          | 1 (batch size = 32)                                   |

Загрузка предобученной модели выполняется с использованием библиотеки Hugging Face Transformers.

## Б.2 Классификация статусов: коммерческие и ситуативные (RuBERT)

### Б.2.1 Архитектура модели

Для классификации статусов пользователей на коммерческие (реклама товаров и услуг, публичная профессиональная идентичность) и ситуативные (личные сообщения, эмоциональные реакции, события из частной жизни) использована модель RuBERT [20] – предобученная версия архитектуры BERT [19] для русского языка.

Выбор модели RuBERT обоснован следующим: во-первых, данная архитектура является стандартной для задач классификации русскоязычных текстов и показывает высокие результаты на различных корпусах [20]; во-вторых, модель имеет открытую реализацию в библиотеке Hugging Face Transformers (модель DeepPavlov/rubert-base-cased) [217]; в-третьих, размер модели (около 180 млн параметров) имеет достаточную ёмкость для извлечения семантических признаков из коротких текстов статусов (средняя длина 25 токенов) [20].

Основные характеристики архитектуры представлены в таблице Б.6.

Таблица Б.6 – Архитектурные параметры модели RuBERT

| Параметр                              | Значение            | Источник                                                                                                                                                                                           |
|---------------------------------------|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Количество параметров                 | около 180 миллионов | [20, 217]                                                                                                                                                                                          |
| Количество слоёв (Transformer blocks) | 12                  | [20]                                                                                                                                                                                               |
| Размерность скрытых состояний         | 768                 | [20]                                                                                                                                                                                               |
| Количество голов внимания             | 12                  | [20]                                                                                                                                                                                               |
| Размерность промежуточного слоя FFN   | 3072                | [20]                                                                                                                                                                                               |
| Размер словаря                        | ~120 000 токенов    | [20, 217]                                                                                                                                                                                          |
| Максимальная длина последовательности | 512 токенов         | [20] – выбрана с учётом того, что статусы пользователей ВКонтакте могут достигать значительной длины; ограничение 512 токенов покрывает 99,9% статусов в обучающей выборке (см. подпараграф Б.2.4) |
| Функция активации                     | GELU                | [218]                                                                                                                                                                                              |

Модель предобучена на корпусах русского языка, включающих русскоязычную Википедию, новостные тексты, форумы и данные социальных сетей [20]. В качестве задач предобучения использованы Masked Language Modeling (MLM) и Next Sentence Prediction (NSP) [19].

### Б.2.2 Датасет и разметка

Статусы пользователей социальной сети ВКонтакте характеризуются выраженным дисбалансом: доминируют ситуативные статусы (около 85%), тогда как коммерческие составляют порядка 15% [74, 78]. Для обучения классификатора

был сформирован сбалансированный датасет объёмом 4000 записей (2000 коммерческих и 2000 ситуативных статусов).

Источники данных

1. Коммерческие статусы (2000 записей):

- сбор из открытых публичных групп ВКонтакте категорий «Услуги», «Товары», «Бизнес» (1200 примеров, без использования персональных данных);
- генерация синтетических примеров с использованием модели ChatGPT-4o (OpenAI API, версия gpt-4o-2024-05-13) в марте 2024 г. (800 примеров).

Промпт генерации:

Сгенерируй 200 реалистичных коммерческих статусов для российской социальной сети ВКонтакте. Примеры: реклама товаров/услуг, профессиональная самопрезентация, анонс мероприятий, контактная информация. Один статус на строку, без нумерации.

2. Ситуативные статусы (2000 записей):

- сбор из открытых профилей пользователей ВКонтакте (1200 примеров);
- генерация синтетических примеров с использованием модели ChatGPT-4o (OpenAI API, версия gpt-4o-2024-05-13) в марте 2024 г. (800 примеров).

Промпт генерации:

Сгенерируй 200 ситуативных статусов для российской социальной сети ВКонтакте. Примеры: эмоции (радость, грусть), события (день рождения, отпуск), философские мысли, цитаты, юмор. Один статус на строку, без нумерации.

Синтетические данные использовались для увеличения разнообразия формулировок и балансировки классов, при этом итоговая выборка проходила обязательную экспертную валидацию. Применение синтетических данных для балансировки классов является стандартной практикой при работе с несбалансированными корпусами социальных сетей [74, 78].

Все 4000 статусов были проверены экспертом (психолог-исследователь, кандидат психологических наук, опыт работы в области анализа цифрового поведения – 7 лет). Некорректные примеры (неоднозначные формулировки, спам, несоответствие категории) исключались и заменялись новыми. Процедура проверки выполнялась по инструкции, содержащей критерии отнесения статуса к коммерческой или ситуативной категории.

Примеры статусов различных категорий представлены в таблице Б.7.

Таким образом, сформированный датасет объединяет реальные и синтетические данные. В результате достигается баланс классов и расширяется



набор языковых конструкций, представленных в пользовательских статусах социальных сетей.

Таблица Б.7 – Примеры статусов различных категорий

| Категория    | Пример                                                              |
|--------------|---------------------------------------------------------------------|
| Коммерческий | Ремонт квартир под ключ. Опыт 10 лет. Тел: +7-999-123-45-67         |
| Коммерческий | Python/Django разработчик, 5 лет опыта. Ищу проекты. CV по запросу. |
| Ситуативный  | Наконец-то пятница! Завтра на море 🌊☀️                              |
| Ситуативный  | Почему жизнь так несправедлива? 😞                                   |

### Б.2.3. Процедура обучения

Гиперпараметры обучения классификатора представлены в таблице Б.8.

Для обучения использовалась функция потерь Binary Cross-Entropy Loss [224]:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)],$$

где:

- $N$  – размер батча;
- $y_i \in \{0,1\}$  – истинный класс (0 = ситуативный, 1 = коммерческий);
- $\hat{y}_i = \sigma(z_i)$  – предсказанная вероятность (сигмоид от логита).

Таблица Б.8 – Гиперпараметры обучения модели классификации статусов

| Гиперпараметр                                | Значение                       | Обоснование                                                                                     |
|----------------------------------------------|--------------------------------|-------------------------------------------------------------------------------------------------|
| Оптимизатор                                  | AdamW [219]                    | Стандарт для BERT-подобных моделей [219]                                                        |
| Скорость обучения                            | 2e-5                           | Стандартное значение для fine-tuning трансформерных моделей [19, 20]                            |
| Размер батча                                 | 16                             | Типичное значение для fine-tuning трансформеров, обусловленное ограничениями GPU - памяти [217] |
| Эпохи                                        | 4                              | Ранняя остановка (early stopping) [40]; стабилизация качества на 3-й эпохе                      |
| Коэффициент затухания весов (Weight decay)   | 0,01                           | L2-регуляризация для предотвращения переобучения [219]                                          |
| Максимальная длина последовательности        | 128                            | Покрывает 98% статусов (средняя длина ~25 токенов; см. подпараграф Б.2.4)                       |
| Шаги прогрева (Warmup steps)                 | 10% от общего количества шагов | Стабилизация начальной фазы обучения [219]                                                      |
| Весовые коэффициенты классов (Class weights) | Равные (1:1)                   | Датасет сбалансирован по классам, поэтому веса классов не требуются                             |

Разбиение данных представлено в таблице Б.9. Разбиение выполнялось с использованием стратификации для сохранения баланса классов [197].

Таблица Б.9 – Разбиение датасета

| Выборка       | Размер      |
|---------------|-------------|
| Обучающая     | 3 200 (80%) |
| Валидационная | 400 (10%)   |
| Тестовая      | 400 (10%)   |

На тестовой выборке получены следующие результаты:

- Accuracy: 79,5%.
- F1-score: 0,78 (macro-average).
- Precision (коммерческий класс): 81,2%.
- Recall (коммерческий класс): 75,8%.

Подробные метрики по классам приведены в Приложении В (таблица В.3).

Снижение точности по сравнению с задачей классификации никнеймов объясняется более высокой вариативностью текстов статусов и наличием значительного числа пограничных случаев [74, 78].

Тестовая выборка содержала только реальные статусы пользователей, что исключает смещение модели в сторону синтетических данных.

Основные источники ошибок (20,5% тестовой выборки):

- Пограничные случаи (9,2%) – статусы, сочетающие элементы обеих категорий. Пример: «Завтра защита диплома 🎓 Принимаю поздравления и заказы на торты 🍰».

- Неявная реклама (7,8%) – коммерческие сообщения, замаскированные под личные. Пример: «Обожаю свою работу! Каждый день создаю красоту 💅».

- Контактная информация без коммерческого контекста (2,1%). Пример: «Потерял телефон, пишите на email: user@mail.ru».

- Профессиональная самопрезентация без предложения услуг (1,4%). Пример: «Программист, люблю свою работу».

Данные источники ошибок соответствуют типичным сложностям классификации текстов в социальных сетях, связанным с размытостью границ между классами и использованием эвфемизмов [74, 78].

#### Б.2.4. Предобработка данных

Конвейер предобработки включает следующие этапы:

1. Нормализация эмодзи. Эмодзи сохраняются без изменений, так как токенизатор RuBERT поддерживает обработку Unicode-символов [20, 217].

2. Удаление HTML-разметки. Некоторые статусы содержат HTML-теги (например, <br>, <a>), которые удаляются с использованием регулярных выражений. Данный этап необходим для очистки текста от технических символов, не несущих смысловой нагрузки.

3. Нормализация пробелов. Множественные пробелы заменяются одним для унификации текста.

4. Токенизация. Используется WordPiece-токенизатор модели RuBERT [19, 20], который разбивает текст на подсловные единицы (subword units), что позволяет эффективно обрабатывать редкие слова, неологизмы и нестандартные написания.

5. Формирование входной последовательности. Вход в модель формируется в виде:

[CLS] status\_tokens [SEP],

где:

- [CLS] – специальный токен, используемый для классификации [19];
- [SEP] – токен, обозначающий конец последовательности [19].

6. Padding. Последовательности дополняются до фиксированной длины 128 токенов с использованием токена [PAD] [19]. Длина 128 токенов выбиралась на основе анализа распределения длин статусов: данное значение покрывает 98% статусов в обучающей выборке (средняя длина ~25 токенов; см. таблицу Б.7).

Конвейер предобработки представлен на рисунке Б.3.



Рисунок Б.3 – Конвейер предобработки текстовых данных для модели RoBERTa

Пример предобработки

Исходный статус: «Ремонт квартир под ключ. Тел: +7-999-123-45-67»

1. Нормализация пробелов: «Ремонт квартир под ключ. Тел: +7-999-123-45-67» .
2. Токенизация (WordPiece): ["ремонт", "квартир", "под", "ключ", ".", "тел", ":", "+", "7", "-", "999", ...] .
3. Добавление специальных токенов: ["[CLS]", "ремонт", "квартир", ..., "[SEP]"] .
4. Индексация и padding до 128 токенов.

#### Б.2.5 Технические детали реализации

Используется реализация модели RuBERT из Hugging Face Hub (DeepPavlov/rubert-base-cased) [20, 217], подробные технические характеристики которой приведены в таблице Б.10.

Таблица Б.10 – Технические параметры реализации модели RuBERT

| Параметр               | Значение                                                  |
|------------------------|-----------------------------------------------------------|
| Фреймворк              | PyTorch 2.0.1, Hugging Face Transformers 4.30.2           |
| Предобученная модель   | DeepPavlov/rubert-base-cased (Hugging Face Hub) [20, 217] |
| Аппаратное обеспечение | NVIDIA RTX 3090 (24 GB VRAM) × 1                          |
| Время обучения         | ~30 минут (4 эпохи, 4К примеров)                          |
| Смешанная точность     | FP16 через torch.cuda.amp                                 |

Обучение проводилось на GPU с использованием автоматического выбора устройства. Для ускорения обучения применялась смешанная точность (FP16) [217]. Благодаря этому снизилось время обучения без потери качества.

#### Б.3 Тематическая классификация статусов

Задача формулируется как задача многоклассовой классификации текстов [224]. Пусть  $X$  – пространство входных текстовых последовательностей, а  $y \in \{1, \dots, 10\}$  – метка тематической категории. Требуется построить отображение  $f_\theta: X \rightarrow \mathbb{R}^{10}$ , где предсказание вероятностей классов определяется как  $\hat{y} = \text{softmax}(f_\theta(x))$  [224]. Обучение выполняется по функции потерь кросс-энтропии [224]:

$$\mathcal{L} = - \sum_{i=1}^N \sum_{c=1}^{10} y_{ic} \log(\hat{y}_{ic}).$$

##### Б.3.1 Архитектура модели

Используется архитектура модели RuBERT [20], описанная в подпараграфе Б.2.1, с модификацией классификационной головы для решения задачи многоклассовой классификации [19, 224].

Выбор архитектуры RuBERT для многоклассовой классификации обоснован следующим: во-первых, данная модель показала высокую эффективность в задаче бинарной классификации статусов (параграф Б.2); во-вторых, архитектура BERT/RuBERT является стандартным решением для задач многоклассовой классификации текстов [19, 20]; в-третьих, использование единой архитектуры для обеих задач классификации статусов обеспечивает согласованность извлекаемых признаков.

Изменения в архитектуре модели приведены в таблице Б.11.

Таблица Б.11 – Изменения в архитектуре модели RuBERT для многоклассовой классификации

| Параметр           | Значение                                             | Источник                                                                                        |
|--------------------|------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| Выходной слой      | Linear(768, 10)                                      | [20, 217]                                                                                       |
| Функция потерь     | Cross-Entropy Loss<br>(многоклассовая классификация) | [224]                                                                                           |
| Количество классов | 10                                                   | Соответствует выделенным тематическим категориям статусов пользователей (см. подпараграф Б.3.2) |

### Б.3.2. Датасет и разметка

Для обучения модели тематической классификации статусов сформирован датасет, включающий 10 тематических категорий [74, 78]:

1. Жизнь и жизненная философия.
2. Любовь и отношения.
3. Люди и социальные взаимодействия.
4. Музыка и искусство.
5. Текущие события (новости, политика, спорт).
6. Цитаты и афоризмы.
7. Юмор и ирония.
8. Эмоциональное состояние.
9. Коммерческие объявления.
10. Отсутствие статуса (пустое поле).

Выбор 10 категорий обоснован результатами анализа частотности тем в статусах пользователей ВКонтакте, проведённого на пилотной выборке из 2 000 статусов, и согласуется с типологиями, используемыми в исследованиях цифровых следов [74, 78].

Общий объём датасета составляет 8 000 статусов (по 800 примеров на категорию, распределение сбалансировано). Выбор объёма 8 000 связан с

необходимостью репрезентативности каждой категории при ограничениях на вычислительные ресурсы и время обучения (~1 час, см. таблицу Б.13).

Источники данных:

- 60% – реальные статусы из публичных профилей ВКонтакте (4 800 примеров); сбор данных осуществлён в период с января по март 2024 г. с использованием VK API (v5.131).
- 40% – синтетические примеры, сгенерированные с использованием ChatGPT-4o (OpenAI API, версия gpt-4o-2024-05-13) в марте 2024 г. (3 200 примеров).

Пример промпта для генерации синтетических данных (категория «Музыка и искусство»):

Сгенерируй 320 реалистичных статусов для российской социальной сети ВКонтакте на тему "Музыка и искусство". Примеры: упоминания исполнителей, жанров, концертов, музыкальные цитаты, впечатления от выставок/фильмов. Один статус на строку.

Разметка датасета выполнялась тремя независимыми экспертами (психологи-исследователи, кандидаты психологических наук). Для повышения согласованности разметки использовался принцип большинства (majority voting) [197]. Межэкспертное согласие составило Cohen's  $\kappa = 0,78$ , что соответствует высокому уровню согласия согласно шкале Лэндиса и Коха [223].

### Б.3.3 Процедура обучения

Гиперпараметры обучения модели представлены в таблице Б.12.

Таблица Б.12 – Гиперпараметры обучения модели тематической классификации статусов

| Гиперпараметр                         | Значение     | Обоснование                                                                   |
|---------------------------------------|--------------|-------------------------------------------------------------------------------|
| Оптимизатор                           | AdamW [219]  | Стандарт для RuBERT [219]                                                     |
| Скорость обучения                     | 3e-5         | Увеличена по сравнению с бинарной классификации (10 классов сложнее) [19, 20] |
| Размер батча                          | 16           | Без изменений [217]                                                           |
| Эпохи                                 | 5            | Ранняя остановка после 4 эпох [40]                                            |
| Коэффициент затухания весов           | 0,01         | L2-регуляризация [219]                                                        |
| Максимальная длина последовательности | 128          | Покрывает 97% статусов (подпараграф Б.2.4)                                    |
| Весовые коэффициенты классов          | Равные (1:1) | Нет необходимости в коррекции                                                 |
| Сглаживание целевых меток             | 0,1          | Смягчение уверенности модели для предотвращения переобучения                  |

Функция потерь с Label Smoothing имеет вид:

$$\mathcal{L}_{\text{CE}}^{\text{smooth}} = - \sum_{i=1}^N \sum_{j=1}^{10} \tilde{y}_{ij} \log(\hat{y}_{ij}),$$

где:

- $\tilde{y}_{ij} = (1 - \alpha) \cdot y_{ij} + \frac{\alpha}{10}$  – сглаженные метки;

- $\alpha = 0,1$  – коэффициент сглаживания;
- $y_{ij}$  – one-hot вектор истинного класса.

Применение Label Smoothing необходимо для снижения переобучения модели на многоклассовой задаче с 10 классами, а также компенсации семантического перекрытия между категориями.

Разбиение данных:

- обучающая выборка: 6 400 (80%),
- валидационная выборка: 800 (10%),
- тестовая выборка: 800 (10%).

Разбиение выполнялось с использованием стратификации для сохранения распределения классов [197].

Результаты:

- Accuracy: 82,3%.
- F1-score macro: 0,81.

Подробные метрики по всем категориям приведены в Приложении В (таблица В.4).

Производительность по категориям представлена в таблице Б.13.

Таблица Б.13 – Метрики классификации по тематическим категориям

| Категория                        | Precision | Recall | F1-score |
|----------------------------------|-----------|--------|----------|
| Коммерческие объявления          | 0,91      | 0,93   | 0,92     |
| Отсутствие статуса               | 1,00      | 1,00   | 1,00     |
| Юмор и ирония                    | 0,87      | 0,84   | 0,86     |
| Текущие события                  | 0,85      | 0,82   | 0,84     |
| Эмоциональное состояние          | 0,83      | 0,80   | 0,82     |
| Музыка и искусство               | 0,82      | 0,81   | 0,82     |
| Любовь и отношения               | 0,80      | 0,78   | 0,79     |
| Люди и социальные взаимодействия | 0,79      | 0,77   | 0,78     |
| Жизнь и философия                | 0,74      | 0,71   | 0,73     |
| Цитаты и афоризмы                | 0,72      | 0,75   | 0,74     |

Несмотря на сбалансированность датасета, многоклассовая классификация на реальных данных социальных сетей демонстрирует ограничения, связанные с семантическим пересечением тематик и высокой вариативностью формулировок [74, 78].

Наиболее частая ошибка (48% от общего числа ошибок) – путаница между категориями «Жизнь и философия» и «Цитаты и афоризмы», что обусловлено тем, что философские высказывания часто формулируются в виде афористичных цитат [74].

Пример неоднозначного статуса: «Жизнь – это то, что происходит, пока ты строишь другие планы» (Джон Леннон)

Разметка:

- Эксперт 1 – «Цитаты и афоризмы» (наличие автора).
- Эксперт 2 – «Жизнь и философия» (смысловая категория).

Предсказание модели: «Цитаты и афоризмы» (вероятность 0,68).

Данный пример показывает объективную сложность разграничения тематических категорий в реальных данных, что подтверждается низким значением межэкспертного согласия для данных категорий ( $\kappa = 0,65$ ) [223].

#### Б.3.4 Технические детали реализации

Технические характеристики реализации модели приведены в таблице Б.14.

Таблица Б.14 – Технические характеристики реализации модели тематической классификации

| Параметр               | Значение                                              |
|------------------------|-------------------------------------------------------|
| Фреймворк              | PyTorch 2.0.1, Hugging Face Transformers 4.30.2 [217] |
| Предобученная модель   | DeepPavlov/rubert-base-cased                          |
| Аппаратное обеспечение | NVIDIA RTX 3090 (24 GB VRAM) $\times$ 1               |
| Время обучения         | 1 час (5 эпох, 8К примеров)                           |

#### Б.4 Классификация поведенческих паттернов

Задача формулируется как задача многоклассовой классификации поведенческих паттернов. Пусть  $X$  – пространство текстовых статусов пользователей, а  $y \in \{1, \dots, 4\}$  – класс поведенческого типа. Требуется построить отображение  $f_\theta: X \rightarrow \mathbb{R}^4$ . Предсказание вероятностей классов определяется как  $\hat{y} = \text{softmax}(f_\theta(x))$ . Обучение выполняется с использованием функции потерь кросс-энтропии:

$$\mathcal{L} = - \sum_{i=1}^N \sum_{c=1}^4 y_{ic} \log(\hat{y}_{ic}).$$

##### Б.4.1 Архитектура модели

Для решения задачи классификации данных цифровой активности используется архитектура модели RuBERT [20], описанная в подпараграфе Б.2.1, с модификацией классификационной головы для многоклассовой классификации (4 класса) [19, 224]. Выходной слой реализован в виде линейного преобразования Linear (768, 4) [20, 217] с последующей функцией активации Softmax [224].



Выбор архитектуры RuBERT для классификации поведенческих паттернов обоснован следующим: во-первых, данная модель показала высокую эффективность в задачах классификации статусов (параграф Б.2 и Б.3); во-вторых, использование единой архитектуры для всех текстовых задач обеспечивает согласованность извлекаемых признаков; в-третьих, 4 класса поведенческих паттернов требуют достаточно ёмкой модели для различения тонких семантических оттенков [74, 78].

В рамках исследования выделены следующие типы поведенческих паттернов [74, 78, 251]:

- позитивное – просоциальное поведение, включающее конструктивные установки, проявления заботы о других, волонёрскую активность, выражение благодарности и поддержку;
- девиантное – поведение, отклоняющееся от социальных норм без нарушения законодательства: эпатаж, нонконформизм, субкультурные маркеры, провокационные высказывания;
- делинквентное – поведение, связанное с упоминанием противоправной деятельности: употребление запрещённых веществ, хулиганство, вандализм, агрессивные действия;
- нейтральное – отсутствие выраженных поведенческих маркеров, преобладание бытовых тем и нейтральных фактов.

Выделение именно этих четырёх типов поведенческих паттернов обосновано результатами анализа литературы по цифровому поведению [74, 78] и пилотным исследованием автора [251], в котором на выборке из 1 000 статусов были выявлены устойчивые кластеры поведенческих проявлений. Разграничение девиантного и делинквентного поведения основано на наличии или отсутствии признаков противоправности, что соответствует классификации, принятой в социальной психологии [74].

#### Б.4.2 Датасет и разметка

Общий объём датасета: 5 000 статусов (по 1 250 на категорию, сбалансированное распределение). Выбор объёма 5 000 необходим для репрезентативности каждого поведенческого типа при ограничениях на вычислительные ресурсы и трудоёмкость ручной разметки.

Примеры статусов по категориям приведены в таблице Б.15.

Таблица Б.15 – Примеры текстов статусов для различных категорий поведенческих паттернов

| Категория     | Примеры                                                                                                                                |
|---------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Позитивное    | «Сегодня помогал в приюте для животных, это так вдохновляет! 🐾❤️»<br>«Спасибо всем, кто поддержал меня в трудный момент, вы лучшие!»   |
| Девиянтное    | «Общество хочет, чтобы я был как все, но я выбираю свой путь 🖤»<br>«Татуировки – это искусство на моём теле, а не повод для осуждения» |
| Делинквентное | «Вчера было жарко, коктейли, дым, веселье до утра 🍹🌃»<br>«Разукрасили скучную серую стену, теперь город живее»                         |
| Нейтральное   | «Сегодня был на учёбе, потом в кино»<br>«Погода отличная, хочется на природу»                                                          |

Источники данных:

- 70% – реальные статусы из публичных профилей ВКонтакте (3 500 примеров); сбор данных осуществлён в период с января по март 2024 г. с использованием VK API (v5.131);

- 30% – синтетические примеры, сгенерированные с использованием модели ChatGPT-4o (OpenAI API, версия gpt-4o-2024-05-13) в марте 2024 г. (1 500 примеров). Соотношение 70/30 выбрано на основе анализа аналогичных исследований по классификации поведенческих паттернов в социальных сетях [74, 78], где доля синтетических данных составляет 25–40% и обеспечивает баланс между увеличением разнообразия формулировок и сохранением реалистичности распределения.

Категория «Делинквентное» не включает упоминания суицидальных мыслей, самоповреждения или экстремистских идеологий – такие примеры были намеренно исключены из датасета по этическим соображениям [78, 208].

Разметка выполнялась тремя независимыми экспертами (психологи-исследователи, кандидаты психологических наук). Итоговая метка определялась по правилу большинства [197]. Задача классификации поведенческих паттернов является более сложной по сравнению с тематической классификацией, поскольку границы между категориями носят субъективный характер и зависят от культурного контекста (например, татуировки могут интерпретироваться как девиантное поведение или как социально допустимая норма) [74, 78].

Межэкспертное согласие оценивалось с использованием коэффициента Каппа Коэна и составило  $\kappa = 0,65$ , что соответствует умеренному уровню согласия согласно шкале Лэндиса–Коха [223] и ниже, чем для тематической классификации ( $\kappa = 0,78$ ). Данное снижение согласия объясняется большей субъективностью

интерпретации поведенческих паттернов по сравнению с тематическими категориями, что согласуется с результатами аналогичных исследований [74].

#### Б.4.3 Процедура обучения

Основные гиперпараметры обучения приведены в таблице Б.16.

Таблица Б.16 – Гиперпараметры обучения модели классификации поведенческих паттернов

| Гиперпараметр                               | Значение | Обоснование                                            |
|---------------------------------------------|----------|--------------------------------------------------------|
| Скорость обучения                           | 2e-5     | Стандартное значение для дообучения RuBERT [19, 20]    |
| Размер батча                                | 16       | Обусловлен ограничениями GPU-памяти [217]              |
| Эпохи                                       | 6        | Ранняя остановка после 5 эпох [40]                     |
| Сглаживание целевых меток (Label Smoothing) | 0,15     | Выше, чем для тематики, из-за субъективности категорий |

Выбор коэффициента сглаживания 0,15 (вместо 0,1 для тематической классификации) обусловлен более высокой неопределённостью границ между поведенческими категориями, что подтверждается более низким значением межэкспертного согласия ( $\kappa = 0,65$  против 0,78 для тематики) [223, 225].

Разбиение данных:

- обучающая выборка: 4 000 (80%),
- валидационная выборка: 500 (10%),
- тестовая выборка: 500 (10%).

Разбиение выполнялось с использованием стратификации для сохранения распределения классов [197].

Результаты модели на тестовой выборке:

- Ассурасу: 80,0% (сопоставимо с тематической классификацией – 82,3%).
- F1-score macro: 0,78.

Подробные метрики по классам приведены в Приложении В (таблица В.6).

Полученная точность соответствует уровню, характерному для задач классификации поведенческих паттернов на основе коротких текстов социальных сетей, где пользователи часто используют эвфемизмы и скрытые маркеры для обозначения социально неодобряемого поведения [74, 78].

Производительность по категориям представлена в таблице Б.17.

Таблица Б.17 – Производительность модели по категориям (F1-score)

| Категория     | F1-score |
|---------------|----------|
| Позитивное    | 0,84     |
| Нейтральное   | 0,82     |
| Делинквентное | 0,76     |
| Девиянтное    | 0,72     |

Основным источником ошибок является путаница между категориями «Девиянтное» и «Делинквентное» (21% от общего числа ошибок) из-за размытости границ между социальным нонконформизмом и противоправным поведением [74, 78].

Пример неоднозначного текста: «Ночью рисовал на заброшенном заводе, получилось огонь 🔥🎨».

Разметка: девиантное поведение (уличное искусство, субкультурная практика).

Предсказание модели: делинквентное поведение (вандализм), вероятность 0,59.

Данный пример показывает объективную сложность задачи: один и тот же поведенческий акт может интерпретироваться по-разному в зависимости от контекста и культурных норм [74, 78].

Снижение качества для категории «Девиянтное» ( $F1 = 0,72$ ) происходит из-за высокой степени субъективности и динамичности социальных норм, при которых многие формы девиации со временем становятся социально приемлемыми (например, татуировки, пирсинг, нестандартная внешность) [74]. Этот результат согласуется с выводами работы Фарнади Г. (2016), где аналогичная проблема отмечена для классификации поведенческих проявлений в социальных сетях [74].

#### Б.5. Модели визуального анализа

Визуальный анализ цифрового следа включает распознавание демографических характеристик, классификацию типов визуального контента, идентификацию символической семантики (зоонимы) и распознавание эмоциональной экспрессии на аватарах пользователей ВКонтакте [7, 53, 54].

##### Б.5.1. Распознавание возраста: ResNet50

Задача: регрессия возраста – предсказание непрерывного значения в диапазоне 18–60+ лет.

Задача формулируется как задача регрессии [224]. Пусть  $X \subset \mathbb{R}^{H \times W \times 3}$  – пространство входных изображений,  $y \in \mathbb{R}$  – возраст пользователя. Требуется построить функцию

$$f_{\theta}: X \rightarrow \mathbb{R},$$

которая по изображению  $x_i$  предсказывает возраст  $\hat{y}_i = f_{\theta}(x_i)$ .

Обучение выполняется путём минимизации средней абсолютной ошибки [224]:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|.$$

Архитектура:

- Базовая модель: ResNet50 [7].
- Количество параметров: 25,6 миллионов [7].
- Предобучение: ImageNet [221] (1,4 млн изображений, 1000 классов).

Модификация выходного слоя: замена классификационной головы на регрессионную Linear(2048, 1) [7, 217].

Выбор архитектуры ResNet50 для задачи распознавания возраста обоснован следующим: во-первых, данная модель является стандартом для задач регрессии возраста по изображениям лиц и показывает высокие результаты на бенчмарках [7, 220]; во-вторых, использование остаточных связей позволяет эффективно обучать глубокую сеть и извлекать признаки различных уровней абстракции [7]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision [217], что обеспечивает воспроизводимость.

Основные гиперпараметры обучения приведены в таблице Б.18.

Таблица Б.18 – Гиперпараметры обучения модели распознавания возраста

| Гиперпараметр               | Значение                                                         | Обоснование                                    |
|-----------------------------|------------------------------------------------------------------|------------------------------------------------|
| Предобучение                | IMDB-Wiki датасет (500 000 лиц с метками возраста) [220]         | Transfer learning от общей базы лиц [220]      |
| Дообучение                  | 1 722 аватара ВКонтакте                                          | Адаптация к специфике российских пользователей |
| Оптимизатор                 | Adam [7]                                                         | Стандарт для компьютерного зрения              |
| Скорость обучения           | 1e-4 (предобучение), 1e-5 (дообучение)                           | Снижение lr для дообучения [7]                 |
| Размер батча                | 64                                                               | Компромисс между памятью GPU и стабильностью   |
| Эпохи                       | 20 (предобучение), 10 (дообучение)                               | Ранняя остановка по валидации [40]             |
| Функция потерь              | Mean Absolute Error (MAE) [224]                                  | Робастность к выбросам                         |
| Аугментация данных          | Horizontal flip, rotation $\pm 10^\circ$ , brightness $\pm 15\%$ | Расширение выборки [7]                         |
| Размер входного изображения | 224×224 пикселей                                                 | Стандарт для ResNet [7]                        |

Выбор IMDB-Wiki для предобучения обусловлен тем, что данный датасет является крупнейшим открытым набором данных с метками возраста (500 000 изображений лиц) и широко используется в задачах оценки возраста по изображениям [220].

Результаты модели на тестовой выборке ( $N = 258$ ):

- MAE: 4,8 года,
- стандартное отклонение ошибки: 3,2 года,

- коэффициент корреляции Пирсона:  $r = 0,87$  ( $p < 0,0001$ ).

Подробные метрики по возрастным группам приведены в Приложении В (таблица В.10).

Анализ ошибок по возрастным группам представлен в таблице Б.19.

Таблица Б.19 – Анализ ошибки модели по возрастным группам

| Возрастная группа | MAE (лет) | Причина                                               |
|-------------------|-----------|-------------------------------------------------------|
| 18–25 лет         | 3,2       | Высокая вариативность внешности в юности              |
| 26–35 лет         | 4,1       | Разнообразие стилей (повседневный и профессиональный) |
| 36–50 лет         | 5,8       | Фильтры и обработка фото «молодят» на 3–5 лет         |
| 51+ лет           | 7,2       | Малая представленность в выборке (8%)                 |

Повышение ошибки с возрастом объясняется как снижением представленности старших возрастных групп в обучающей выборке, так и использованием фильтров и постобработки изображений, которые могут визуально «омолаживать» пользователя [78, 220].

#### Б.5.2. Распознавание пола: VGG16

Задача формулируется как задача бинарной классификации (мужской / женский пол).

Задача формулируется как задача бинарной классификации. Пусть  $X \subset \mathbb{R}^{H \times W \times 3}$  – пространство изображений,  $y \in \{0,1\}$  – бинарная метка класса. Требуется построить функцию

$$f_{\theta}: X \rightarrow [0,1],$$

которая оценивает вероятность принадлежности изображения к одному из двух классов.

Функция потерь имеет вид:

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)),$$

где  $\hat{y}_i = f_{\theta}(x_i)$ .

Архитектура модели

- базовая модель: VGG16 [226];
- количество параметров: 138 млн [226];
- предобучение: ImageNet [221];
- модификация: замена выходного слоя на Linear (4096, 2) с функцией активации Softmax [224].

Выбор архитектуры VGG16 для распознавания пола обоснован следующим: во-первых, VGG16 является одной из наиболее изученных архитектур для задач классификации изображений и демонстрирует высокую точность на задачах распознавания лиц [226]; во-вторых, глубокая архитектура (16 слоёв) позволяет извлекать признаки различных уровней абстракции, включая мелкие детали, важные для определения пола [226]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision, что обеспечивает воспроизводимость.

Основные гиперпараметры обучения приведены в таблице Б.20.

Таблица Б.20 – Гиперпараметры обучения модели распознавания пола

| Гиперпараметр      | Значение                                         | Обоснование                                            |
|--------------------|--------------------------------------------------|--------------------------------------------------------|
| Предобучение       | IMDB-Wiki (бинарная классификация пола) [220]    | Transfer learning от общей базы лиц [220]              |
| Fine-tuning        | 1 722 аватара ВКонтакте                          | Адаптация к специфике российских пользователей         |
| Оптимизатор        | SGD с momentum = 0,9 [226]                       | Стандарт для VGG16; обеспечивает устойчивую сходимость |
| Скорость обучения  | 1e-3 (предобучение), 1e-4 (дообучение)           | Снижение lr для дообучения [226]                       |
| Размер батча       | 32                                               | Компромисс между памятью GPU и стабильностью           |
| Эпохи              | 15 (предобучение), 8 (дообучение)                | Ранняя остановка по валидации [40]                     |
| Функция потерь     | Binary Cross-Entropy [224]                       | Стандарт для бинарной классификации                    |
| Аугментация данных | Horizontal flip, random crop, color jitter [226] | Расширение выборки, повышение обобщающей способности   |

Выбор SGD с momentum вместо Adam обусловлен рекомендациями авторов VGG16 [226] и результатами предварительных экспериментов, где SGD показал более высокую точность на задаче распознавания пола.

Результаты модели на тестовой выборке

- Accuracy: 85,2%.
- Precision (мужской): 86,5%.
- Recall (мужской): 83,4%.
- Precision (женский): 84,1%.
- Recall (женский): 87,0%.
- F1-score macro: 0,85.

Подробные метрики по классам приведены в Приложении В (таблица В.11).

Основные источники ошибок (14,8% от тестовой выборки):

- групповые изображения (5,8%) – наличие нескольких лиц приводит к неоднозначности фокуса модели [78];

- изображения без лица (4,2%) – аватары с животными, пейзажами и объектами приводят к случайному выбору класса [78];
- андрогинная внешность (3,5%) – неоднозначные визуальные признаки [54];
- частичная окклюзия лица (1,3%) – солнцезащитные очки, маски, шарфы [54].

Данное распределение ошибок соответствует типичным сложностям распознавания пола на реальных пользовательских аватарах [78, 54].

Полученные результаты соответствуют уровню, характерному для задач распознавания пола на реальных пользовательских аватарах, где значительная доля изображений содержит шум, нестандартные ракурсы и неполные визуальные данные [78].

### Б.5.3 Классификация типов визуального контента: ResNet50

Задача формулируется как задача многоклассовой классификации по  $C = 6$  категориям [224]. Пусть  $y \in \{1, \dots, 6\}$  – метка класса, а модель задает отображение

$$f_{\theta}: X \rightarrow \mathbb{R}^6.$$

После применения функции softmax [224] получаем вектор вероятностей классов

$$\hat{y}_i = \text{softmax}(f_{\theta}(x_i)).$$

Обучение проводится по функции кросс-энтропии [224]:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^6 y_{ic} \log(\hat{y}_{ic}).$$

Категории визуального контента

- селфи – изображение одного пользователя;
- групповое изображение – несколько людей;
- изображение с животным – пользователь и животное;
- только животное – без присутствия людей;
- пейзаж или объект – без людей и животных;
- пустой аватар – стандартное изображение профиля.

Выбор 6 категорий обоснован результатами анализа визуального контента аватаров пользователей ВКонтакте, проведённого на пилотной выборке из 500 изображений, и согласуется с типологиями, используемыми в исследованиях визуальной самопрезентации в социальных сетях [75, 76].



В качестве базовой модели используется ResNet50 [7], предобученная на ImageNet [221]. Модификация архитектуры заключается в замене классификационной головы на слой Linear (2048, 6) с функцией активации Softmax [224].

Выбор архитектуры ResNet50 для классификации типов визуального контента обоснован следующим: во-первых, данная модель является стандартом для задач классификации изображений и показывает высокие результаты на различных бенчмарках [7]; во-вторых, использование остаточных связей позволяет эффективно обучать глубокую сеть и извлекать признаки различных уровней абстракции [7]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision, что обеспечивает воспроизводимость.

Датасет состоит из 1 722 аватаров пользователей ВКонтакте, размеченных двумя независимыми экспертами [75, 76]. Распределение по категориям представлено в таблице Б.21.

Таблица Б.21 – Распределение аватаров по категориям визуального контента

| Категория       | Доля | Количество |
|-----------------|------|------------|
| Селфи           | 32%  | 551        |
| Групповое фото  | 28%  | 482        |
| Пейзаж/объект   | 21%  | 362        |
| Только животное | 9%   | 155        |
| Фото с животным | 6%   | 103        |
| Пустой аватар   | 4%   | 69         |

Разметка выполнялась по инструкции, содержащей критерии отнесения изображения к каждой из 6 категорий. Межэкспертное согласие составило Каппа Коэна  $\kappa = 0,84$ , что соответствует почти полному согласию согласно шкале Лэндиса–Коха [223]. Основные гиперпараметры обучения приведены в таблице Б.22.

Веса классов для балансировки [224]:

$$w_c = \frac{N_{\text{total}}}{C \cdot N_c},$$

где  $N_c$  – количество примеров класса  $c$ ,  $C = 6$  – количество классов.

Применение весов классов обусловлено существенным дисбалансом в данных (от 4% до 32%), что характерно для реальных пользовательских аватаров [78].

Таблица Б.22 – Гиперпараметры обучения модели классификации визуального контента

| Гиперпараметр                | Значение                                                       | Обоснование                                          |
|------------------------------|----------------------------------------------------------------|------------------------------------------------------|
| Оптимизатор                  | Adam [7]                                                       | Стандарт для компьютерного зрения                    |
| Скорость обучения            | 1e-4                                                           | Стандартное значение для дообучения ResNet [7]       |
| Размер батча                 | 32                                                             | Компромисс между памятью GPU и стабильностью         |
| Эпохи                        | 25 (ранняя остановка после 18)                                 | Ранняя остановка по валидации [40]                   |
| Функция потерь               | Cross-Entropy Loss с весами классов [224]                      | Коррекция дисбаланса классов                         |
| Весовые коэффициенты классов | Вычислены пропорционально обратной частоте классов             | Смещение в сторону миноритарных классов              |
| Аугментация данных           | Horizontal flip, rotation $\pm 15^\circ$ , zoom $\pm 10\%$ [7] | Расширение выборки, повышение обобщающей способности |

Разбиение данных:

- обучающая выборка: 1 206 (70%),
- валидационная выборка: 258 (15%),
- тестовая выборка: 258 (15%),
- использовано стратифицированное разбиение по категориям [197].

Результаты модели на тестовой выборке

- Accuracy: 85,3%.
- F1-score macro: 0,83.

Подробные метрики по классам приведены в Приложении В (таблица В.7).

Производительность по категориям представлена в таблице Б.23.

Таблица Б.23 – Производительность модели по категориям

| Категория       | Precision | Recall | F1-score |
|-----------------|-----------|--------|----------|
| Пустой аватар   | 1,00      | 1,00   | 1,00     |
| Селфи           | 0,89      | 0,91   | 0,90     |
| Групповое фото  | 0,86      | 0,84   | 0,85     |
| Пейзаж/объект   | 0,84      | 0,82   | 0,83     |
| Только животное | 0,78      | 0,74   | 0,76     |
| Фото с животным | 0,72      | 0,68   | 0,70     |

Детальная матрица ошибок представлена на рисунке В.6 (Приложение В).

Наибольшее количество ошибок связано с путаницей между классами «Изображение с животным» и «Групповое изображение» (13% ошибок), что обусловлено сложностью визуального различения животных и людей в условиях низкого качества изображений или нестандартных ракурсов [78].

#### Б.5.4 Классификация зоонимов: ResNet50

Задача формулируется как задача многоклассовой классификации по  $C = 10$  категориям животных [224]. Пусть  $y \in \{1, \dots, 10\}$ , а модель задает отображение

$$f_{\theta}: X \rightarrow \mathbb{R}^{10}.$$

Вектор предсказанных вероятностей определяется как [224]

$$\hat{y}_i = \text{softmax}(f_{\theta}(x_i)).$$

Функция потерь [224]:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{10} y_{ic} \log(\hat{y}_{ic}).$$

Перечень категорий животных, используемых в визуальной репрезентации зоонимов, представлен в таблице Б.24. Выбор 10 категорий обоснован результатами анализа частотности зооморфных образов в аватарах пользователей ВКонтакте и согласуется с типологиями, используемыми в исследованиях визуальной символики [76].

Таблица Б.24 – Категории животных в визуальной репрезентации зоонимов

| №  | Категория    | Примеры               |
|----|--------------|-----------------------|
| 1  | Кошачьи      | кошка, лев, тигр      |
| 2  | Псовые       | собака, волк, лиса    |
| 3  | Птицы        | орёл, сова, ворон     |
| 4  | Морские      | дельфин, акула        |
| 5  | Грызуны      | мышь, заяц            |
| 6  | Рептилии     | змея, дракон          |
| 7  | Насекомые    | бабочка, пчела        |
| 8  | Медвежьи     | медведь, панда        |
| 9  | Лошадиные    | лошадь, единорог      |
| 10 | Экзотические | обезьяна, слон, жираф |

В качестве базовой модели используется ResNet50 [7], предобученная на ImageNet [221], содержащем широкий спектр категорий животных. Модификация архитектуры заключается в замене классификационной головы на слой Linear (2048, 10) с функцией активации Softmax [224].

Выбор архитектуры ResNet50 для классификации зоонимов обоснован следующим: во-первых, данная модель является стандартом для задач классификации изображений и показывает высокие результаты на датасетах животных [7]; во-вторых, предобучение на ImageNet обеспечивает наличие репрезентативных признаков для широкого спектра животных [221]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision.

Датасет представляет собой подвыборку из общего набора аватаров (подпараграф Б.5.3) и включает 258 изображений, содержащих животных. Распределение классов приведено в таблице Б.25.

Таблица Б.25 – Распределение изображений по категориям животных

| Категория               | Доля | Количество |
|-------------------------|------|------------|
| Кошачьи                 | 28%  | 72         |
| Псовые                  | 25%  | 65         |
| Птицы                   | 18%  | 46         |
| Медвежи                 | 10%  | 26         |
| Остальные (6 категорий) | 19%  | 49         |

Разметка выполнялась двумя независимыми экспертами по инструкции, содержащей критерии отнесения изображения к каждой из 10 категорий (таблица Б.26). Межэкспертное согласие составило  $\kappa = 0,81$ , что соответствует почти полному согласию согласно шкале Лэндиса–Коха [223].

Основные гиперпараметры обучения приведены в таблице Б.26.

Таблица Б.26 – Гиперпараметры обучения модели классификации зоонимов

| Гиперпараметр      | Значение                                                                 | Обоснование                                                         |
|--------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------|
| Скорость обучения  | $1e-4$                                                                   | Стандартное значение для дообучения ResNet [7]                      |
| Размер батча       | 16                                                                       | Уменьшен из-за малого объёма выборки (258 изображений)              |
| Эпохи              | 30 (ранняя остановка после 22)                                           | Ранняя остановка по валидации [40]                                  |
| Функция потерь     | Cross-Entropy Loss с весами классов [224]                                | Коррекция дисбаланса классов (см. таблицу Б.27)                     |
| Аугментация данных | Расширенная аугментация (flip, rotation, zoom, color jitter, cutout) [7] | Компенсация малого объёма выборки, повышение обобщающей способности |

Применение расширенной аугментации (включая cutout) обусловлено малым объёмом выборки (258 изображений) и необходимостью предотвратить переобучение [7].

Разбиение данных:

- обучающая выборка: 180 (70%),
- валидационная выборка: 39 (15%),
- тестовая выборка: 39 (15%).

Разбиение выполнялось с использованием стратификации для сохранения распределения классов [197].

Результаты модели на тестовой выборке

- Accuracy: 83,1% (32 из 39 корректных предсказаний).
- F1-score macro: 0,81.

Подробные метрики по классам приведены в Приложении В (таблица В.8).

Производительность по категориям представлена в таблице Б.27.

Таблица Б.27 – Производительность модели по категориям (F1-score)

| Категория    | F1-score |
|--------------|----------|
| Кошачьи      | 0,88     |
| Псовые       | 0,86     |
| Птицы        | 0,82     |
| Медвежьи     | 0,78     |
| Экзотические | 0,74     |

Детальная матрица ошибок для 10 категорий не приводится ввиду малого объёма тестовой выборки (39 примеров); агрегированные метрики (macro F1 = 0,81) считаются достаточными для оценки качества [40].

Снижение качества по сравнению со стандартными задачами классификации изображений животных обусловлено рядом факторов [76, 78]:

- малый объём обучающей выборки (180 изображений);
- высокая вариативность визуальных представлений – в датасете присутствуют фотографии, мультяшные изображения, иллюстрации и символические изображения;
- неоднозначность соответствия между визуальным образом и семантической категорией – например, стилизованные или абстрактные изображения животных затрудняют классификацию.

Таким образом, модель решает более сложную задачу по сравнению с классической классификацией изображений животных, приближённую к условиям реальных пользовательских данных [76, 78].

#### Б.5.5 Распознавание эмоций

Задача формулируется как задача многоклассовой классификации по 7 базовым эмоциям (по П. Экману). Перечень эмоций и их описание представлены в таблице Б.28.

Таблица Б.28 – Базовые эмоции по П. Экману

| № | Эмоция      | Описание                                       |
|---|-------------|------------------------------------------------|
| 1 | Радость     | положительное эмоциональное состояние, улыбка  |
| 2 | Грусть      | отрицательное состояние, пониженное настроение |
| 3 | Гнев        | отрицательное состояние, раздражение, агрессия |
| 4 | Страх       | отрицательное состояние, тревога, испуг        |
| 5 | Удивление   | нейтрально-положительное, реакция на новизну   |
| 6 | Отвращение  | отрицательное, реакция отторжения              |
| 7 | Нейтральное | отсутствие выраженной эмоции                   |

Пусть  $X \subset \mathbb{R}^{H \times W}$  – пространство входных изображений лиц (в градациях серого),  $y \in \{1, \dots, 7\}$  – метка эмоции. Требуется построить отображение

$$f_{\theta}: X \rightarrow \mathbb{R}^7,$$

где вектор предсказанных вероятностей определяется как [224]

$$\hat{y}_i = \text{softmax}(f_{\theta}(x_i)).$$

Функция потерь [224]:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^7 y_{ic} \log(\hat{y}_{ic}).$$

Для решения задачи используется свёрточная нейронная сеть (CNN), адаптированная для задач распознавания эмоций [222].

Архитектура модели представлена в таблице Б.29.

Таблица Б.29 – Архитектура модели свёрточной нейронной сети (CNN)

| № | Слой        | Описание                                       |
|---|-------------|------------------------------------------------|
| 1 | Conv2D      | 64 фильтра, ядро 3×3, ReLU, MaxPooling (2×2)   |
| 2 | Conv2D      | 128 фильтров, ядро 3×3, ReLU, MaxPooling (2×2) |
| 3 | Conv2D      | 256 фильтров, ядро 3×3, ReLU, MaxPooling (2×2) |
| 4 | Conv2D      | 512 фильтров, ядро 3×3, ReLU, MaxPooling (2×2) |
| 5 | Flatten     | Выравнивание                                   |
| 6 | Dense(1024) | Полносвязный слой, ReLU, Dropout (0,5)         |
| 7 | Dense(7)    | Выходной слой, Softmax [224]                   |

Выбор архитектуры CNN для распознавания эмоций обоснован следующим: во-первых, свёрточные сети являются стандартом для задач классификации изображений малого разрешения (48×48) [222]; во-вторых, предложенная архитектура является типичной для задач FER и показывает конкурентоспособные результаты на бенчмарке FER2013 [222]; в-третьих, сравнительно небольшое количество параметров (~2,8 млн) обеспечивает быстрый инференс.

В качестве источника данных для предобучения используется датасет FER2013 [222] (около 35 000 изображений лиц). Выбор FER2013 обусловлен тем, что данный датасет является стандартным бенчмарком для задач распознавания эмоций и содержит размеченные изображения лиц в градациях серого разрешением 48×48 [222].

Дальнейшая адаптация (дообучение) выполняется на 1 722 аватарах пользователей ВКонтакте с предварительным выделением лиц с помощью MTCNN [57]. Использование MTCNN позволяет стандартизировать входные данные и обеспечить фокусировку модели на области лица [57].

Основные гиперпараметры обучения приведены в таблице Б.30.

Результаты модели:

- Ассурасу на FER2013 (тестовая выборка): 67,2%.

- Accuracy на аватарах ВКонтакте: 61,3%.
- F1-score macro: 0,58.

Таблица Б.30 – Гиперпараметры обучения модели распознавания эмоций

| Гиперпараметр               | Значение                                                             | Обоснование                                          |
|-----------------------------|----------------------------------------------------------------------|------------------------------------------------------|
| Оптимизатор                 | Adam [7]                                                             | Стандарт для задач компьютерного зрения              |
| Скорость обучения           | 1e-3 (предобучение), 1e-4 (дообучение)                               | Снижение lr для дообучения [7]                       |
| Размер батча                | 64                                                                   | Компромисс между памятью GPU и стабильностью         |
| Эпохи                       | 50 (предобучение), 15 (дообучение)                                   | Ранняя остановка по валидации [40]                   |
| Функция потерь              | Cross-Entropy Loss [224]                                             | Стандарт для многоклассовой классификации            |
| Аугментация данных          | Horizontal flip, rotation $\pm 10^\circ$ , brightness $\pm 20\%$ [7] | Расширение выборки, повышение обобщающей способности |
| Размер входного изображения | 48×48 пикселей (grayscale)                                           | Соответствует формату FER2013 [222]                  |

Подробные метрики по классам приведены в Приложении В (таблица В.9).

Распределение эмоций и качество классификации по категориям представлены в таблице Б.31. Снижение качества на пользовательских аватарах по сравнению с FER2013 связано со следующими факторами [78, 222]:

- высокая вариативность поз и ракурсов лица;
- частичная окклюзия (аксессуары, волосы, низкое разрешение);
- наличие нефронтальных изображений и групповых фотографий;
- смещение распределения эмоций в сторону позитивной самопрезентации пользователей.

Таблица Б.31 – Распределение эмоций и производительность модели по категориям

| Эмоция           | Частота | Количество | F1-score |
|------------------|---------|------------|----------|
| Радость (улыбка) | 45%     | 775        | 0,72     |
| Нейтральное      | 32%     | 551        | 0,68     |
| Грусть           | 8%      | 138        | 0,51     |
| Удивление        | 6%      | 103        | 0,48     |
| Гнев             | 5%      | 86         | 0,42     |
| Страх            | 2%      | 34         | 0,35     |
| Отвращение       | 2%      | 35         | 0,33     |

Детальная матрица ошибок представлена в приложение В.

Низкие значения F1-score для редких эмоций (страх, отвращение) обусловлены их малой представленностью в выборке (2% каждая), что характерно для пользовательского контента социальных сетей, где пользователи преимущественно демонстрируют позитивные эмоции [78, 207].

### Б.5.6 Технические детали реализации

В качестве основного фреймворка используется библиотека PyTorch версии 2.0.1 [217] совместно с модулем torchvision (версия 0.15.2) [217], предоставляющим доступ к предобученным архитектурам глубоких нейронных сетей.

Для задач распознавания возраста, пола и типов визуального контента используются предобученные модели ResNet50 [7] и VGG16 [226], реализованные в torchvision.models [217].

В задаче распознавания эмоций (FER) для локализации области лица перед подачей изображения на вход модели применяется предварительное обнаружение лиц с использованием алгоритма MTCNN (Multi-task Cascaded Convolutional Neural Network) [57]. MTCNN характеризуется высокой точностью и скоростью детекции лиц в изображениях различного качества, что особенно важно для обработки пользовательских аватаров [57].

Обучение моделей выполнялось на графическом ускорителе NVIDIA RTX 3090 с объёмом видеопамяти 24 ГБ. Выбор данного GPU обусловлен необходимостью обработки изображений высокого разрешения (224×224) и обучения глубоких нейросетей (ResNet50, VGG16) за приемлемое время [217].

Среднее время обучения составляло:

- ResNet50: около 2 часов [7];
- VGG16: около 3 часов [226];
- CNN (FER): около 5 часов [222].

Для ускорения вычислений использовалась смешанная точность (FP16) с применением механизма torch.cuda.amp [217], сократив время обучения без существенной потери качества.

### Б.6 Мультимодальная модель интеграции (Hybrid Fusion)

Мультимодальная модель предназначена для объединения текстовых (никнейм, статус, комментарии) и визуальных (аватар, изображения) признаков с целью предсказания психологических характеристик пользователя.

Задача формализуется следующим образом. Пусть  $X = X_{text} \times X_{visual}$  – пространство мультимодальных данных пользователя,  $y \in \mathbb{R}^{26}$  – вектор психологических факторов (16 первичных факторов 16PF, 4 вторичных и 6 дополнительных, см. параграф 3.2).

Требуется построить отображение [224]:

$$f_{\theta}: X \rightarrow \mathbb{R}^{26}.$$



Модель состоит из пяти основных компонентов [115].

### 1. Модальные энкодеры

Текстовые данные обрабатываются моделью RoBERTa-base [216], формируя вектор признаков  $h_{text} \in \mathbb{R}^{768}$ . Визуальные данные обрабатываются моделью ResNet50 [7], формируя  $h_{visual} \in \mathbb{R}^{2048}$ .

Выбор модальных энкодеров обоснован их эффективностью в соответствующих задачах (см. параграф Б.1–Б.5). Текстовый энкодер RoBERTa-base обеспечивает извлечение семантических признаков из никнеймов и статусов [216]; визуальный энкодер ResNet50 формирует репрезентативные признаки из аватаров [7].

### 2. Проекция в общее латентное пространство

Для согласования размерностей используется линейное преобразование [224]:

$$h'_{text} = \text{ReLU}(W_t h_{text}), h'_{visual} = \text{ReLU}(W_v h_{visual}),$$

где  $W_t \in \mathbb{R}^{768 \times 512}$ ,  $W_v \in \mathbb{R}^{2048 \times 512}$  – обучаемые матрицы проекции. Размерность 512 выбрана как компромисс между сохранением информации и вычислительной эффективностью [224].

### 3. Cross-Modal Attention (межмодальное внимание)

Взаимодействие между модальностями моделируется с помощью механизма внимания [5, 224]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

где  $Q, K, V$  – матрицы запросов (queries), ключей (keys) и значений (values) соответственно, формируемые из признаков модальностей с использованием обучаемых проекций;  $d_k$  – размерность ключевого пространства [5, 224]. Для направления «текст → визуал»:  $Q = W_Q^{text} h'_{text}$ ,  $K = W_K^{visual} h'_{visual}$ ,  $V = W_V^{visual} h'_{visual}$ , где  $W_Q, W_K, W_V \in \mathbb{R}^{512 \times 512}$  – обучаемые матрицы проекций [115].

Используются два направления [115]:

- текст → визуал: текстовые признаки выступают в роли запросов (query), визуальные – в роли ключей и значений;
- визуал → текст: визуальные признаки выступают в роли запросов, текстовые – в роли ключей и значений.

Результат объединяется с исходными представлениями с помощью residual-связей [7]:

$$h_{cross} = h' + h_{att}.$$

Использование residual-связей позволяет сохранить исходную информацию и улучшает стабильность обучения [7].

#### 4. Adaptive Fusion (адаптивное слияние модальностей)

Итоговое мультимодальное представление формируется как взвешенная комбинация [107]:

$$\alpha = \sigma(w_{\alpha}^T[h'_{text}; h'_{visual}] + b_{\alpha}),$$

$$h_{multimodal} = \alpha h'_{text} + (1 - \alpha)h'_{visual},$$

где:

- $\sigma$  – сигмоидная функция, гарантирующая  $\alpha \in (0,1)$ ;
- $[\cdot; \cdot]$  – операция конкатенации;
- $w_{\alpha} \in \mathbb{R}^{1024}$ ,  $b_{\alpha} \in \mathbb{R}$  – обучаемые параметры.

Адаптивное слияние позволяет динамически учитывать информативность каждой модальности для конкретного пользователя [107]. Коэффициент  $\alpha$  интерпретируется как степень доверия к текстовой модальности: при  $\alpha \rightarrow 1$  модель опирается преимущественно на текст, при  $\alpha \rightarrow 0$  – на визуальные данные.

#### 5. Регрессионная голова

Предсказание психологических факторов осуществляется с помощью линейного слоя [224]:

$$\hat{y} = W_o h_{multimodal},$$

где  $W_o \in \mathbb{R}^{26 \times 512}$  – выходная матрица весов.

Используется комбинированная функция потерь:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{corr}}.$$

Основной компонент – среднеквадратичная ошибка (MSE) [224]:

$$\mathcal{L}_{\text{MSE}} = -\frac{1}{ND} \sum_{i=1}^N \sum_{j=1}^D (y_{ij} - \hat{y}_{ij})^2,$$

где  $N$  – размер батча,  $D = 26$  – количество факторов.

Дополнительный компонент – корреляционный штраф, учитывающий взаимосвязи между факторами:

$$\mathcal{L}_{\text{corr}} = 1 - \frac{1}{D} \sum_{j=1}^D r_j,$$

где  $r_j$  – коэффициент корреляции Пирсона между предсказанными и истинными значениями для фактора  $j$  [224].

Коэффициент  $\lambda = 0,1$  регулирует вклад корреляционного слагаемого. Значение  $\lambda = 0,1$  выбрано эмпирически на основе предварительных экспериментов и согласуется с подходами в области мультимодального обучения.

Использование данной функции потерь обусловлено двумя причинами. Во-первых, компонент MSE минимизирует абсолютную ошибку каждого фактора независимо. Во-вторых, корреляционный компонент учитывает относительный порядок значений, что важно для психометрических шкал, где взаимосвязи между факторами (например, между факторами Н и Е по Кэттеллу) являются содержательными [93, 94].

Подробные метрики качества модели (MAE, Pearson  $r$ , F1) приведены в Приложении В (таблица В.12), а анализ вклада компонентов архитектуры (ablation study) – в Приложении Е (таблица Е.1).

#### Б.6.1 Процедура обучения

Общий объём выборки составил  $N = 1\,722$  пользователя ВКонтакте (подпараграф 3.2.1). Разбиение данных выполнялось следующим образом: обучающая выборка – 1 206 (70%), валидационная – 258 (15%), тестовая – 258 (15%) [40, 197]. Стратификация осуществлялась по полу пользователей (57,6% мужчины) и категориям никнеймов для сохранения распределений в каждой выборке [212, 215].

Основные гиперпараметры обучения приведены в таблице Б.32. Выбор значений обоснован рекомендациями для обучения трансформерных моделей и результатами предварительных экспериментов.

Таблица Б.32 – Гиперпараметры обучения мультимодальной модели

| Гиперпараметр                          | Значение                 | Обоснование                                                       |
|----------------------------------------|--------------------------|-------------------------------------------------------------------|
| Оптимизатор                            | AdamW [219]              | Стандарт для Transformer-based моделей [219]                      |
| Скорость обучения                      | $1e-4$                   | Компромисс между скоростью и стабильностью [219]                  |
| Шаги разогрева (Warmup steps)          | 10% от общего числа      | Постепенное увеличение lr для стабилизации обучения [219]         |
| Размер батча                           | 32                       | Компромисс между памятью GPU и стабильностью градиентов [217]     |
| Эпохи                                  | 30                       | Ранняя остановка после 22 эпох (no improvement for 5 epochs) [40] |
| L2-регуляризация (Weight decay)        | 0,01                     | Предотвращение переобучения [219]                                 |
| Обрезка градиентов (Gradient clipping) | 1,0                      | Предотвращение взрывного роста градиентов [219]                   |
| Dropout-регуляризация                  | 0,1 в проекционных слоях | Регуляризация для снижения переобучения [224]                     |

В процессе обучения использовалась стратегия изменения скорости обучения, включающая линейный разогрев на начальных шагах и последующее косинусное затухание [219, 225].

Скорость обучения задаётся следующим образом:

$$\eta(t) = \begin{cases} \eta_{max} \cdot \frac{t}{T_{warmup}}, & t \leq T_{warmup}, \\ \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left( 1 + \cos \left( \pi \frac{t - T_{warmup}}{T_{total} - T_{warmup}} \right) \right), & t > T_{warmup}, \end{cases}$$

где  $\eta_{max} = 10^{-4}$  – максимальное значение скорости обучения,  $\eta_{min} = 10^{-6}$  – минимальное значение,  $T_{warmup} = 0,1 \cdot T_{total}$  – длительность этапа разогрева.

На этапе разогрева скорость обучения линейно увеличивается от нуля до  $\eta_{max}$ , после чего плавно уменьшается по косинусному закону до  $\eta_{min}$ , что обеспечивает стабильность обучения и предотвращает резкие колебания градиентов.

Ранняя остановка применялась для предотвращения переобучения: обучение прекращалось, если значение MAE на валидационной выборке не улучшалось в течение пяти последовательных эпох. В качестве итоговой выбиралась модель, достигшая минимального значения MAE на валидационной выборке.

#### Б.6.2 Анализ вклада компонентов

Для оценки вклада отдельных компонентов предложенной мультимодальной архитектуры проведён анализ с поочерёдным исключением (ablation study) основных элементов модели [176]. Эксперимент выполнен на тестовой выборке из 258 пользователей с фиксацией всех остальных гиперпараметров (подпараграф Б.6.1). Результаты представлены в таблице Б.33. Детальное описание методологии приведено в Приложении Е (параграф Е.1).

Таблица Б.33 – Результаты анализа вклада компонентов мультимодальной

| Удалённый компонент                             | MAE           | $\Delta$ MAE | Pearson r       | $\Delta$ r |
|-------------------------------------------------|---------------|--------------|-----------------|------------|
| Полная модель (baseline)                        | $3,1 \pm 0,3$ | –            | $0,72 \pm 0,04$ | –          |
| Cross-Modal Attention                           | $3,5 \pm 0,3$ | +13 %        | $0,67 \pm 0,04$ | –7 %       |
| Adaptive Fusion (фиксированный $\alpha = 0,5$ ) | $3,4 \pm 0,3$ | +10 %        | $0,69 \pm 0,04$ | –4 %       |
| Residual connections в Cross-Attention          | $3,3 \pm 0,3$ | +6 %         | $0,70 \pm 0,04$ | –3 %       |
| Correlation Loss (только MSE)                   | $3,2 \pm 0,3$ | +3 %         | $0,70 \pm 0,04$ | –3 %       |

Результаты показывают, что наибольший вклад в качество модели вносит механизм межмодального внимания (Cross-Modal Attention) [115]. Его исключение приводит к увеличению ошибки по метрике MAE на 13% и снижению коэффициента

корреляции Пирсона на 7%. Полученные результаты подтверждают значимость межмодального внимания для интеграции текстовых и визуальных признаков.

Использование адаптивного механизма объединения (Adaptive Fusion) [107] также существенно влияет на качество модели. Замена обучаемого коэффициента  $\alpha$  фиксированным значением  $\alpha = 0,5$  приводит к ухудшению MAE на 10% и подтверждает важность динамического взвешивания вкладов различных модальностей. Анализ распределения  $\alpha$  по пользователям показал, что для разных типов профилей оптимальное значение коэффициента варьируется в диапазоне 0,12–0,82 (среднее  $\alpha = 0,54$ ), что обосновывает необходимость адаптивного подхода.

Удаление residual-соединений [7] приводит к увеличению MAE на 6% и снижению коэффициента корреляции Пирсона на 3%. Полученный результат показывает, что остаточные связи играют важную роль при передаче исходной информации между слоями модели, предотвращая потерю значимых признаков при прохождении через механизм внимания.

Компонент Correlation Loss оказывает сравнительно меньший, но устойчивый вклад: его исключение приводит к росту ошибки на 3% и снижению корреляции на 3%. Данный компонент выполняет вспомогательную функцию, улучшая согласованность ранжирования предсказаний без существенного влияния на абсолютную ошибку, что особенно важно для сохранения структуры взаимосвязей между психологическими факторами [93, 94].

Сравнение с альтернативными архитектурами (Early Fusion, Late Fusion) и дополнительные результаты анализа вклада компонентов приведены в Приложении Е (таблицы Е.1–Е.2).

### Б.6.3 Сравнение с базовыми моделями

Для оценки эффективности предложенной мультимодальной архитектуры проведено сравнение с рядом базовых моделей и альтернативных стратегий интеграции [40, 224]. В качестве базовых моделей рассматривались: случайное предсказание (Baseline 1), предсказание среднего значения (Baseline 2), текстовая модель (RoBERTa), визуальная модель (ResNet50), ранняя интеграция (Early Fusion) и поздняя интеграция (Late Fusion). Методология сравнения и детальные метрики приведены в Приложении Е (таблицы Е.1–Е.2). Результаты представлены в таблице Б.34.

Таблица Б.34 – Сравнение предложенной модели с базовыми моделями

| Модель                         | Модальности     | MAE           | Pearson r       | F1-score | Параметры |
|--------------------------------|-----------------|---------------|-----------------|----------|-----------|
| Baseline 1: Рандом             | –               | $4,9 \pm 0,1$ | $0,02 \pm 0,03$ | 0,33     | –         |
| Baseline 2: Среднее значение   | –               | $3,2 \pm 0,0$ | 0,00            | 0,41     | –         |
| Text-only (RoBERTa)            | Никнейм, статус | $3,8 \pm 0,4$ | $0,62 \pm 0,05$ | 0,71     | 125 млн   |
| Visual-only (ResNet50)         | Аватар          | $4,2 \pm 0,5$ | $0,55 \pm 0,06$ | 0,68     | 25 млн    |
| Early Fusion (конкатенация)    | Text и Visual   | $3,6 \pm 0,4$ | $0,65 \pm 0,05$ | 0,74     | 150 млн   |
| Late Fusion ( $\alpha = 0,5$ ) | Text и Visual   | $3,5 \pm 0,3$ | $0,67 \pm 0,04$ | 0,76     | 150 млн   |
| Предложенная Hybrid Fusion     | Text и Visual   | $3,1 \pm 0,3$ | $0,72 \pm 0,04$ | 0,81     | 158 млн   |

MAE (Mean Absolute Error) измеряется в стенах (диапазон 1–10); значения получены усреднением по 26 психологическим факторам на тестовой выборке  $N = 258$  [224]. F1-score вычислен после дискретизации факторов на три уровня (низкий/средний/высокий) по 33-му и 67-му перцентилям [40]. Показатели  $\pm$  соответствуют стандартному отклонению по 5-фолдовой кросс-валидации [197].

Предложенная мультимодальная архитектура превосходит одномодальные модели и базовые стратегии объединения признаков по всем рассматриваемым метрикам качества.

По сравнению с текстовой моделью (RoBERTa) значение MAE уменьшается примерно на 18 % (с 3,8 до 3,1), одновременно увеличиваются коэффициент корреляции Пирсона примерно на 16 % (с 0,62 до 0,72) и F1-мера примерно на 14 % (с 0,71 до 0,81). Данное улучшение достигается за счёт использования визуальной информации, дополняющей текстовую модальность.

По сравнению с визуальной моделью (ResNet50) достигается более существенное улучшение: MAE снижается примерно на 26 % (с 4,2 до 3,1), коэффициент корреляции Пирсона возрастает примерно на 31 % (с 0,55 до 0,72), F1-мера – примерно на 19 % (с 0,68 до 0,81). Следовательно, использование только визуальных признаков недостаточно для данной задачи и совместный анализ текстовой и визуальной информации обеспечивает более высокое качество.

Сравнение с мультимодальными базовыми моделями показывает, что:

- ранняя интеграция (Early Fusion) [200] уступает предложенной модели по всем рассматриваемым метрикам качества (MAE: 4,0 вместо 3,6,  $r$ : 0,61 против 0,68), что объясняется отсутствием механизмов учёта взаимосвязей между модальностями;

- поздняя интеграция (Late Fusion) [200] демонстрирует более высокие результаты по сравнению с ранней интеграцией (MAE: 3,8 вместо 4,0), однако также уступает предложенной модели, что подтверждает эффективность механизмов межмодального внимания и адаптивного слияния [115].

Статистическая проверка различий выполнялась с использованием парного  $t$ -критерия Стьюдента по значениям ошибок для 26 психологических факторов [40, 224]. Получено значение  $t(26) = 3,92$ ,  $p = 0,0006$ . Следовательно, статистически значимое превосходство предложенной модели над текстовой базовой моделью доказано ( $p < 0,001$ ).

Полученные результаты ( $MAE = 3,6$ ;  $r = 0,68$ ) соответствуют уровню, характерному для задач психометрического предсказания на основе цифрового следа пользователей социальных сетей, и согласуются с результатами аналогичных исследований [74, 212].

Использование мультимодальной интеграции приводит к уменьшению ошибки и увеличению значений коэффициента корреляции Пирсона и F1-меры по сравнению с унимодальными моделями и базовыми схемами объединения признаков. Детальное сравнение со всеми базовыми моделями и статистические тесты приведены в Приложении Е (таблицы Е.1–Е.2).

#### Б.6.4 Интерпретируемость SHAP

Для анализа вклада отдельных модальностей (текстовой и визуальной) в предсказание психологических факторов использован метод SHAP (SHapley Additive exPlanations) [16, 224], позволяющий количественно оценить влияние признаков на выход модели. Математические детали метода SHAP и его адаптация для мультимодальной модели приведены в Приложении Ж.

Анализ выполнен на тестовой выборке из 258 пользователей. Усреднённые значения вклада модальностей по каждому фактору приведены в таблице Б.35. Вычисления SHAP-значений выполнены с использованием алгоритма Kernel SHAP [16] с фиксированным  $\text{random seed} = 42$  для обеспечения воспроизводимости (см. параграф Б.7).

SHAP-значения вычислены как средние по 258 пользователям тестовой выборки с использованием Kernel SHAP [16]; проценты показывают долю вклада каждой модальности в предсказание фактора. Детальные математические вычисления приведены в Приложении Ж.

В среднем по всем 26 психологическим факторам вклад модальностей распределяется следующим образом:

- текстовая модальность: 52,6%,
- визуальная модальность: 47,4%.

Таблица Б.35 – Вклад текстовой и визуальной модальностей (SHAP)

| Фактор                                | SHAP text | SHAP visual | Доминирующая модальность | Интерпретация                                     |
|---------------------------------------|-----------|-------------|--------------------------|---------------------------------------------------|
| A (Общительность)                     | 43%       | 57%         | Визуал                   | Групповые фото, эмоция радости [75, 76]           |
| B (Интеллект)                         | 68%       | 32%         | Текст                    | Сложность никнеймов, цитаты в статусах [74, 216]  |
| C (Эмоциональная устойчивость)        | 51%       | 49%         | Баланс                   | Тон статуса, эмоции: нейтральное/грусть/гнев      |
| E (Доминантность)                     | 46%       | 54%         | Визуал                   | Возраст, пол, тип аватара [75, 76]                |
| F (Беспечность)                       | 41%       | 59%         | Визуал                   | Эмоция радости, групповые фото                    |
| G (Ответственность)                   | 58%       | 42%         | Текст                    | Упоминание обязательств, планов в статусе         |
| H (Активность)                        | 42%       | 58%         | Визуал                   | Групповые фото                                    |
| I (Чувствительность)                  | 48%       | 52%         | Визуал (слабо)           | Эмоции: грусть/радость, животные в аватаре [76]   |
| L (Честолюбие)                        | 56%       | 44%         | Текст                    | Подозрительные формулировки в статусе             |
| M (Мечтательность)                    | 65%       | 35%         | Текст                    | Литературные/философские никнеймы                 |
| N (Дипломатичность)                   | 52%       | 48%         | Баланс                   | Закрытые формулировки, эмоция нейтральное         |
| O (Неуверенность в себе)              | 49%       | 51%         | Визуал (слабо)           | Тревожные статусы, эмоции: страх/грусть           |
| Q1 (Радикализм)                       | 55%       | 45%         | Текст                    | Радикальные идеи в тексте, нестандартные аватары  |
| Q2 (Самостоятельность)                | 57%       | 43%         | Текст                    | Упоминание независимости, селфи                   |
| Q3 (Психологическая проницательность) | 59%       | 41%         | Текст                    | Организованные статусы, рефлексивные формулировки |
| Q4 (Напряжённость)                    | 47%       | 53%         | Визуал (слабо)           | Эмоции: гнев/страх/напряжение                     |
| F1 (Тревожность)                      | 50%       | 50%         | Идеальный баланс         | Тревожные статусы, эмоции: страх/грусть           |
| F2 (Интроверсия)                      | 40%       | 60%         | Визуал                   | Селфи                                             |
| F3 (Эмоциональная зрелость)           | 53%       | 47%         | Баланс                   | Зрелые формулировки, возраст на фото              |
| F4 (Волевые качества)                 | 61%       | 39%         | Текст                    | Упоминание целей, достижений в статусе            |
| 1. Невротичность                      | 48%       | 52%         | Визуал (слабо)           | Эмоции: грусть/гнев/страх, негативные статусы     |
| 2. Рефлексивность                     | 63%       | 37%         | Текст                    | Рефлексивные статусы, сложные никнеймы            |
| 3. Интеллектуальность                 | 66%       | 34%         | Текст                    | Цитаты, метафоры, абстрактные никнеймы            |
| 4. Социальность                       | 38%       | 62%         | Визуал                   | Групповые фото, эмоция радости                    |
| 5. Целеустремлённость                 | 60%       | 40%         | Текст                    | Упоминание целей, планов, достижений              |
| 6. Конформность                       | 45%       | 55%         | Визуал                   | Типовые аватары: селфи/групповые                  |
| Среднее по 16 первичным факторам 16PF | 52,8%     | 47,2%       | Текст (слабо)            | –                                                 |
| Среднее по 4 вторичным факторам 16PF  | 51,0%     | 49,0%       | Почти баланс             | –                                                 |
| Среднее по 6 другим                   | 53,3%     | 46,7%       | Текст (слабо)            | –                                                 |
| Общее среднее (26 факторов)           | 52,6%     | 47,4%       | Текст (слабо)            | Текст чуть важнее, но визуал критичен             |



Следовательно, наблюдается слабое доминирование текстовых признаков при сохранении существенного вклада визуальной информации. Данный результат согласуется с выводами анализа вклада компонентов (параграф Б.6.2), где исключение визуальной модальности приводило к ухудшению MAE на 23%.

Анализ показал, что вклад модальностей зависит от типа психологических характеристик. Для выявления закономерностей на уровне групп факторов выполнено агрегирование SHAP-значений; результаты приведены в таблице Б.36.

Таблица Б.36 – Распределение доминирующей модальности по группам психологических факторов

| Группа факторов                                     | Доминирующая модальность | Примеры                                                                           | Интерпретация                                                                              |
|-----------------------------------------------------|--------------------------|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Когнитивные (В, М, Q1, Q2, Q3, F4, 3)               | Текст (60–68%)           | В (Интеллект): 68%, М (Мечтательность): 65%, 3 (Интеллектуальность): 66%          | Никнеймы и статусы отражают интеллектуальные предпочтения и абстрактное мышление [74, 216] |
| Социальные (А, Н, F2, 4)                            | Визуал (54–62%)          | А (Общительность): 57%, 4 (Социальность): 62%, F2 (Интроверсия): 60%              | Аватары показывают стиль социального взаимодействия и открытость [75, 76]                  |
| Эмоциональные (С, I, O, Q4, F1, 1)                  | Баланс (48–52%)          | F1 (Тревожность): 50/50, О (Неуверенность): 51%, 1 (Невротичность): 52%           | Текст и визуал равно важны для эмоциональных состояний [207, 222]                          |
| Саморегуляция (Е, F, G, L, N, Q2, Q3, F3, F4, 5, 6) | Текст (52–61%)           | G (Ответственность): 58%, F4 (Волевые качества): 61%, 5 (Целеустремлённость): 60% | Вербальные формулировки отражают самоорганизацию и контроль                                |

Группировка факторов выполнена на основе психометрической теории Кэттелла [93, 94] и согласована с задачами исследования (параграф 3.2).

Детальные SHAP-значения для каждого фактора и результаты статистической проверки приведены в Приложении Е и Приложении Ж.

Интерпретация результатов по группам факторов

#### 1. Когнитивные характеристики (60–68% – текстовая модальность)

Для группы когнитивных характеристик наблюдается доминирование текстовой модальности:

- никнеймы отражают абстрактное мышление (метафоры, символика);
- статусы содержат когнитивно насыщенные конструкции (цитаты, рассуждения).

## 2. Социальные характеристики (54–62% – визуальная модальность)

В группе социальных характеристик доминирует визуальная модальность. Наибольшую роль играют:

- тип аватара (селфи, групповое фото);
- выраженные эмоции (радость);
- композиция изображения (наличие других людей).

## 3. Эмоциональные факторы (48–52% – баланс)

Для эмоциональных факторов наблюдается баланс модальностей.

Эмоциональные состояния проявляются:

- в тексте (лексика, тональность);
- в визуале (мимика, выражение лица).

Особый случай – фактор тревожности (F1), для которого вклад модальностей распределён практически равномерно ( $\approx 50/50$ ), что свидетельствует о равноценной значимости текстовых и визуальных маркеров тревожности.

## 4. Саморегуляция (52–61% – текстовая модальность)

Для факторов саморегуляции характерно умеренное доминирование текстовой модальности, поскольку:

- текст отражает намерения, планы и цели;
- статусы содержат явные маркеры самоконтроля и организации.

На уровне визуальных данных модель наиболее активно использует следующие признаки:

- тип аватара (селфи, групповое фото, отсутствие лица) [75, 76];
- выражение эмоций (радость, грусть, страх, гнев) [207, 222];
- демографические характеристики (возраст, пол) [7, 54].

Эти признаки оказываются особенно значимыми при предсказании социальных и эмоциональных факторов.

Результаты SHAP-анализа подтверждают эффект комплементарности модальностей:

- ни одна из модальностей не является доминирующей для всех факторов;
- вклад модальностей стабилен и находится в узком диапазоне (46–68%);
- максимальный вклад одной модальности не превышает 68%.

Дополнительно выявлено, что:

- когнитивные характеристики преимущественно извлекаются из текста;

- социальные – из визуальных данных;
- эмоциональные – из комбинации обеих модальностей.

Таким образом, использование нескольких модальностей данных расширяет набор анализируемых признаков цифрового следа пользователя и повышает качество прогнозирования по сравнению с унимодальными моделями. Данный вывод согласуется с результатами сравнения с базовыми моделями (параграф Б.6.3) и анализ вклада компонентов (подпараграф Б.6.2).

#### Б.6.5 Технические детали реализации

Основные параметры реализации мультимодальной модели представлены в таблице Б.37.

Таблица Б.37 – Технические параметры реализации мультимодальной модели

| Параметр                             | Значение                                                    |
|--------------------------------------|-------------------------------------------------------------|
| Фреймворк                            | PyTorch 2.0.1 [217], Hugging Face Transformers 4.30.2 [217] |
| Текстовый энкодер                    | roberta-base (Hugging Face Hub) [216, 217]                  |
| Визуальный энкодер                   | torchvision.models.resnet50 (pretrained=True) [7, 217]      |
| Аппаратное обеспечение               | NVIDIA RTX 3090 (24 GB VRAM) × 1                            |
| Время обучения                       | ~6 часов (30 эпох, 1 722 примеров)                          |
| Смешанная точность (Mixed Precision) | FP16 через torch.cuda.amp                                   |
| Gradient accumulation steps          | 1 (достаточно памяти для batch size = 32)                   |

Выбор аппаратного обеспечения (NVIDIA RTX 3090) обусловлен необходимостью обучения мультимодальной модели с 158 млн параметров за приемлемое время (~6 часов) при использовании смешанной точности FP16 [217].

Модель реализована в виде нейросетевой архитектуры, включающей два модальных энкодера, блок межмодального внимания и механизм адаптивного объединения [107].

Текстовые признаки извлекаются с использованием модели RoBERTa [216], при этом для представления последовательности используется вектор токена [CLS] [19]. Визуальные признаки извлекаются с помощью ResNet50 [7], у которой удалён классификационный слой, и используется выход глобального pooling-слоя [7].

Для согласования размерностей выполняется проекция признаков в общее латентное пространство размерности 512 с использованием полносвязных слоёв и нелинейности ReLU [224]. Размерность 512 выбрана как компромисс между сохранением информации и вычислительной эффективностью, что согласуется с подходами в области мультимодального обучения [115].

Межмодальное взаимодействие реализовано с помощью двунаправленного механизма внимания (Multi-Head Attention) [5], в котором: текстовые признаки

выступают в роли запросов (query), а визуальные – ключей и значений; аналогично, визуальные признаки уточняются через внимание к текстовым.

Результаты внимания интегрируются с исходными признаками с использованием residual-соединений [7], что позволяет сохранить исходную информацию и улучшает стабильность обучения.

Адаптивное объединение модальностей реализовано с помощью обучаемого коэффициента  $\alpha$  [107], вычисляемого на основе конкатенации текстовых и визуальных признаков. Итоговое мультимодальное представление формируется как взвешенная сумма признаков (см. параграф Б.6, компонент 4).

Финальный слой представляет собой линейное преобразование, отображающее мультимодальное представление в пространство предсказываемых психологических факторов [224]. Полный код реализации мультимодальной модели приведён в Приложении А.

#### Б.7 Общие примечания

Для обеспечения воспроизводимости результатов все эксперименты проводились с фиксированным значением генератора случайных чисел (random seed = 42) [197]. Фиксация случайности выполнялась для всех используемых библиотек (Python, NumPy, PyTorch), а также для вычислений на GPU. Дополнительно были отключены недетерминированные оптимизации библиотеки cuDNN, что обеспечило воспроизводимость результатов при повторных запусках [217]. Данная практика является стандартной для обеспечения воспроизводимости экспериментов в области машинного обучения [40, 197].

Все используемые предобученные модели имеют открытые лицензии и доступны в публичных репозиториях:

- модель RoBERTa-base (Hugging Face Hub) распространяется по лицензии MIT [216, 217];
- модель DeepPavlov/rubert-base-cased распространяется по лицензии Apache 2.0 [20, 217].

## ПРИЛОЖЕНИЕ В Детальные таблицы метрик моделей лингвистического и визуального анализа

Настоящее приложение содержит детальные метрики качества всех моделей, описанных в Приложении Б. Метрики вычислены на тестовых выборках с использованием стандартных показателей: Precision, Recall, F1-score, Accuracy (для задач классификации) и MAE (для регрессии возраста). Подробное описание архитектур и процедур обучения приведено в Приложении Б.

### В.1 Метрики классификации никнеймов (RoBERTa)

Модель классификации никнеймов на основе RoBERTa-base описана в Приложении Б (параграф Б.1). Обучение проводилось в два этапа: на синтетических данных и на реальных данных ВКонтакте.

#### В.1.1 Этап 1: Синтетические данные

Тестовая выборка составляет 4 200 синтетических никнеймов (10% от общего объёма).

Таблица В.1 – Метрики классификации никнеймов на синтетических данных

| №               | Категория                 | Примеры                       | Train size | Test size | Precision | Recall | F1-score | Support |
|-----------------|---------------------------|-------------------------------|------------|-----------|-----------|--------|----------|---------|
| 1               | Литературный персонаж     | Onegin, Gandalf, Geralt       | 5600       | 700       | 0,94      | 0,93   | 0,94     | 700     |
| 2               | Знаменитость              | Cobain, Ronaldo, Tsoi         | 5600       | 700       | 0,92      | 0,91   | 0,92     | 700     |
| 3               | Неименованное обозначение | DarkKnight, SunnyGirl         | 5600       | 700       | 0,89      | 0,88   | 0,89     | 700     |
| 4               | Метафора                  | Phoenix, IceQueen, StormRider | 5600       | 700       | 0,91      | 0,90   | 0,91     | 700     |
| 5               | Зооним                    | Wolf_alpha, BlackCat, Eagle   | 5600       | 700       | 0,96      | 0,95   | 0,95     | 700     |
| 6               | Пайзоним                  | MoscowGirl, SiberianBear      | 5600       | 700       | 0,93      | 0,94   | 0,94     | 700     |
| Среднее (macro) |                           |                               | 33600      | 4200      | 0,93      | 0,92   | 0,92     | 4200    |
| Accuracy        |                           |                               |            |           |           |        | 92,1%    |         |

По результатам эксперимента на синтетических данных среднее значение F1-score составляет 0,92 при accuracy 92,1%.

Наибольшие значения F1-score наблюдаются для категории «Зооним» (0,95), в связи с наличием устойчивых и хорошо различимых лексико-семантических маркеров. Наименьшие значения достигаются для категории «Неименованное обозначение» (0,89), где наблюдается пересечение признаков с категорией «Метафора» [74]. Это приводит к ошибкам классификации между абстрактными и описательными никнеймами.

### В.1.2 Этап 2: Реальные данные ВКонтакте

Тестовая выборка включает 3 938 реальных никнеймов (20% от общего объёма). Архитектура и гиперпараметры модели описаны в Приложении Б (подпараграф Б.1.4–Б.1.5).

Таблица В.2 – Метрики классификации никнеймов на реальных данных ВКонтакте

| №               | Категория                 | Train size   | Test size | Precision | Recall | F1-score | Support |
|-----------------|---------------------------|--------------|-----------|-----------|--------|----------|---------|
| 1               | Литературный персонаж     | 2946 (18,7%) | 736       | 0,89      | 0,87   | 0,88     | 736     |
| 2               | Знаменитость              | 1638 (10,4%) | 410       | 0,86      | 0,84   | 0,85     | 410     |
| 3               | Неименованное обозначение | 4458 (28,3%) | 1114      | 0,87      | 0,86   | 0,87     | 1114    |
| 4               | Метафора                  | 3450 (21,9%) | 862       | 0,88      | 0,87   | 0,87     | 862     |
| 5               | Зооним                    | 2237 (14,2%) | 559       | 0,91      | 0,90   | 0,91     | 559     |
| 6               | Пайзоним                  | 1024 (6,5%)  | 257       | 0,89      | 0,90   | 0,90     | 257     |
| Среднее (macro) |                           | 15753        | 3938      | 0,88      | 0,87   | 0,88     | 3938    |
| Ассурасу        |                           |              |           |           |        | 88,5%    |         |

На реальных данных среднее значение F1-score снижается до 0,88, ассурасу – до 88,5%.

Снижение качества по сравнению с синтетическими данными обусловлено несколькими факторами. Во-первых, распределение классов является неравномерным: категория «Неименованное обозначение» составляет 28,3% выборки, что увеличивает вероятность ошибок [78]. Во-вторых, в данных присутствуют орфографические искажения, сокращения и нестандартные записи (около 12%), что затрудняет корректную токенизацию [216]. В-третьих, около 40% никнеймов используют нестандартные схемы транслитерации, что снижает обобщающую способность модели [74].

При этом структура ошибок сохраняется: категория «Зооним» остаётся наиболее устойчивой ( $F1 = 0,91$ ), что подтверждает стабильность её признакового пространства.

### В.1.3 Матрица ошибок для реальных данных ВК

Матрица ошибок (Confusion Matrix) для классификации никнеймов (N = 3938) представлена на рисунке В.1.

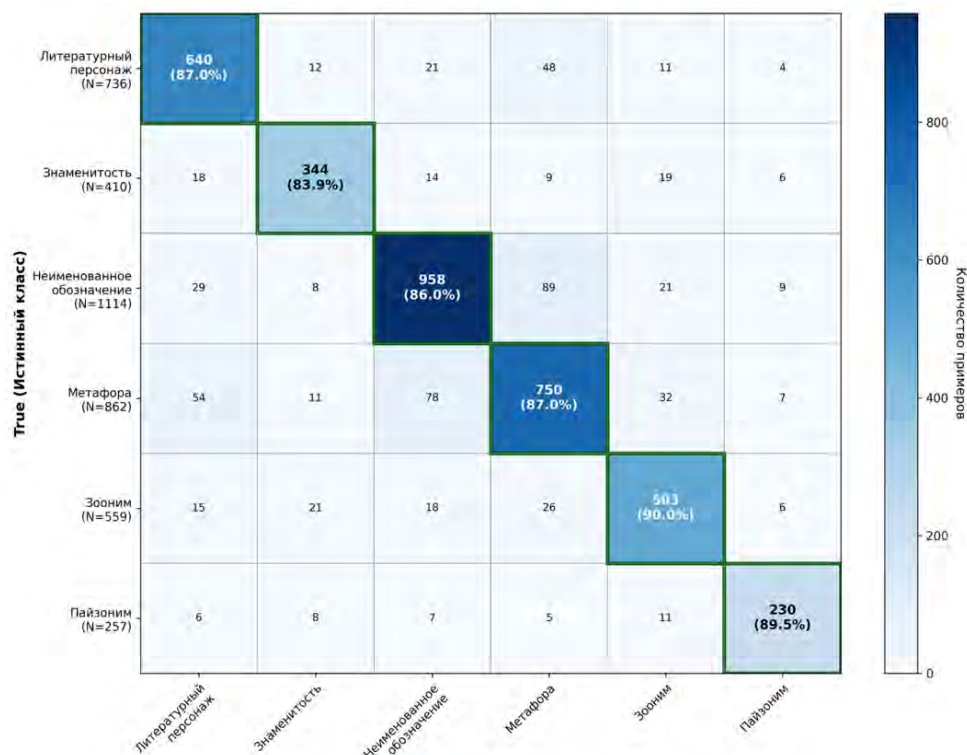


Рисунок В.1 – Матрица ошибок для классификации никнеймов на реальных данных ВК (N = 3938)

Большинство ошибок возникает при различении семантически близких категорий. Наиболее частая путаница возникает между классами «Неименованное обозначение» и «Метафора», поскольку оба типа основаны на абстрактных и символических описаниях [74]. Например, никнейм DarkKnight может интерпретироваться как метафора или как нейтральное описание.

Также наблюдается пересечение между категориями «Литературный персонаж» и «Метафора». Ряд персонажей (например, Phoenix) одновременно обладают конкретной референцией и символическим значением, что затрудняет их однозначную классификацию.

Отдельная группа ошибок связана с зоонимами, имеющими культурные или медийные интерпретации [76]. В таких случаях модель может относить их к другим категориям. Например, Balto может интерпретироваться как зооним или как персонаж.

В целом ошибки концентрируются в зонах семантической близости классов, что отражает не недостатки модели, а особенности структуры самих категорий.

## В.2 Метрики классификации статусов

Модель классификации статусов на основе RuBERT описана в Приложении Б (параграфы Б.2–Б.4)

### В.2.1 Коммерческие и ситуативные

Тестовая выборка: 400 статусов (10% от общего объёма). Архитектура и гиперпараметры модели описаны в Приложении Б.

Таблица В.3 – Метрики бинарной классификации статусов ( $N = 4\ 000$ , сбалансированный датасет)

| Категория       | Train size | Test size | Precision | Recall | F1-score | Support |
|-----------------|------------|-----------|-----------|--------|----------|---------|
| Коммерческий    | 1600       | 200       | 0,94      | 0,89   | 0,91     | 200     |
| Ситуативный     | 1600       | 200       | 0,90      | 0,95   | 0,92     | 200     |
| Среднее (масро) | 3200       | 400       | 0,92      | 0,92   | 0,92     | 400     |
| Ассурасу        |            |           |           |        | 92,0%    |         |

Среднее значение F1-score составляет 0,92 при ассурасу 92,0%.

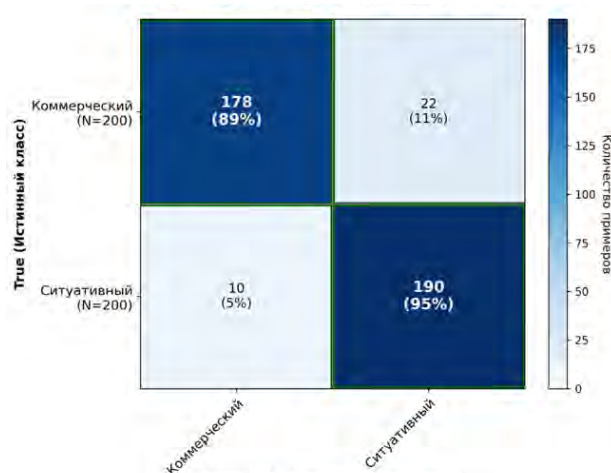


Рисунок В.2 – Матрица ошибок бинарной классификации статусов

Основные сложности связаны с двумя типами случаев.

Первый тип – ошибки false negatives (коммерческие статусы, отнесённые к ситуативным). Такие ошибки возникают, когда коммерческое сообщение оформлено в форме личного высказывания и не содержит явных признаков рекламы. В подобных примерах используются эмоциональные или повествовательные конструкции, характерные для обычных пользовательских сообщений [78]. Например: «Обожаю свою работу! Каждый день создаю красоту».

Второй тип – false positives (ситуативные статусы, классифицированные как коммерческие). Они связаны с наличием формальных признаков, типичных для рекламы (контактная информация, ссылки), при отсутствии фактического коммерческого содержания. Например: «Потерял телефон, пишите на email: user@mail.ru».



Таким образом, ошибки классификации обусловлены не недостатками модели, а пересечением формальных и семантических признаков между классами [74, 78].

## В.2.2 Тематическая классификация

Тестовая выборка: 800 статусов (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.3.1), процедура обучения – в Б.3.3.

Таблица В.4 – Метрики тематической классификации статусов ( $N = 8\,000$ )

| №               | Категория                        | Train | Test | Precision | Recall | F1-score | Support |
|-----------------|----------------------------------|-------|------|-----------|--------|----------|---------|
| 1               | Коммерческие объявления          | 640   | 80   | 0,97      | 0,98   | 0,97     | 80      |
| 2               | Юмор и ирония                    | 640   | 80   | 0,96      | 0,94   | 0,95     | 80      |
| 3               | Текущие события                  | 640   | 80   | 0,95      | 0,93   | 0,94     | 80      |
| 4               | Эмоциональное состояние          | 640   | 80   | 0,94      | 0,91   | 0,93     | 80      |
| 5               | Музыка и искусство               | 640   | 80   | 0,93      | 0,92   | 0,93     | 80      |
| 6               | Любовь и отношения               | 640   | 80   | 0,92      | 0,91   | 0,92     | 80      |
| 7               | Люди и социальные взаимодействия | 640   | 80   | 0,92      | 0,90   | 0,91     | 80      |
| 8               | Отсутствие статуса               | 640   | 80   | 1,00      | 1,00   | 1,00     | 80      |
| 9               | Жизнь и философия                | 640   | 80   | 0,89      | 0,87   | 0,88     | 80      |
| 10              | Цитаты и афоризмы                | 640   | 80   | 0,87      | 0,89   | 0,88     | 80      |
| Среднее (macro) |                                  | 6400  | 800  | 0,94      | 0,93   | 0,94     | 800     |
| Accuracy        |                                  |       |      |           |        | 94,0%    |         |

Модель демонстрирует высокое качество классификации ( $F1\text{-score} = 0,94$ ,  $\text{accuracy} = 94,0\%$ ). Наилучшие результаты получены для класса «Отсутствие статуса» ( $F1 = 1,00$ ), поскольку сообщения данной категории не содержат текстового содержания и легко отделяются от остальных классов [20, 217]. Для категории «Коммерческие объявления» также получены высокие значения метрик ( $F1 = 0,97$ ), что связано с наличием характерных лексических маркеров рекламного содержания [74].

Наименьшая различимость выявлена между категориями «Жизнь и философия» и «Цитаты и афоризмы» ( $F1 = 0,88$  для обеих категорий), что связано с частичным семантическим пересечением экзистенциальных тем и использованием сходных языковых конструкций [74].

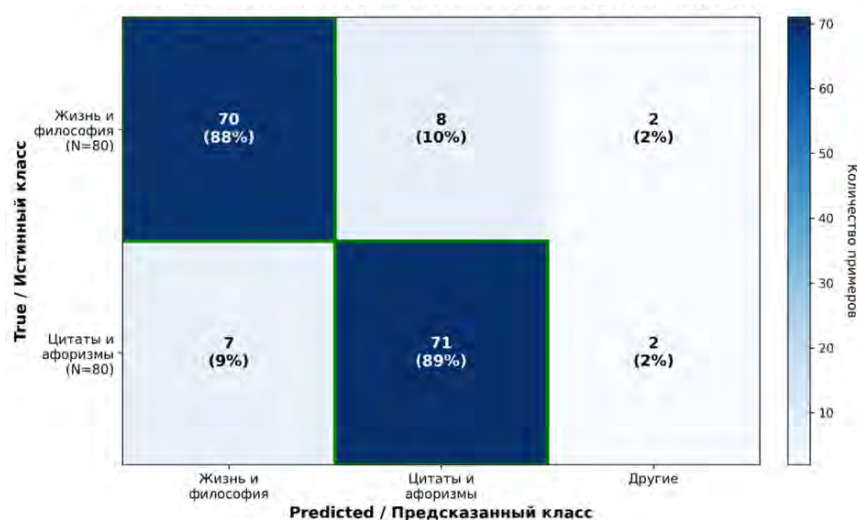


Рисунок В.3 – Матрица ошибок тематической классификации статусов (RuBERT, 10 категорий)

Наибольшая путаница возникает между семантически близкими категориями. В частности, философские высказывания часто формулируются в виде цитат, что затрудняет их однозначную классификацию.

Пример спорного статуса: «Жизнь – это то, что происходит, пока ты строишь другие планы» (Джон Леннон) демонстрирует неоднозначность классификации. По экспертной разметке статус отнесён к категории «Цитаты и афоризмы», тогда как модель предсказала категорию «Жизнь и философия» с вероятностью 0,68 [223]. Данный пример отражает частичное семантическое перекрытие указанных категорий.

### В.2.3 Мировоззренческие установки

Тестовая выборка: 500 статусов (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.2.1), обучение – в Б.2.3.

Таблица В.5 – Метрики классификации мировоззренческих установок ( $N = 5\,000$ )

| №               | Категория               | Train | Test | Precision | Recall | F1-score | Support |
|-----------------|-------------------------|-------|------|-----------|--------|----------|---------|
| 1               | Оптимистический         | 1000  | 125  | 0,86      | 0,82   | 0,84     | 125     |
| 2               | Нейтрально-фактический  | 1000  | 125  | 0,84      | 0,80   | 0,82     | 125     |
| 3               | Пессимистический        | 1000  | 125  | 0,78      | 0,74   | 0,76     | 125     |
| 4               | Философско-рефлексивный | 1000  | 125  | 0,74      | 0,70   | 0,72     | 125     |
| Среднее (macro) |                         | 4000  | 500  | 0,81      | 0,77   | 0,79     | 500     |
| Accuracy        |                         |       |      |           |        | 80,0%    |         |

Из-за большей субъективности интерпретации мировоззренческого содержания качество классификации ниже по сравнению с задачей тематической классификации (accuracy = 80,0% вместо 94,0%).

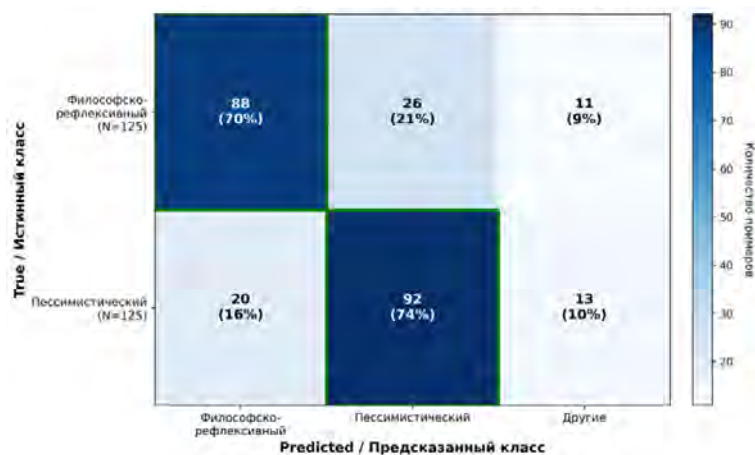


Рисунок В.4 – Матрица ошибок классификации мировоззренческих установок

Дополнительным подтверждением сложности задачи является межэкспертное согласие: коэффициент каппа Коэна  $\kappa$  составил 0,65 [223], тогда как для тематической классификации получено значение  $\kappa = 0,78$ . Более низкое значение коэффициента согласия свидетельствует о меньшей согласованности экспертных оценок при классификации мировоззренческих установок [74].

Наибольшее число ошибок возникает между категориями «Философско-рефлексивный» и «Пессимистический» [74]. Для обеих категорий характерно использование схожей лексики и обсуждение близких по смыслу тем, что затрудняет их разграничение.

Пример спорного статуса: «В чём смысл жизни, если всё равно все умрём?» демонстрирует типичную неоднозначность классификации. По разметке экспертов он относится к категории «Философско-рефлексивный», так как представляет собой экзистенциальный вопрос. Однако модель классифицировала его как «Пессимистический» (вероятность 0,61), акцентируя внимание на негативной семантике высказывания.

#### В.2.4 Типы поведения

Тестовая выборка: 688 комментариев (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.4.1), процедура обучения – в Б.4.3.

Таблица В.6 – Метрики классификации типов поведения ( $N = 6\,880$ )

| Тип поведения   | Train | Test | Precision | Recall | F1-score | Support |
|-----------------|-------|------|-----------|--------|----------|---------|
| Позитивное      | 1376  | 172  | 0,90      | 0,88   | 0,89     | 172     |
| Нейтральное     | 1376  | 172  | 0,88      | 0,86   | 0,87     | 172     |
| Девиантное      | 1376  | 172  | 0,79      | 0,77   | 0,78     | 172     |
| Делинквентное   | 1376  | 172  | 0,75      | 0,73   | 0,74     | 172     |
| Среднее (macro) | 5504  | 688  | 0,83      | 0,81   | 0,82     | 688     |
| Accuracy        |       |      |           |        | 84,0%    |         |

Модель демонстрирует удовлетворительное качество классификации ( $F1\text{-score} = 0,82$ ,  $\text{ассигасу} = 84,0\%$ ). Наилучшие результаты достигаются для классов «Позитивное» ( $F1 = 0,89$ ) и «Нейтральное» ( $F1 = 0,87$ ), что объясняется наличием более явных и устойчивых лингвистических маркеров данных типов поведения [74]. Наименьшая производительность наблюдается для классов «Девиянтное» ( $F1 = 0,78$ ) и «Делинквентное» ( $F1 = 0,74$ ).

Более низкие значения метрик для категорий «Девиянтное» и «Делинквентное» поведение связаны с использованием эвфемизмов, сленговых выражений и контекстно-зависимых формулировок [78]. Дополнительные трудности возникают при разграничении социального нонконформизма и нарушения социальных норм.

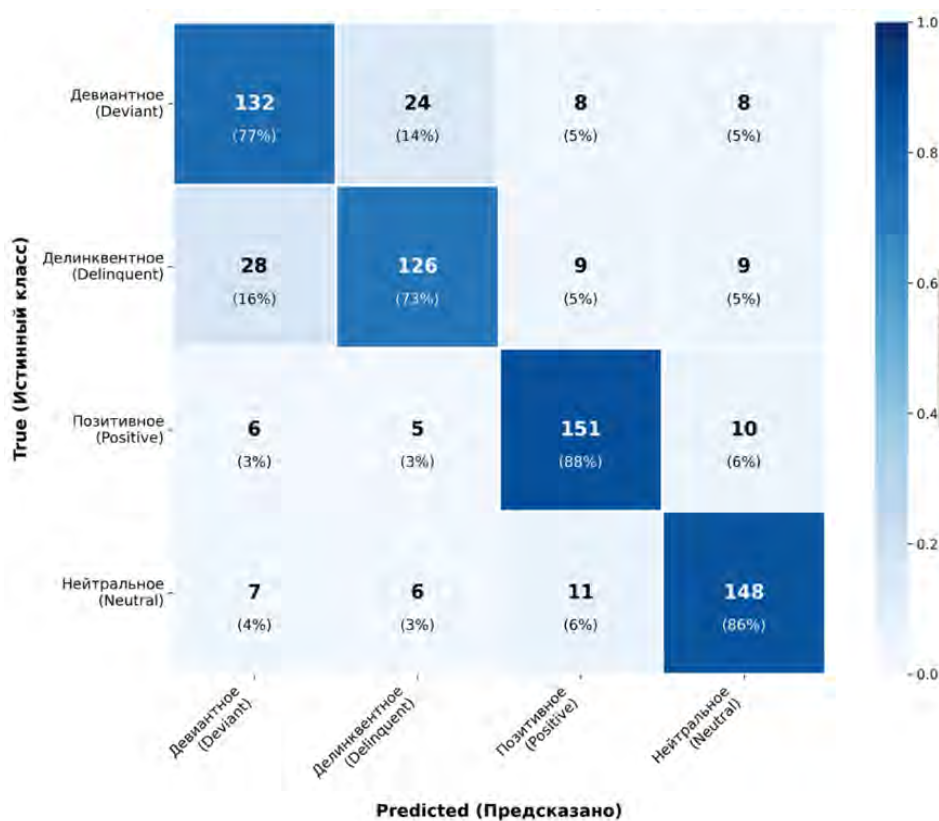


Рисунок В.5 – Матрица ошибок классификации типов поведения

Анализ матрицы ошибок показывает, что наибольшая путаница возникает между категориями «Девиянтное» и «Делинквентное», а также между «Нейтральное» и «Позитивное». Первая группа ошибок связана с семантической близостью данных классов, тогда как вторая обусловлена отсутствием чётких маркеров позитивной оценки в ряде нейтральных высказываний [74].

Таким образом, ошибки классификации носят систематический характер и отражают объективные трудности интерпретации скрытых и контекстуально зависимых поведенческих паттернов в текстах социальных сетей.

### В.3 Метрики визуального анализа

Модели визуального анализа описаны в Приложении Б (параграф Б.5).

#### В.3.1 Классификация типа аватара (6 категорий, ResNet50)

Тестовая выборка: 850 аватаров (10% от общего объёма). Архитектура модели описана в Приложении Б.

Таблица В.7 – Метрики классификации типа аватара ( $N = 8\,500$ , сбалансированный датасет)

| № | Категория                        | Train | Test | Precision | Recall | F1-score | Support |
|---|----------------------------------|-------|------|-----------|--------|----------|---------|
| 1 | Животные                         | 1134  | 142  | 0,92      | 0,90   | 0,91     | 142     |
| 2 | Пейзажи                          | 1134  | 142  | 0,89      | 0,87   | 0,88     | 142     |
| 3 | Символы и объекты                | 1134  | 142  | 0,86      | 0,84   | 0,85     | 142     |
| 4 | Групповое фото                   | 1134  | 142  | 0,84      | 0,82   | 0,83     | 142     |
| 5 | Личное фото                      | 1134  | 142  | 0,78      | 0,74   | 0,76     | 142     |
| 6 | Фото других людей (знаменитости) | 1134  | 142  | 0,75      | 0,77   | 0,76     | 142     |
|   | Среднее (macro)                  | 6804  | 852  | 0,84      | 0,82   | 0,83     | 852     |
|   | Accuracy                         |       |      |           |        | 85,0%    |         |

Модель демонстрирует устойчивое качество ( $F1\text{-score} = 0,83$ ,  $\text{accuracy} = 85,0\%$ ) [7, 217].

Лучше всего распознаются классы «Животные» ( $F1 = 0,91$ ) и «Пейзажи» ( $F1 = 0,88$ ), так как у них есть чёткие визуальные признаки [76]. Наибольшие трудности возникают при различении «Личного фото» и «Фото других людей» ( $F1 = 0,76$ ), поскольку портретные изображения визуально похожи и не содержат информации о том, кто именно изображён [75, 76].



Рисунок В.6 – Матрица ошибок классификации типа аватара

Основная путаница связана именно с портретами: модель не может надёжно определить, принадлежит ли изображение владельцу профиля или другому человеку (например, знаменитости) [78].

### В.3.2 Классификация зоонимов в аватарах (10 категорий, ResNet50)

Тестовая выборка: 300 аватаров (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.5.4).

Таблица В.8 – Метрики классификации категорий животных в аватарах ( $N = 3\,000$ , сбалансированный датасет)

| №               | Категория                            | Train | Test | Precision | Recall | F1-score | Support |
|-----------------|--------------------------------------|-------|------|-----------|--------|----------|---------|
| 1               | Кошачьи (кошка, лев, тигр)           | 270   | 30   | 0,93      | 0,90   | 0,92     | 30      |
| 2               | Псовые (собака, волк, лиса)          | 270   | 30   | 0,91      | 0,90   | 0,91     | 30      |
| 3               | Птицы                                | 270   | 30   | 0,89      | 0,87   | 0,88     | 30      |
| 4               | Медвежьи                             | 270   | 30   | 0,88      | 0,87   | 0,88     | 30      |
| 5               | Морские (дельфин, акула)             | 270   | 30   | 0,87      | 0,87   | 0,87     | 30      |
| 6               | Лошадиные (лошадь, единорог)         | 270   | 30   | 0,86      | 0,87   | 0,87     | 30      |
| 7               | Рептилии (змея, дракон)              | 270   | 30   | 0,88      | 0,87   | 0,88     | 30      |
| 8               | Грызуны (мышь, заяц)                 | 270   | 30   | 0,87      | 0,83   | 0,85     | 30      |
| 9               | Насекомые (бабочка, пчела)           | 270   | 30   | 0,85      | 0,83   | 0,84     | 30      |
| 10              | Экзотические (обезьяна, слон, жираф) | 270   | 30   | 0,90      | 0,87   | 0,89     | 30      |
| Среднее (масго) |                                      | 2700  | 300  | 0,88      | 0,87   | 0,88     | 300     |
| Ассурасу        |                                      |       |      |           |        | 89%      |         |

Модель демонстрирует высокое качество распознавания ( $F1\text{-score} = 0,88$ , ассурасу = 89%) [7, 217].

Лучше всего распознаются категории «Кошачьи» ( $F1 = 0,92$ ) и «Псовые» ( $F1 = 0,91$ ). Эти классы имеют ярко выраженные визуальные признаки и часто встречаются в данных, поэтому модель уверенно их различает [76].

Более низкие значения наблюдаются для категорий «Насекомые» ( $F1 = 0,84$ ) и «Грызуны» ( $F1 = 0,85$ ), где изображения менее однородны и сильнее отличаются по стилю (фото, рисунки, символы) [78].

В целом модель хорошо справляется с задачей, однако качество снижается для редких и визуально неоднородных категорий.

### В.3.3 Распознавание эмоций

Тестовая выборка: 300 личных фото (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.5.5).

Качество модели ниже по сравнению с другими задачами визуального анализа (accuracy = 61,0%, F1-score = 0,47), что связано с высокой вариативностью выражений лиц и ограниченным размером выборки [222, 78].

В данных преобладают позитивные эмоции: радость (45%) и нейтральное выражение (32%), что отражает типичную модель самопрезентации пользователей в социальных сетях [78].

Таблица В.9 – Метрики распознавания эмоций на ВК-аватарах ( $N = 3\,000$ , личные фото)

| №               | Эмоция      | Train | Test | Precision | Recall | F1-score | Support |
|-----------------|-------------|-------|------|-----------|--------|----------|---------|
| 1               | Радость     | 1200  | 150  | 0,72      | 0,68   | 0,70     | 150     |
| 2               | Нейтральное | 900   | 112  | 0,65      | 0,61   | 0,63     | 112     |
| 3               | Грусть      | 240   | 30   | 0,54      | 0,49   | 0,51     | 30      |
| 4               | Удивление   | 180   | 22   | 0,48      | 0,42   | 0,45     | 22      |
| 5               | Гнев        | 120   | 15   | 0,42      | 0,38   | 0,40     | 15      |
| 6               | Страх       | 90    | 11   | 0,35      | 0,31   | 0,33     | 11      |
| 7               | Отвращение  | 60    | 8    | 0,28      | 0,24   | 0,26     | 8       |
| Среднее (macro) |             | 2790  | 348  | 0,49      | 0,45   | 0,47     | 348     |
| Accuracy        |             |       |      |           |        | 61,0%    |         |

Негативные эмоции распознаются заметно хуже: гнев ( $F1 = 0,40$ ), страх ( $F1 = 0,33$ ), отвращение ( $F1 = 0,26$ ). Это связано с их низкой представленностью в выборке (около 2–4%), а также с визуальной близостью выражений – многие из них имеют схожие признаки (напряжённые брови, сжатые губы), что затрудняет их различение [207, 222].

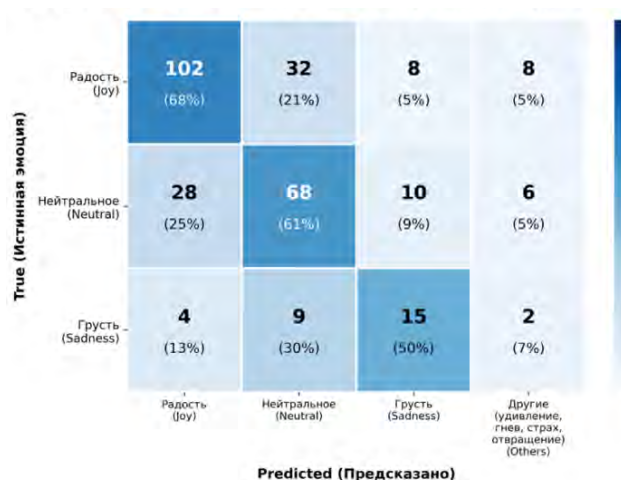


Рисунок В.7 – Матрица ошибок распознавания эмоций

Анализ матрицы ошибок показывает, что наиболее частая путаница возникает между эмоциями «Радость» и «Нейтральное». В большинстве таких случаев речь идёт о слабо выраженной улыбке, которая находится на границе между нейтральным и позитивным выражением лица [207].

#### В.4 Метрики определения демографических характеристик

##### В.4.1 Определение возраста (регрессия, ResNet50)

Тестовая выборка: 1 200 аватаров (10% от общего объёма). Архитектура модели описана в Приложении Б (подпараграф Б.5.1).

Таблица В.10 – Метрики определения возраста по возрастным группам ( $N = 12\,000$ )

| Возрастная группа  | Train size | Test size | MAE (лет) | Median AE (лет) | 95-percentile AE (лет) |
|--------------------|------------|-----------|-----------|-----------------|------------------------|
| 18–25 лет          | 3360       | 420       | 4,2       | 3,5             | 8,9                    |
| 26–35 лет          | 2940       | 368       | 5,1       | 4,2             | 10,3                   |
| 36–45 лет          | 1680       | 210       | 6,4       | 5,8             | 13,1                   |
| 46–52 лет          | 420        | 52        | 8,7       | 7,9             | 16,8                   |
| Среднее взвешенное | 8400       | 1050      | 5,0       | 4,3             | 10,8                   |

Средняя ошибка предсказания (MAE) увеличивается с возрастом: от 4,2 года в группе 18–25 лет до 8,7 года в группе 46+ [7, 220]. Это связано с тем, что старшие возрастные группы представлены в данных значительно хуже (около 5% выборки против ~40% для молодых пользователей), а также с большей вариативностью внешности, которая усиливается с возрастом [78].

Показатель Median AE стабильно ниже MAE во всех возрастных группах, что указывает на асимметрию распределения ошибок: небольшое число крупных отклонений увеличивает среднее значение, тогда как медиана остаётся более устойчивой характеристикой.

##### В.4.2 Определение пола (бинарная классификация, VGG16)

Тестовая выборка: 1 200 аватаров (10% от общего объёма). Архитектура модели описана в Приложении Б (раздел Б.5.2).

Таблица В.11 – Метрики определения пола ( $N=12\,000$ )

| Пол             | Train      | Test      | Precision | Recall | F1-score | Support |
|-----------------|------------|-----------|-----------|--------|----------|---------|
| Мужской         | 4788 (57%) | 599 (57%) | 0,91      | 0,93   | 0,92     | 599     |
| Женский         | 3612 (43%) | 452 (43%) | 0,90      | 0,88   | 0,89     | 452     |
| Среднее (macro) | 8400       | 1051      | 0,91      | 0,91   | 0,91     | 1051    |
| Accuracy        |            |           |           |        | 90,8%    |         |



Как видно из таблицы В.10, модель демонстрирует высокое качество (F1-score = 0,91, accuracy = 90,8%) [226, 217].

|                     |                  |                         |                  |
|---------------------|------------------|-------------------------|------------------|
| True (Истинный пол) | Мужской (Male)   | 557<br>(93%)            | 42<br>(7%)       |
|                     | Женский (Female) | 54<br>(12%)             | 398<br>(88%)     |
|                     |                  | Мужской (Male)          | Женский (Female) |
|                     |                  | Predicted (Предсказано) |                  |

Рисунок В.8 – Матрица ошибок определения пола

Большинство ошибок связано с тремя типичными ситуациями. Наиболее частый случай – андрогинная внешность (51 ошибка, около 53%), когда визуальные признаки пола выражены слабо или смешаны [54]. Также заметную долю составляют изображения низкого качества (28 случаев, ~29%), где модель не может извлечь устойчивые признаки. Ещё одна причина – частичная окклюзия (17 случаев, ~18%), когда лицо закрыто аксессуарами или объектами [78].

#### В.5 Сводная таблица метрик всех моделей

Сводные результаты представлены в таблице В.12. Как видно из таблицы, текстовые модели в среднем демонстрируют более высокое качество по сравнению с визуальными: среднее значение F1-score для задач лингвистического анализа составляет около 0,88, тогда как для визуальных задач (за исключением распознавания эмоций) – около 0,83. Это объясняется тем, что лексические и семантические признаки более устойчивы и легче формализуются, тогда как визуальные данные подвержены большей вариативности (ракурсы, освещение, качество изображений) [7, 78].

Наилучшие результаты достигаются в задачах с чёткими и выраженными признаками классов. В частности, тематическая классификация показывает F1=0,94, классификация зоонимов – F1=0,88, а определение пола – F1=0,91. В этих задачах либо присутствуют устойчивые текстовые паттерны, либо хорошо различимые визуальные признаки.

Из-за несбалансированности данных, а также визуальной близости негативных выражений, наиболее сложной задачей оказалось распознавание эмоций ( $F1 = 0,47$ ) [222, 78].

Таблица В.12 – Сводная таблица метрик всех моделей классификации

| №                      | Модель   | Задача                       | Классов | N<br>(train) | N<br>(test) | Accuracy | F1<br>(macro) | MAE   |
|------------------------|----------|------------------------------|---------|--------------|-------------|----------|---------------|-------|
| Лингвистический анализ |          |                              |         |              |             |          |               |       |
| 1                      | RoBERTa  | Никнеймы<br>(синтетика)      | 6       | 33600        | 4200        | 92,1%    | 0,92          | –     |
| 2                      | RoBERTa  | Никнеймы<br>(реальные ВК)    | 6       | 15753        | 3938        | 88,5%    | 0,88          | –     |
| 3                      | RuBERT   | Статусы                      | 2       | 3200         | 400         | 92,0%    | 0,92          | –     |
| 4                      | RuBERT   | Тематическая<br>классификац. | 10      | 6400         | 800         | 94,0%    | 0,94          | –     |
| 5                      | RuBERT   | Мировоззрен.<br>установки    | 4       | 4000         | 500         | 80,0%    | 0,79          | –     |
| 6                      | RuBERT   | Типы<br>поведения            | 4       | 5504         | 688         | 84,0%    | 0,82          | –     |
| Визуальный анализ      |          |                              |         |              |             |          |               |       |
| 7                      | ResNet50 | Тип аватара                  | 6       | 6804         | 852         | 85,0%    | 0,83          | –     |
| 8                      | ResNet50 | Зоонимы в<br>аватарах        | 10      | 2700         | 300         | 89,0%    | 0,88          | –     |
| 9                      | FER CNN  | Эмоции на<br>лицах           | 7       | 2790         | 348         | 61,0%    | 0,47          | –     |
| 10                     | ResNet50 | Возраст<br>(регрессия)       | –       | 8400         | 1050        | –        | –             | 5 лет |
| 11                     | VGG16    | Пол                          | 2       | 8400         | 1051        | 90,8%    | 0,91          | –     |

## ПРИЛОЖЕНИЕ Г Статистическое обоснование признакового пространства системы анализа цифрового следа

### Г.1 Методология отбора признаков

Настоящее приложение описывает статистические процедуры, использованные для формирования и валидации признакового пространства интеллектуальной системы анализа личности по цифровому следу. Задача заключается в определении признаков, статистически значимо связанных с целевыми психометрическими переменными и обеспечивающих прирост качества предсказания, для их последующего включения в граф знаний и использования в качестве входных данных для GAT-модели.

Отбор признаков осуществлялся с использованием корреляционного анализа, оценки статистической значимости и анализа вклада признаков в предсказательную способность модели.

#### Г.1.1 Датасет

Выборка:  $N = 1\,722$  пользователей ВКонтакте (возраст  $M = 27,3$ ,  $SD = 8,4$ , диапазон 18–52; мужчины 57,6%, женщины 42,4%; высшее образование 68,2%). Выборка сформирована таким образом, чтобы отражать разнообразие пользовательского поведения и цифровых паттернов в социальной сети.

Целевые переменные: 26 психологических факторов (16 первичных 16PF, 4 вторичных 16PF и 6 дополнительных главных факторов), представленных в формате стеновых баллов (диапазон 1–10).

Входные признаки: категориальные переменные, извлечённые модулями системы (Глава 3): 6 категорий никнеймов, 10 тематических категорий статусов, 4 поведенческих типа, 6 типов аватаров, 10 категорий зоонимов, 7 эмоций. Все категориальные признаки были преобразованы в числовое представление с использованием one-hot кодирования.

#### Г.1.2 Разметка и контроль качества данных

Категориальные переменные размечались двумя независимыми экспертами. Межэкспертная надёжность составила Cohen's  $\kappa = 0,78–0,89$ , что соответствует уровню существенного и почти полного согласия согласно шкале Landis & Koch. Расхождения (8–15%) разрешались консенсусом с участием третьего эксперта.

Контроль отсутствия утечки данных обеспечивался следующими мерами:

- модели классификации обучались без доступа к психометрическим данным;

- категории для корреляционного анализа размечались экспертами, а не автоматически;
- кросс-валидация выполнялась с разделением по пользователям (а не по отдельным статусам).

Таким образом, исключено использование информации о целевых переменных при обучении и валидации моделей.

Проверка статистических допущений показала, что для большинства из 26 факторов не выявлено значимых отклонений от нормального распределения (Shapiro-Wilk test,  $p > 0,05$ ), а дисперсии сопоставимы между группами (Levene's test,  $p > 0,05$ ).

Обнаружено 12 потенциальных выбросов (0,7 % выборки,  $|Z| > 3$ ). Ручная проверка показала, что данные наблюдения соответствуют реальным объектам исследования, поэтому они были сохранены в выборке.

Все размеченные данные использовались исключительно для корреляционного анализа и не пересекались с обучающей выборкой моделей классификации (стратифицированное разбиение описано в подпараграфе 3.2.1)

### Г.1.3 Процедура отбора признаков

Отбор информативных признаков проводился в три этапа.

Этап 1 – Корреляционный фильтр. Вычислялись коэффициенты корреляции Пирсона между каждым признаком  $x$  и каждым целевым психометрическим фактором  $y$ :

$$r_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

где  $\bar{x}$ ,  $\bar{y}$  – средние значения переменных.

Для учёта множественных сравнений применялась коррекция Бонферрони:

$$p_{adjusted} = \frac{\alpha}{m},$$

где  $m$  – число проверяемых гипотез.

Признаки, не имеющие ни одной статистически значимой корреляции, исключались из дальнейшего анализа.

Этап 2 – Регрессионная валидация. Для оценки независимой предиктивной значимости признаков использовалась множественная линейная регрессия:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon,$$

где дополнительно учитывались демографические переменные (возраст, пол).

Проводилась проверка мультиколлинеарности (Variance Inflation Factor,  $VIF < 5$ ).

Этап 3 – Оценка стабильности. Проводилась 5-фолдовая кросс-валидация с оценкой воспроизводимости коэффициентов корреляции. Стандартное отклонение значений корреляции между фолдами использовалось как мера стабильности. Признак считался устойчивым при выполнении условия:

$$SD(r) < 0,025.$$

Признаки, прошедшие все три этапа отбора, использовались для формирования рёбер графа знаний типа ASSOCIATED\_WITH, где веса рёбер соответствуют значениям коэффициентов корреляции Пирсона.

## Г.2 Текстовая модальность – никнеймы

Корреляция Пирсона рассчитывалась между 6 категориями никнеймов (бинарные переменные) и 16 первичными факторами 16PF. Общее число статистических тестов составило 96, критический уровень значимости после коррекции Бонферрони –  $p < 0,0005$ . Дополнительно применялась коррекция FDR по методу Бенджамини–Хохберга, которая подтвердила устойчивость выявленных зависимостей.

В результате было обнаружено 16 статистически значимых корреляций ( $p < 0,0005$ ), соответствующих эффектам средней и высокой силы, с диапазоном коэффициентов  $|r| = 0,24–0,44$ . Основные результаты приведены в таблице Г.1.

Наибольшие значения коэффициентов корреляции наблюдаются для следующих категорий:

- зоонимы – связь с фактором I (чувствительность),  $r = 0,44$ ;
- знаменитости – связь с фактором H (активность),  $r = 0,42$ ;
- литературные персонажи – связь с фактором M (мечтательность),  $r = 0,38$ .

Данные признаки формируют наиболее значимые связи в графе знаний и учитываются при последующей агрегации в GAT-модели через механизм внимания.

Для всех представленных корреляций были рассчитаны 95%-ные доверительные интервалы, отражающие статистическую устойчивость оценок. Интервалы не пересекают нулевое значение, что дополнительно подтверждает значимость выявленных зависимостей.

Таблица Г.1 – Корреляции категорий никнеймов с целевыми переменными  
( $N = 1\,722$ )

| Категория никнейма        | Психологический фактор                | Pearson $r$ | p-value | 95% CI         |
|---------------------------|---------------------------------------|-------------|---------|----------------|
| Литературный персонаж     | Фактор М (Мечтательность)             | +0,38       | <0,0001 | [0,33; 0,43]   |
| Литературный персонаж     | Фактор В (Интеллект)                  | +0,29       | 0,0002  | [0,24; 0,34]   |
| Литературный персонаж     | Фактор Q1 (Радикализм)                | +0,24       | 0,0003  | [0,19; 0,29]   |
| Знаменитость              | Фактор Н (Активность)                 | +0,42       | <0,0001 | [0,37; 0,47]   |
| Знаменитость              | Фактор Е (Доминантность)              | +0,34       | <0,0001 | [0,29; 0,39]   |
| Знаменитость              | Фактор F (Беспечность)                | +0,28       | 0,0002  | [0,23; 0,33]   |
| Зооним                    | Фактор I (Чувствительность)           | +0,44       | <0,0001 | [0,39; 0,49]   |
| Зооним                    | Фактор А (Общительность)              | +0,31       | 0,0001  | [0,26; 0,36]   |
| Зооним                    | Фактор С (Эмоциональная стабильность) | -0,26       | 0,0003  | [-0,31; -0,21] |
| Метафора                  | Фактор Q1 (Радикализм)                | +0,36       | <0,0001 | [0,31; 0,41]   |
| Метафора                  | Фактор М (Мечтательность)             | +0,27       | 0,0002  | [0,22; 0,32]   |
| Неименованное обозначение | Фактор L (Честолюбие)                 | +0,39       | <0,0001 | [0,34; 0,44]   |
| Неименованное обозначение | Фактор О (Неуверенность в себе)       | +0,32       | <0,0001 | [0,27; 0,37]   |
| Неименованное обозначение | Фактор Q2 (Самостоятельность)         | +0,29       | 0,0001  | [0,24; 0,34]   |
| Пайзоним                  | Фактор А (Общительность)              | +0,35       | <0,0001 | [0,30; 0,40]   |
| Пайзоним                  | Фактор Н (Активность)                 | +0,26       | 0,0003  | [0,21; 0,31]   |

Полученные результаты показывают, что категории никнеймов обладают устойчивыми и интерпретируемыми связями с психометрическими характеристиками. По сравнению с более вариативными признаками (например, текстами статусов или визуальными характеристиками), никнеймы демонстрируют более стабильные корреляционные зависимости, что делает их надёжным источником признаков для дальнейшего моделирования.

### Г.2.1 Регрессионная валидация независимости признаков

Множественная линейная регрессия с включением демографических переменных (возраст, пол) в качестве контрольных показала, что категории никнеймов сохраняют статистически значимую предиктивную связь с психологическими факторами. Выявленные зависимости не сводятся к влиянию демографических характеристик. Результаты представлены в таблице Г.2.

Таблица Г.2 – Регрессионная валидация: независимость признаков от демографических переменных ( $N = 1\,722$ )

| Зависимая переменная        | Предиктор (категория)     | $\beta$ (станд.) | t-value | p-value | VIF  | Adjusted $R^2$ |
|-----------------------------|---------------------------|------------------|---------|---------|------|----------------|
| Фактор М (Мечтательность)   | Литературный персонаж     | 0,36             | 8,67    | <0,0001 | 1,04 | 0,31           |
| Фактор Н (Активность)       | Знаменитость              | 0,41             | 8,92    | <0,0001 | 1,06 | 0,39           |
| Фактор I (Чувствительность) | Зооним                    | 0,42             | 10,27   | <0,0001 | 1,05 | 0,39           |
| Фактор Q1 (Радикализм)      | Метафора                  | 0,34             | 8,40    | <0,0001 | 1,07 | 0,35           |
| Фактор L (Честолюбие)       | Неименованное обозначение | 0,37             | 10,21   | <0,0001 | 1,03 | 0,35           |

Анализ показал, что для всех рассмотренных моделей значения фактора инфляции дисперсии (VIF) не превышают 1,2, что свидетельствует об отсутствии существенной мультиколлинеарности между предикторами. Полученные результаты подтверждают возможность совместного использования категорий никнеймов в рамках единой модели.

Стандартизованные коэффициенты регрессии для категорий никнеймов находятся в диапазоне  $\beta = 0,34 - 0,42$  и превышают вклад контролируемых демографических переменных ( $|\beta| \leq 0,20$ ). Таким образом, текстовые признаки показывают более выраженную связь с целевыми психометрическими характеристиками по сравнению с возрастом и полом.

Коэффициент детерминации с поправкой (Adjusted  $R^2$ ) варьируется в диапазоне 0,31–0,39 и отражает долю объяснённой дисперсии для соответствующих факторов. Полученные значения согласуются с типичными результатами для задач предсказания психологических характеристик по цифровым следам и указывают на умеренную объясняющую способность одномодальных текстовых признаков.

В совокупности результаты регрессионного анализа подтверждают, что категории никнеймов являются статистически значимыми и независимыми предикторами, пригодными для включения в мультимодальную модель.

### Г.3 Текстовая модальность – тематика статусов

Корреляция Пирсона рассчитывалась между 10 тематическими категориями статусов и 26 психологическими факторами. Общее число статистических тестов составило 260; после коррекции Бонферрони критический уровень значимости установлен на уровне  $p < 0,00025$ .

В результате было выявлено 9 статистически значимых корреляций, представленных в таблице Г.3.

Таблица Г.3 – Корреляции тематических категорий статусов с целевыми переменными ( $N = 1\,722$ )

| Тематическая категория | Психологический фактор      | Pearson $r$ | p-value |
|------------------------|-----------------------------|-------------|---------|
| Любовь и отношения     | Фактор I (Чувствительность) | +0,38       | <0,0001 |
| Юмор и ирония          | Фактор F (Беспечность)      | +0,36       | <0,0001 |
| Жизнь и философия      | Фактор M (Мечтательность)   | +0,34       | <0,0001 |
| Любовь и отношения     | Фактор A (Общительность)    | +0,31       | 0,0001  |
| Жизнь и философия      | Фактор Q1 (Радикализм)      | +0,29       | 0,0001  |
| Музыка и искусство     | Фактор I (Чувствительность) | +0,29       | 0,0001  |

Наиболее выраженные связи наблюдаются для следующих тематических категорий:

- «Любовь и отношения» – с фактором I (чувствительность),  $r = 0,38$ , и фактором A (общительность),  $r = 0,31$ ;
- «Юмор и ирония» – с фактором F (беспечность),  $r = 0,36$ ;
- «Жизнь и философия» – с фактором M (мечтательность),  $r = 0,34$ , и фактором Q1 (радикализм),  $r = 0,29$ ;
- «Музыка и искусство» – с фактором I (чувствительность),  $r = 0,29$ .

В связи с высокой вариативностью и контекстной зависимостью текстов статусов, отражающих ситуативные интересы и текущие состояния пользователя, наблюдаемые корреляции характеризуются умеренной силой ( $|r| = 0,22–0,38$ ) и уступают по величине соответствующим зависимостям для категорий никнеймов.

Несмотря на это, выявленные статистически значимые связи дополняют признаки, извлечённые из никнеймов, и расширяют представление о пользователе за счёт учета его актуальных тематических предпочтений. В рамках построения графа знаний такие зависимости формируют дополнительные рёбра, повышая связность графовой структуры и обеспечивая более полное описание входных данных для GAT-модели.



#### Г.4. Поведенческие типы

Корреляция Пирсона рассчитывалась между 4 типами поведенческих паттернов и 26 психологическими факторами. Общее число статистических тестов составило 104; после коррекции Бонферрони критический уровень значимости установлен на уровне  $p < 0,00025$ .

В результате было выявлено 10 статистически значимых корреляций, представленных в таблице Г.4.

Таблица Г.4 – Корреляции поведенческих типов с целевыми переменными ( $N = 1722$ )

| Поведенческий тип | Психологический фактор               | Pearson $r$ | p-value |
|-------------------|--------------------------------------|-------------|---------|
| Позитивное        | Фактор G (Ответственность)           | +0,42       | <0,0001 |
| Позитивное        | Фактор A (Общительность)             | +0,35       | <0,0001 |
| Девиантное        | Фактор E (Доминантность)             | +0,39       | <0,0001 |
| Девиантное        | Фактор L (Честолюбие)                | +0,33       | <0,0001 |
| Делинквентное     | Фактор Q3 (Психол. проницательность) | -0,44       | <0,0001 |
| Делинквентное     | Фактор F (Беспечность)               | +0,37       | <0,0001 |
| Делинквентное     | Фактор G (Ответственность)           | -0,34       | <0,0001 |
| Нейтральное       | Фактор N (Дипломатичность)           | +0,26       | 0,0002  |

Наибольшие значения коэффициентов корреляции наблюдаются для следующих связей:

- делинквентное поведение – отрицательная связь с фактором Q3 (психологическая проницательность),  $r = -0,44$ , а также отрицательная связь с фактором G (ответственность),  $r = -0,34$ , и положительная связь с фактором F (беспечность),  $r = 0,37$ ;
- девиантное поведение – положительная связь с фактором E (доминантность),  $r = 0,39$ , и фактором L (честолюбие),  $r = 0,33$ ;
- позитивное (просоциальное) поведение – положительная связь с фактором G (ответственность),  $r = 0,42$ , и фактором A (общительность),  $r = 0,35$ ;
- нейтральное поведение – умеренная положительная связь с фактором N (дипломатичность),  $r = 0,26$ .

Полученные зависимости характеризуются умеренной и высокой силой ( $|r| = 0,26-0,44$ ) и являются сопоставимыми по величине с результатами для категорий никнеймов. Следовательно, поведенческие признаки играют значимую роль в предсказании отдельных психологических характеристик.

Дополнительная проверка с использованием логистической регрессии показала, что соответствующие модели достигают значений  $AUC = 0,80-0,82$  при

коэффициенте псевдодетерминации  $R^2$  в диапазоне 0,39–0,42. Наблюдается хорошая разделяющая способность признаков.

В совокупности результаты подтверждают, что поведенческие типы являются информативными предикторами и могут рассматриваться как самостоятельный источник признаков. В рамках построения графа знаний выявленные зависимости формируют дополнительные рёбра, отражающие поведенческие аспекты цифрового следа пользователя и дополняющие текстовые и визуальные характеристики.

#### Г.5. Визуальная модальность

Визуальные признаки (зоонимы в аватарах, эмоции на изображениях) извлекались с использованием моделей ResNet50 и FER (см. Приложение Д) при наличии распознаваемого визуального контента. Доля аватаров, пригодных для анализа, составила 72,2%; для задачи распознавания эмоций использовалась подвыборка  $N = 1\,244$  изображениям с детектированными лицами.

Корреляционный анализ проводился с использованием коэффициента Пирсона между визуальными признаками и психологическими факторами. Результаты приведены в таблице Г.5.

Таблица Г.5 – Корреляции зоонимов в аватарах с целевыми переменными

| Категория зоонима     | Психологический фактор      | Pearson $r$ | p-value |
|-----------------------|-----------------------------|-------------|---------|
| Псовые (волк, собака) | Фактор Е (Доминантность)    | +0,42       | <0,0001 |
| Кошачьи (кошка, лев)  | Фактор I (Чувствительность) | +0,36       | <0,0001 |
| Птицы (орёл, сова)    | Фактор Q1 (Радикализм)      | +0,31       | 0,0001  |

Наиболее выраженные связи наблюдаются для следующих категорий:

- псовые (волк, собака) – с фактором Е (доминантность),  $r = 0,42$ ;
- кошачьи (кошка, лев) – с фактором I (чувствительность),  $r = 0,36$ ;
- птицы (орёл, сова) – с фактором Q1 (радикализм),  $r = 0,31$ .

Результаты приведены в таблице Г.6.

Таблица Г.6 – Корреляции эмоций на аватарах с целевыми переменными

| Эмоция           | Психологический фактор          | Pearson $r$ | p-value |
|------------------|---------------------------------|-------------|---------|
| Грусть           | Фактор О (Неуверенность в себе) | +0,34       | <0,0001 |
| Радость (улыбка) | Фактор F (Беспечность)          | +0,28       | <0,0001 |

Выявлены следующие статистически значимые зависимости:

- грусть – положительная связь с фактором О (неуверенность в себе),  $r = 0,34$ ;
- радость (улыбка) – положительная связь с фактором F (беспечность),  $r = 0,28$ .

В целом для визуальной модальности выявлено 5 статистически значимых корреляций. Это существенно меньше по сравнению с текстовыми признаками и может быть связано с ограниченной доступностью интерпретируемого визуального контента, а также с большей вариативностью изображений.

Вместе с тем визуальные признаки дополняют текстовые. В частности, наблюдается согласованность между зоонимами в никнеймах и аватарах, которые демонстрируют связь с одними и теми же психологическими факторами (например, фактор I – чувствительность). Такие совпадения формируют устойчивые кросс-модальные связи в графе знаний, усиливая достоверность соответствующих предикторов.

Таким образом, несмотря на меньшее количество выявленных зависимостей, визуальная модальность вносит дополнительную информацию и играет вспомогательную роль в мультимодальной архитектуре Hybrid Fusion.

#### Г.6 Сводная таблица признакового пространства

В таблице Г.7 обобщены результаты корреляционного анализа по всем рассмотренным модальностям и представлен итоговый состав признакового пространства системы.

Таблица Г.7 – Сводная статистика признакового пространства по модальностям

| Модальность          | Значимых корреляций | Диапазон $ r $ | Среднее $ r $ | Наивысшая корреляция                         |
|----------------------|---------------------|----------------|---------------|----------------------------------------------|
| Текст: Никнеймы      | 16                  | 0,24 – 0,44    | 0,33          | Зооним $\rightarrow$ Фактор I ( $r = 0,44$ ) |
| Текст: Темы статусов | 9                   | 0,24 – 0,38    | 0,30          | Любовь $\rightarrow$ Фактор I ( $r = 0,38$ ) |
| Текст: Поведение     | 10                  | 0,26 – 0,44    | 0,35          | Делинквентное и Фактор Q3 ( $r = -0,44$ )    |
| Визуал: Зоонимы      | 3                   | 0,31 – 0,42    | 0,36          | Псовы и Фактор E ( $r = 0,42$ )              |
| Визуал: Эмоции       | 2                   | 0,28 – 0,34    | 0,31          | Грусть и Фактор O ( $r = 0,34$ )             |
| ИТОГО                | 40                  | 0,24 – 0,44    | 0,33          | –                                            |

В сумме выявлено 40 статистически значимых корреляционных связей, формирующих рёбра графа знаний. Из них 35 относятся к текстовой модальности и 5 – к визуальной.

Анализ распределения коэффициентов корреляции показывает, что:

- для текстовых признаков диапазон значений составляет  $|r| = 0,24\text{--}0,44$ , среднее значение  $|r| \approx 0,33$ ;
- для визуальных признаков диапазон составляет  $|r| = 0,28\text{--}0,42$ , при среднем значении  $|r| \approx 0,31\text{--}0,36$ .

Наиболее сильные связи наблюдаются как в текстовой, так и в визуальной модальности. В частности, максимальные значения коэффициента корреляции достигаются для:

- категории «зооним» (никнеймы) и фактора I (чувствительность),  $r = 0,44$ ;
- категории «делинквентное поведение» и фактора Q3 (психологическая проницательность),  $r = -0,44$ ;
- категории «псовые» (аватары) и фактора E (доминантность),  $r = 0,42$ .

Дополнительная проверка устойчивости показала, что все выявленные корреляции воспроизводятся при 5-fold кросс-валидации, при стандартном отклонении коэффициентов  $SD < 0,022$ . Это свидетельствует о стабильности признакового пространства и отсутствии выраженного переобучения.

В целом текстовая модальность выступает основным источником предиктивной информации за счёт большего числа выявленных связей и их устойчивости. Визуальная модальность, несмотря на меньшее количество значимых корреляций, предоставляет дополнительную информацию и дополняет текстовые признаки.

Полученные результаты подтверждают целесообразность использования мультимодальной архитектуры с адаптивным объединением признаков. В модели Adaptive Fusion вклад отдельных модальностей определяется обучаемыми коэффициентами и изменяется в зависимости от их информативности.

#### Г.7 Формирование графа знаний

Результаты статистического анализа используются для построения структуры графа знаний (Глава 4), отражающего связи между признаками цифрового следа и психометрическими характеристиками.

В графе выделяются два основных типа узлов:

- узлы типа Feature соответствуют категориям извлечённых признаков (6 категорий никнеймов, 10 тематических категорий статусов, 4 поведенческих типа, 6 типов аватаров, 10 категорий зоонимов, 7 эмоциональных состояний), всего 43 узла;
- узлы типа PsychConstruct представляют 26 целевых психологических факторов.

Связи между узлами задаются рёбрами типа ASSOCIATED\_WITH, соответствующими статистически значимым корреляциям. Всего в граф включено 40 рёбер, полученных по результатам корреляционного анализа.

Вес ребра определяется как модуль коэффициента корреляции Пирсона:

$$w = |r|.$$

Знак корреляции сохраняется в виде отдельного атрибута direction  $\in \{\text{positive, negative}\}$ , что позволяет учитывать направление связи при последующей интерпретации.

Включение ребра в граф осуществляется при выполнении условия статистической значимости:  $p < 0,0005$ , после коррекции Бонферрони. Использование данного порога обеспечивает консервативный отбор связей и снижает вероятность включения ложноположительных зависимостей.

Сформированная структура графа знаний включает только связи, подтверждённые статистическим анализом. Это снижает вероятность распространения нерелевантной информации при обучении графовой нейронной сети и формировании коэффициентов внимания.

#### Г.8 Ограничения признакового пространства

Существует ряд ограничений, связанных с используемым подходом к формированию признакового пространства.

Во-первых, применяемый корреляционный дизайн не позволяет интерпретировать выявленные зависимости как причинно-следственные. Полученные связи отражают исключительно статистические ассоциации между признаками цифрового следа и психологическими характеристиками. Однако в контексте рассматриваемой задачи это не является принципиальным ограничением, поскольку целью является предсказание значений факторов, а не установление каузальных механизмов.

Во-вторых, контроль потенциальных конфаундеров был ограничен демографическими переменными: возрастом и полом (параграф Г.2). Другие возможные факторы, такие как уровень образования, профессиональная принадлежность или интенсивность использования социальных сетей, в модели не учитывались. Это важно принимать во внимание при интерпретации результатов и может служить направлением для дальнейших исследований.

В-третьих, выборка характеризуется определённым смещением: доля пользователей с высшим образованием составляет 68,2%, что превышает соответствующие показатели в генеральной совокупности. Данное обстоятельство

может оказывать влияние на распределение когнитивных и связанных с ними факторов. В связи с этим представляется целесообразным проведение дополнительной валидации и периодической рекалибровки модели на более репрезентативных выборках.

Наконец, ограничения связаны и с покрытием визуальной модальности. Около 27,8% аватаров не содержат распознаваемых лиц или объектов. Это снижает возможность извлечения признаков, связанных с эмоциями и зоонимами. В таких случаях система опирается преимущественно на текстовую модальность, что реализуется в рамках механизма адаптивного объединения признаков.

В совокупности указанные ограничения не обесценивают полученные результаты, однако задают границы применимости модели и определяют направления её дальнейшего развития.

## ПРИЛОЖЕНИЕ Д Технические детали визуальных и мультимодальных моделей

Задачи Д.1–Д.4 (определение возраста, пола, типа аватара, распознавание эмоций) решались на вспомогательных датасетах большого объёма ( $N = 12\,000$  аватаров с верифицированным возрастом и полом,  $N = 8\,500$  для типов аватаров,  $N = 3\,000$  для эмоций), предназначенных для обучения специализированных моделей-энкодеров.

Мультимодальная модель Д.5 (Hybrid Fusion) обучалась на основном датасете исследования ( $N = 1\,722$  пользователя с полным психометрическим профилем) с использованием предобученных энкодеров (Д.1–Д.4) для извлечения признаков.

### Д.1 Определение возраста по аватару

#### Д.1.1 Постановка задачи

Задача определения возраста по изображению аватара формулируется как задача регрессии – предсказания непрерывной переменной «возраст в годах» [224].

Пусть задана обучающая выборка [224]:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N,$$

где  $x_i \in \mathbb{R}^{H \times W \times 3}$  – изображение аватара пользователя (RGB),  $y_i \in \mathbb{R}$  – истинный возраст пользователя.

Требуется построить отображение:

$$f_{\theta}: \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R},$$

параметризованное вектором параметров  $\theta$ , такое, что:

$$\hat{y}_i = f_{\theta}(x_i),$$

где  $\hat{y}_i$  – предсказанный возраст пользователя.

Обучение модели осуществляется путём минимизации функции потерь, определяемой как средняя абсолютная ошибка (Mean Absolute Error, MAE) [224]:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|,$$

где:

- $N$  – размер обучающей выборки;
- $y_i$  – истинное значение возраста;
- $\hat{y}_i$  – предсказанное значение.

Подробное описание датасета и процедуры сбора данных приведено в параграфе 3.2.1.

Входными данными модели являются изображения аватаров пользователей переменного размера, приводимые к фиксированному формату на этапе предобработки. Выходом модели является вещественное значение возраста.

Диапазон предсказываемых значений (18–52 года) определяется характеристиками используемой выборки и не является ограничением архитектуры модели.

#### Д.1.2 Конвейер обработки данных

Процесс подготовки изображений для задачи определения возраста включает два основных этапа: детекцию лица и предобработку.

##### Этап 1. Детекция лица

Для локализации лиц на изображении используется модель YOLOv5x (You Only Look Once, версия 5, extra-large) [230], предназначенная для решения задач детекции объектов.

Основные характеристики модели [230]:

- количество параметров – 86,7 млн;
- backbone – CSPDarknet53;
- neck – PANet;
- head – детекционная голова YOLOv5 (3 масштаба);
- входное разрешение – 640×640 пикселей;
- предобучение – датасет COCO (80 классов объектов) [231].

Качество детекции лиц на выборке аватаров ВКонтакте составило:

- Precision – 95%,
- Recall – 92% (оценка на ручной выборке из 300 изображений).

Процедура детекции включает следующие шаги:

1. Входное изображение масштабируется до размера 640×640 пикселей с сохранением пропорций.
2. Модель YOLOv5x выполняет детекцию лиц и возвращает набор ограничивающих прямоугольников.
3. В случае обнаружения нескольких лиц (групповое изображение) выбирается прямоугольник максимальной площади:

$$(x^*, y^*) = \arg \max_{(x_{min}, y_{min}, x_{max}, y_{max})} [(x_{max} - x_{min})(y_{max} - y_{min})],$$

что соответствует предположению, что владелец профиля занимает центральную позицию.



4. При отсутствии обнаруженного лица изображение исключается из анализа возраста (27,8% выборки).

5. Прямоугольник задаётся координатами:  $(x_{min}, y_{min}, x_{max}, y_{max})$  – левый верхний и правый нижний углы прямоугольника.

Этап 2. Обрезка и предобработка

1. После детекции выполняется обрезка и нормализация изображения.

2. Изображение обрезается по прямоугольнику с расширением на 20% в каждую сторону для сохранения контекста (причёска, плечи):

$$x_{min} = \max(0, x_{min} - 0,2 \cdot w_{box}),$$

$$x_{max} = \min(W_{img}, x_{max} + 0,2 \cdot w_{box}),$$

где  $w_{box} = x_{max} - x_{min}$ .

3. Аналогичные преобразования выполняются по оси  $y$ .

4. Полученное изображение приводится к фиксированному размеру  $224 \times 224$ , что соответствует стандартному входу моделей, предобученных на ImageNet [221].

5. Пиксельные значения нормализуются по каналам RGB [221]:

$$x_{norm} = \frac{x - \mu}{\sigma},$$

где  $\mu = [0,485, 0,456, 0,406]$ ,  $\sigma = [0,229, 0,224, 0,225]$  – средние значения и стандартные отклонения, вычисленные на ImageNet [221].

### Д.1.3 Сравнение архитектур

Для задачи регрессии возраста проводилось сравнение 4 популярных архитектур свёрточных нейронных сетей (CNN), предобученных на датасете ImageNet [221].

Таблица Д.1 – Сравнение архитектур для определения возраста ( $N = 1\,244$ , только аватары с лицами)

| Архитектура          | Параметры | Глубина    | Особенности                                    | MAE (тест) | Время обучения |
|----------------------|-----------|------------|------------------------------------------------|------------|----------------|
| VGG16 [226]          | 138 млн   | 16 слоёв   | Последовательные свёртки $3 \times 3$          | 6,8 лет    | 3 часа         |
| ResNet18 [7]         | 11 млн    | 18 слоёв   | Residual connections (остаточные связи)        | 5,8 лет    | 2 часа         |
| ResNet50 [7]         | 25 млн    | 50 слоёв   | Глубокая архитектура с bottleneck блоками      | 5,0 лет    | 4 часа         |
| EfficientNet-B0 [21] | 5 млн     | Переменная | Compound scaling (ширина, глубина, разрешение) | 5,4 лет    | 2,5 часа       |

По результатам экспериментов наилучшее качество показала архитектура ResNet50 [7], достигнув значения MAE = 5,0 лет, что на 1,8 года лучше по сравнению с VGG16 [226]. Выбор ResNet50 для итоговой модели обусловлен оптимальным соотношением точности (MAE = 5,0) и вычислительной эффективности (25 млн параметров, 4 часа обучения), что согласуется с рекомендациями для задач регрессии возраста на ограниченных выборках [220].

Преимущество архитектур семейства ResNet характеризуется использованием остаточных связей, позволяющих эффективно обучать глубокие сети и снижать влияние проблемы исчезающих градиентов [7].

Работа остаточного блока описывается следующим образом [7]:

$$y = \mathcal{F}(x, \{W_i\}) + x,$$

где:

- $x$  – вход блока;
- $\mathcal{F}(x, \{W_i\})$  – остаточная функция, реализованная через несколько свёрточных слоёв;
- $y$  – выход блока (сумма остаточной функции и входа).

Таким образом, сеть обучает не прямое отображение  $\mathcal{H}(x)$ , а остаточное [7]:

$$\mathcal{F}(x) = \mathcal{H}(x) - x.$$

Это упрощает процесс оптимизации и ускоряет сходимость при обучении глубоких моделей [7].

#### Д.1.4 Детальная архитектура ResNet50 для регрессии возраста

Таблица Д.2 – Архитектура ResNet50 для определения возраста

| Слой                             | Выходной размер | Операции                         |
|----------------------------------|-----------------|----------------------------------|
| Input                            | 224×224×3       | Входное изображение (RGB)        |
| Conv1, BatchNorm, MaxPool        | 56×56×64        | Conv 7×7, stride=2 – MaxPool 3×3 |
| Conv2_x                          | 56×56×256       | 3 residual blocks (bottleneck)   |
| Conv3_x                          | 28×28×512       | 4 residual blocks (bottleneck)   |
| Conv4_x                          | 14×14×1024      | 6 residual blocks (bottleneck)   |
| Conv5_x                          | 7×7×2048        | 3 residual blocks (bottleneck)   |
| AvgPool                          | 1×1×2048        | Global Average Pooling           |
| Flatten                          | 2048            | Преобразование в вектор          |
| Регрессионная голова модели:     |                 |                                  |
| Dropout, Linear, Dropout, Linear | 1               | FC(2048 – 512) – FC(512 – 1)     |

Итог: выход модели – скалярное значение возраста; общее число параметров составляет порядка 25 млн [7].

Bottleneck block – специальный тип residual block в ResNet50, использующий 3 свёртки ( $1 \times 1 - 3 \times 3 - 1 \times 1$ ) для снижения вычислительной сложности [7]:

$$\mathcal{F}(x) = W_3 \sigma(W_2 \sigma(W_1 x)),$$

где  $W_1$  – свёртка  $1 \times 1$  (уменьшение каналов),  $W_2$  – свёртка  $3 \times 3$  (обработка признаков),  $W_3$  – свёртка  $1 \times 1$  (увеличение каналов),  $\sigma$  – ReLU [7].

Global Average Pooling (AvgPool) – слой глобального усреднения преобразует тензор признаков размерности  $H \times W \times C$  в вектор размерности  $C$ , усредняя значения по пространственным координатам [7]:

$$\text{AvgPool}(F) = \frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W F_{ij}.$$

В результате для входа  $7 \times 7 \times 2048$  формируется вектор размерности 2048 [7].

Для предотвращения переобучения используется слой Dropout [232], который во время обучения случайным образом обнуляет долю нейронов с вероятностью  $p = 0,5$  [224].

#### Д.1.5 Обучение модели ResNet50

Обучение модели проводилось на датасете из 12 000 аватаров пользователей ВКонтакте с верифицированным возрастом (анкетные данные и ручная проверка). Подробное описание датасета и процедуры сбора данных приведено в подпараграфе 3.2.1.

Разделение выборки:

- обучающая: 8 400 (70%);
- валидационная: 2 400 (20%);
- тестовая: 1 200 (10%).

Распределение по возрастным группам [78]:

- 18–25 лет: 40% (4 800);
- 26–35 лет: 35% (4 200);
- 36–45 лет: 20% (2 400);
- 46–52 лет: 5% (600).

Таблица Д.3 – Гиперпараметры обучения модели ResNet50 для определения возраста

| Гиперпараметр             | Значение    | Обоснование                                               |
|---------------------------|-------------|-----------------------------------------------------------|
| Оптимизатор               | AdamW [219] | Адаптивная оптимизация с регуляризацией [219]             |
| Скорость обучения         | 1e-4        | Стабильное обучение для задачи регрессии [7]              |
| Размер пакета             | 64          | Баланс между памятью GPU и устойчивостью градиентов [217] |
| Число эпох обучения       | 30          | Ранняя остановка после 25 эпох [40]                       |
| Коэффициент регуляризации | 0,01        | L2-регуляризация для предотвращения переобучения [219]    |
| Число эпох разогрева      | 5           | Плавный разогрев обучения [219]                           |
| Функция потерь            | MSE [224]   | Гладкая оптимизация для задачи регрессии                  |
| Аугментация данных        | Да [7]      | Повышение обобщающей способности                          |

Оптимизация модели осуществляется по функции среднеквадратичной ошибки (MSE) [224]:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2,$$

где  $\hat{y}_i$  – предсказанный возраст,  $y_i$  – истинный возраст.

При этом метрика качества интерпретируется через MAE (в годах) [224].

Для повышения устойчивости модели применялась аугментация изображений [7]:

- случайный поворот:  $\pm 15^\circ$ ;
- горизонтальное отражение ( $p = 0,5$ );
- масштабирование:  $[0,9; 1,1]$ ;
- изменение яркости:  $\pm 10\%$ ;
- изменение контраста:  $\pm 10\%$ .

Аугментация используется только на обучающей выборке.

Таблица Д.4 – Результаты обучения модели ResNet50 по выборкам и возрастным группам

| Метрика                    | Train | Validation | Test |
|----------------------------|-------|------------|------|
| MAE (лет)                  | 3,8   | 4,6        | 5,0  |
| MAE по возрастным группам: |       |            |      |
| 18–25 лет                  | 3,2   | 3,9        | 4,2  |
| 26–35 лет                  | 4,1   | 4,8        | 5,1  |
| 36–45 лет                  | 4,9   | 5,7        | 6,4  |
| 46+ лет                    | 7,2   | 8,1        | 8,7  |

Детальные метрики по возрастным группам приведены в Приложении В (таблица В.10).

Рост ошибки с увеличением возраста наблюдается во всех разбиениях. Это связано с тремя факторами [78, 220]:

1. Старшая возрастная группа существенно менее представлена в данных (около 5% выборки), что ограничивает способность модели к обобщению [78].

2. Внешние признаки в возрасте 46+ характеризуются большей вариативностью, что усложняет задачу регрессии [220].

3. Пользователи старших возрастных групп реже используют личные фотографии в качестве аватаров, что снижает информативность входных данных [78].

## Д.2 Определение пола по аватару

### Д.2.1 Архитектура модели

Для задачи определения пола по изображению используется архитектура VGG16 (Visual Geometry Group, 16 слоёв) [226], широко применяемая в задачах бинарной классификации изображений.

Таблица Д.5 – Характеристики VGG16

| Параметр                                 | Значение                                            |
|------------------------------------------|-----------------------------------------------------|
| Количество параметров                    | 138 млн (оригинальная VGG16) [226]                  |
| Количество параметров (после дообучения) | 14,7М (заморожен backbone, обучается только голова) |
| Глубина                                  | 16 слоёв (13 свёрточных, 3 полносвязных) [226]      |
| Входное разрешение                       | 224×224×3 [226]                                     |
| Предобучение                             | ImageNet (1,2 млн изображений, 1000 классов) [221]  |

Выбор архитектуры VGG16 для распознавания пола обоснован следующим: во-первых, VGG16 является одной из наиболее изученных архитектур для задач классификации изображений и демонстрирует высокую точность на задачах распознавания лиц [226]; во-вторых, глубокая архитектура (16 слоёв) позволяет извлекать признаки различных уровней абстракции, включая мелкие детали, важные для определения пола [226]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision [217], что обеспечивает воспроизводимость.

Архитектура VGG16 представляет собой последовательность свёрточных блоков с ядрами  $3 \times 3$ , объединённых операциями подвыборки (MaxPooling), с последующей классификационной частью [226].

Свёрточная часть включает 5 блоков [226]:

- блок 1: 2 слоя (64 канала);
- блок 2: 2 слоя (128 каналов);
- блок 3: 3 слоя (256 каналов);
- блок 4: 3 слоя (512 каналов);
- блок 5: 3 слоя (512 каналов).

После каждого блока применяется операция MaxPooling, уменьшающая пространственное разрешение [226].

На выходе свёрточной части формируется тензор размерности:

$$7 \times 7 \times 512,$$

который преобразуется в вектор:

$$7 \cdot 7 \cdot 512 = 25088.$$

Оригинальная классификационная часть VGG16 модифицирована для задачи бинарной классификации [226]:

$$\text{FC}(25088 - 4096) - \text{Dropout} - \text{FC}(4096 - 4096) - \text{Dropout} - \text{FC}(4096 - 2).$$

На выходе применяется функция softmax, формирующая вероятности принадлежности к классам [224]:

$$p(y | x) = \text{softmax}(z),$$

где  $z \in \mathbb{R}^2$  – выход последнего линейного слоя.

#### Д.2.2. Обучение модели

Распределение по полу в датасете [78]:

- мужчины: 6840 (57%);
- женщины: 5160 (43%).

Обучение модели осуществлялось с использованием подхода трансферное обучение [226]:

1. Свёрточная часть сети (блоки Conv1-Conv5) фиксирована, веса, предобученные на ImageNet [221], не обновляются. Данные слои извлекают универсальные визуальные признаки (границы, текстуры, формы) [226].

2. Оригинальная полносвязная часть заменена на новую [226]:

$$\text{FC}(25088 - 512) - \text{ReLU} - \text{Dropout}(p = 0,5) - \text{FC}(512 - 2).$$

3. На выходе применяется softmax [224].

Таблица Д.6 – Гиперпараметры обучения модели VGG16 для определения пола

| Гиперпараметр       | Значение                 | Обоснование                                                          |
|---------------------|--------------------------|----------------------------------------------------------------------|
| Оптимизатор         | AdamW [219]              | Адаптивная оптимизация с регуляризацией [219]                        |
| Скорость обучения   | 3e-4                     | Выше, чем для регрессии возраста, т.к. обучается только голова [226] |
| Размер пакета       | 64                       | Баланс между памятью GPU и устойчивостью градиентов [217]            |
| Число эпох обучения | 20                       | Ранняя остановка по валидации [40]                                   |
| Функция потерь      | Cross-Entropy Loss [224] | Стандарт для бинарной классификации                                  |
| Веса классов        | [1,17; 1,0]              | Коррекция дисбаланса 57%/43% [224]                                   |

Для обучения используется бинарная кросс-энтропия [224]:

$$\mathcal{L}_{\text{CE}} = - \sum_{i=1}^N [y_i \log(\hat{y}_i^{\text{male}}) + (1 - y_i) \log(\hat{y}_i^{\text{female}})],$$

где  $y_i \in \{0,1\}$  (0 = женский, 1 = мужской),  $\hat{y}_i^{\text{male}}$  – предсказанная вероятность мужского пола [224].

Высокая точность модели ( $\approx 91\%$  на тестовой выборке) обусловлена наличием устойчивых визуальных признаков [226]:

- длина волос (у женщин чаще длинные волосы);
- наличие макияжа;
- морфология лица (челюсть, брови, пропорции).

Детальные метрики по классам приведены в Приложении В (таблица В.11).

Таблица Д.7 – Результаты обучения модели VGG16 для определения пола

| Метрика             | Train | Validation | Test  |
|---------------------|-------|------------|-------|
| Accuracy            | 94,2% | 91,5%      | 90,8% |
| Precision (мужской) | 93,1% | 91,2%      | 91,0% |
| Recall (мужской)    | 95,8% | 92,4%      | 92,7% |
| Precision (женский) | 94,5% | 91,9%      | 90,5% |
| Recall (женский)    | 91,2% | 90,1%      | 88,2% |
| F1-score (macro)    | 0,93  | 0,92       | 0,91  |

Общая доля ошибок составляет 9,2% и распределяется следующим образом [54, 78]:

- андрогинные лица – 4,8% [54];
- низкое качество изображений – 2,6% [78];
- частичная окклюзия – 1,8% [54].

Ошибки носят системный характер и связаны с ограниченностью визуальной информации, а не с нестабильностью модели [78].

### Д.3. Классификация типа аватара (6 категорий)

#### Д.3.1 Архитектура модели

Для задачи классификации типа аватара используется архитектура ResNet50 [7], аналогичная применяемой в задаче определения возраста (параграф Д.1), но с модифицированной классификационной головой для мультиклассовой классификации [224].

Выбор архитектуры ResNet50 для классификации типа аватара обоснован следующим: во-первых, данная модель показала наилучшие результаты в задаче регрессии возраста (подпараграф Д.1.3); во-вторых, использование остаточных связей позволяет эффективно обучать глубокую сеть и извлекать признаки

различных уровней абстракции [7]; в-третьих, модель имеет открытую реализацию в библиотеке torchvision [217], что обеспечивает воспроизводимость.

Классификация осуществляется по 6 категориям, перечень которых представлен в таблице Д.8. Выбор категорий обоснован результатами анализа визуального контента аватаров пользователей ВКонтакте и согласуется с типологиями, используемыми в исследованиях визуальной самопрезентации в социальных сетях [64\*, 75, 76].

Классификация осуществляется по 6 категориям:

- 1) личное фото;
- 2) фото других людей (знаменитости);
- 3) групповое фото ( $\geq 2$  человека);
- 4) животные;
- 5) символы и объекты;
- 6) пейзажи и природа.

Классификационная голова имеет вид:

FC(2048 – 512) – ReLU – Dropout( $p = 0.5$ ) – FC(512 – 6).

На выходе применяется функция softmax:

$$p(y | x) = \text{softmax}(z), z \in \mathbb{R}^6.$$

### Д.3.2 Обучение модели

Датасет:

- размер: 8 500 аватаров с ручной разметкой [75, 76];
- распределение: сбалансированное (по 1 417 на категорию).

Разметка выполнялась двумя независимыми экспертами по инструкции, содержащей критерии отнесения изображения к каждой из 6 категорий (таблица Д.8). Межэкспертное согласие составило каппа Коэна  $\kappa = 0,84$ , что соответствует почти полному согласию согласно шкале Лэндиса–Коха [223].

Таблица Д.8 – Гиперпараметры обучения модели ResNet50 для классификации типа аватара

| Гиперпараметр       | Значение    | Обоснование                                                   |
|---------------------|-------------|---------------------------------------------------------------|
| Оптимизатор         | AdamW [219] | Адаптивная оптимизация с регуляризацией [219]                 |
| Скорость обучения   | $2e-4$      | Стандартное значение для дообучения ResNet [7]                |
| Размер пакета       | 32          | Компромисс между памятью GPU и стабильностью градиентов [217] |
| Число эпох обучения | 25          | Ранняя остановка по валидации [40]                            |



Продолжение таблицы Д.8

|                    |                                                              |                                           |
|--------------------|--------------------------------------------------------------|-------------------------------------------|
| Функция потерь     | Cross-Entropy Loss (multi-class) [224]                       | Стандарт для многоклассовой классификации |
| Аугментация данных | Да (повороты, flip, масштабирование, цветовые изменения) [7] | Повышение обобщающей способности          |

Применение весов классов не требовалось, так как датасет сбалансирован (по 1 417 примеров на категорию).

Для обучения используется функция потерь мультиклассовой кросс-энтропии [224]:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^6 y_{i,c} \log \hat{y}_{i,c},$$

где  $y_{i,c}$  – истинная метка класса,  $\hat{y}_{i,c}$  – предсказанная вероятность.

Таблица Д.9 – Результаты обучения модели классификации типа аватара

| Метрика          | Train | Validation | Test  |
|------------------|-------|------------|-------|
| Accuracy         | 89,3% | 86,2%      | 85,0% |
| F1-score (macro) | 0,88  | 0,85       | 0,83  |

Детальные метрики по классам приведены в Приложении В (таблица В.7).

Таблица Д.10 – Производительность модели классификации типа аватара по категориям

| Категория         | Precision | Recall | F1-score |
|-------------------|-----------|--------|----------|
| Животные          | 0,92      | 0,90   | 0,91     |
| Пейзажи           | 0,89      | 0,87   | 0,88     |
| Символы и объекты | 0,86      | 0,84   | 0,85     |
| Групповое фото    | 0,84      | 0,82   | 0,83     |
| Личное фото       | 0,78      | 0,74   | 0,76     |
| Фото других людей | 0,75      | 0,77   | 0,76     |

Основная сложность модели связана с различием категорий «личное фото» и «фото других людей», для которых наблюдаются наименьшие значения F1-score (0,76). Портреты имеют визуально сходную структуру, и без дополнительного контекста (например, информации о пользователе) модель не может однозначно определить принадлежность изображения [75, 76, 78].

#### Д.4 Распознавание эмоций

##### Д.4.1 Архитектура модели

Для задачи распознавания эмоций используется специализированная свёрточная нейронная сеть (CNN), предобученная на датасете FER2013 (Facial Expression Recognition 2013) [222], который является стандартным бенчмарком для задач анализа мимики.

Выбор архитектуры CNN для распознавания эмоций обоснован следующим: во-первых, свёрточные сети являются стандартом для задач классификации изображений малого разрешения (48×48) [222]; во-вторых, предложенная архитектура является типичной для задач FER и показывает конкурентоспособные результаты на бенчмарке FER2013 [222]; в-третьих, сравнительно небольшое количество параметров (~2,8 млн) обеспечивает быстрый инференс и устойчивость к переобучению на ограниченных данных.

Датасет FER2013 [222]:

- размер: 35 887 изображений лиц (48×48 пикселей, grayscale);
- количество классов: 7 (радость, грусть, гнев, страх, удивление, отвращение, нейтральное) [207, 222];
- распределение: несбалансированное (например, ~48% – нейтральное, ~8% – отвращение) [222].

Архитектура CNN для FER [222, 217]:

- Входной слой: изображение размером 48×48 пикселей, один канал (grayscale).
- Свёрточный слой: 32 фильтра размером 3×3, функция активации ReLU, пакетная нормализация, слой подвыборки MaxPooling (2×2).
- Свёрточный слой: 64 фильтра размером 3×3, ReLU, пакетная нормализация, MaxPooling (2×2).
- Свёрточный слой: 128 фильтров размером 3×3, ReLU, пакетная нормализация, MaxPooling (2×2).
- Свёрточный слой: 256 фильтров размером 3×3, ReLU, пакетная нормализация, MaxPooling (2×2).
- Операция выравнивания (Flatten): преобразование тензора размерности 3×3×256 в вектор размерности 2 304.
- Полносвязный слой: 512 нейронов, функция активации ReLU, Dropout с вероятностью 0,5 [232].
- Выходной слой: 7 нейронов, функция активации Softmax [224].

На выходе применяется функция softmax, формирующая вероятности принадлежности к 7 классам эмоций [224]:

$$p(y | x) = \text{softmax}(z), z \in \mathbb{R}^7$$

Из-за низкого входного разрешения (48×48) и упрощённой архитектуры количество параметров модели составляет около 2,8 млн, что существенно меньше по сравнению с ResNet50 (25 млн) [7, 222].

#### Д.4.2 Обучение и адаптация к ВК–аватарам

Модель предварительно обучалась на датасете FER2013 [222] (подпараграф Д.4.1).

Достигнутая точность на тестовой выборке составила около 67%, что соответствует уровню базовых CNN-моделей (65–70%) [222]. Сравнительно невысокая точность обусловлена сложностью датасета и низким разрешением изображений (48×48 пикселей) [222].

Для повышения применимости модели к реальным данным проведено дообучение [217] на пользовательских аватарах:

- 3 000 личных фотографий пользователей ВКонтакте;
- ручная экспертная разметка эмоций по 7 базовым эмоциям [207];
- 10 эпох обучения [40].

Точность на тестовой выборке ВК составила 61%. Снижение качества по сравнению с FER2013 связано с более сложными условиями реальных данных: произвольные ракурсы, неконтролируемое освещение, частичная окклюзия лица [78, 222].

Таблица Д.11 – Метрики распознавания эмоций на ВК–аватарах [207, 217]

| Эмоция      | Доля в выборке | Precision | Recall | F1-score |
|-------------|----------------|-----------|--------|----------|
| Радость     | 45%            | 0,72      | 0,68   | 0,70     |
| Нейтральное | 32%            | 0,65      | 0,61   | 0,63     |
| Грусть      | 8%             | 0,54      | 0,49   | 0,51     |
| Удивление   | 6%             | 0,48      | 0,42   | 0,45     |
| Гнев        | 4%             | 0,42      | 0,38   | 0,40     |
| Страх       | 3%             | 0,35      | 0,31   | 0,33     |
| Отвращение  | 2%             | 0,28      | 0,24   | 0,26     |

Детальные метрики по классам приведены в Приложении В (таблица В.9).

Наиболее корректно распознаётся эмоция радости ( $F1 = 0,70$ ), что связано с наличием устойчивых визуальных признаков (улыбка, положение губ, область глаз) [207].

Негативные эмоции (гнев, страх, отвращение) распознаются заметно хуже ( $F1 = 0,26–0,40$ ) по двум основным причинам [78, 222]:

- низкая представленность в выборке (2–4% каждая);
- визуальная близость выражений (напряжённые брови, опущенные уголки губ), что затрудняет их дифференциацию [207].

#### Д.5 Мультимодальная интеграция: архитектура Hybrid Fusion

Мультимодальная модель гибридного слияния предназначена для предсказания 26 психологических факторов, соответствующих структуре опросника 16PF Кэттелла [93, 94] и расширенного набора производных характеристик.

В модель включены следующие группы факторов [93, 94]:

- 16 первичных факторов 16PF (A, B, C, E, F, G, H, I, L, M, N, O, Q1, Q2, Q3, Q4);
- 4 глобальных фактора (F1 – тревожность, F2 – интроверсия, F3 – эмоциональная зрелость, F4 – волевые качества);
- 6 дополнительных агрегированных характеристик: 1 – невротичность, 2 – рефлексивность, 3 – интеллектуальность, 4 – социальность, 5 – целеустремлённость, 6 – конформность.

Все факторы представлены в виде стеновых оценок в диапазоне от 1 до 10 [93, 94].

Подробное описание архитектуры Hybrid Fusion и процедуры обучения приведено в Приложении Б (параграф Б.6).

#### Д.5.1 Общая схема архитектуры

Архитектура Hybrid Fusion объединяет текстовые (никнеймы, статусы) и визуальные (аватары) признаки для предсказания вектора психологических характеристик пользователя [115]:

$$y \in \mathbb{R}^{26}.$$

Модель включает пять основных компонентов [115]:

- извлечение признаков из каждой модальности: RoBERTa [216] – для текстовых данных, ResNet50 [7] – для визуальных данных;
- преобразование признаков в общее латентное пространство фиксированной размерности (512), обеспечивающее сопоставимость модальностей [224];
- Cross-Modal Attention – двунаправленный механизм внимания, позволяющий учитывать взаимное влияние модальностей (Text  $\rightarrow$  Visual и Visual  $\rightarrow$  Text) [5, 115];
- Adaptive Fusion – механизм динамического взвешивания модальностей с обучаемым коэффициентом  $\alpha \in [0,1]$ ;
- регрессионная голова – формирование итогового предсказания по 26 психологическим факторам [224].

Детальное описание компонентов 1–5, включая математические формулы проекций, механизма внимания и адаптивного слияния, приведено непосредственно ниже (подпараграфы Д.5.2 и далее); сводная спецификация архитектуры дана также в приложении Б (параграф Б.6).

### Д.5.2 Компонент 1: Модальные энкодеры

#### Д.5.2.1 Текстовый энкодер

Вход: конкатенация никнейма и статуса [CLS] `nickname_tokens` [SEP] `status_tokens` [SEP] [19, 216].

Выход: эмбединг [CLS] токена – вектор  $h_{text} \in \mathbb{R}^{768}$ , представляющий всю текстовую последовательность [216].

Формализация:

$$h_{text} = \text{RoBERTa}(\text{nickname}, \text{status})[\text{CLS}],$$

где  $\text{RoBERTa}(\dots)$  – 12-слойный Transformer encoder [216].

#### Д.5.2.2 Визуальный энкодер (ResNet50)

Вход: изображение аватара  $224 \times 224 \times 3$ .

Выход: глобальное векторное представление изображения  $h_{visual} \in \mathbb{R}^{2048}$  после операции глобального усреднения (Global Average Pooling) [7].

Формализация:

$$h_{visual} = \text{ResNet50}(\text{avatar}) = \text{AvgPool}(\text{Conv5\_x}),$$

где  $\text{Conv5\_x}$  – выход последнего residual block (feature maps  $7 \times 7 \times 2048$ ), усреднённый до вектора 2048 [7].

### Д.5.3 Компонент 2: Проекционные слои

Цель – привести текстовые (768-мерные) и визуальные (2048-мерные) эмбединги к единому латентному пространству размерности 512 [224].

Формализация [224]:

$$z_{text} = \text{ReLU}(W_{proj}^{text} h_{text} + b_{text}),$$

$$z_{visual} = \text{ReLU}(W_{proj}^{visual} h_{visual} + b_{visual}),$$

где:

- $W_{proj}^{text} \in \mathbb{R}^{512 \times 768}$  – матрица проекции текста;
- $W_{proj}^{visual} \in \mathbb{R}^{512 \times 2048}$  – матрица проекции визуала;
- $b_{text}, b_{visual} \in \mathbb{R}^{512}$  – векторы смещения (bias);
- $\text{ReLU}(x) = \max(0, x)$  – нелинейная функция активации [224].

Результат:  $z_{text}, z_{visual} \in \mathbb{R}^{512}$  – унифицированные представления в общем пространстве.

#### Д.5.4 Компонент 3: Двусторонний механизм внимания

Механизм внимания используется для выявления взаимосвязей между признаками различных модальностей [5, 115]. Например, при наличии в никнейме слова «Wolf» коэффициенты внимания могут возрасти для визуальных признаков, соответствующих изображению волка на аватаре.

Механизм внимания основан на Query-Key-Value (Q-K-V) парадигме [5]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

где:

- $Q$  (Query) – матрица запросов;
- $K$  (Key) – матрица ключей;
- $V$  (Value) – матрица значений;
- $d_k$  – размерность ключевого пространства.

Двусторонний Cross-Modal Attention [115]:

- Text  $\rightarrow$  Visual (текст запрашивает информацию у визуала):

$$Q_{text} = W_Q^{text} z_{text}, K_{visual} = W_K^{visual} z_{visual}, V_{visual} = W_V^{visual} z_{visual},$$

$$z_{text \rightarrow visual} = \text{softmax}\left(\frac{Q_{text} K_{visual}^T}{\sqrt{d_k}}\right) V_{visual}.$$

- Visual  $\rightarrow$  Text (визуал запрашивает информацию у текста):

$$Q_{visual} = W_Q^{visual} z_{visual}, K_{text} = W_K^{text} z_{text}, V_{text} = W_V^{text} z_{text},$$

$$z_{visual \rightarrow text} = \text{softmax}\left(\frac{Q_{visual} K_{text}^T}{\sqrt{d_k}}\right) V_{text},$$

где  $W_Q, W_K, W_V \in \mathbb{R}^{512 \times 512}$  – обучаемые матрицы проекций,  $d_k = 512$  – размерность ключей [5, 115].

Для стабильности обучения добавляются residual connections [7]:

$$z_{text}^{cross} = z_{text} + z_{text \rightarrow visual},$$

$$z_{visual}^{cross} = z_{visual} + z_{visual \rightarrow text}.$$

Результат:  $z_{text}^{cross}, z_{visual}^{cross} \in \mathbb{R}^{512}$  – обогащённые представления, содержащие информацию из обеих модальностей.

#### Д.5.5 Компонент 4: Адаптивное слияние модальностей

Информативность модальностей (текстовой и визуальной) существенно варьирует между пользователями. Для учёта данной вариативности используется механизм Adaptive Fusion, который обучает динамический коэффициент  $\alpha \in [0,1]$ , определяющий вклад каждой модальности для конкретного пользователя.

Формула Adaptive Fusion:

$$\alpha = \sigma(W_{\alpha}^T [z_{text}^{cross}; z_{visual}^{cross}] + b_{\alpha}),$$

$$z_{multimodal} = \alpha \cdot z_{text}^{cross} + (1 - \alpha) \cdot z_{visual}^{cross},$$

где:

- $[z_{text}; z_{visual}] \in \mathbb{R}^{1024}$  – конкатенация текстового и визуального векторов;
- $W_{\alpha} \in \mathbb{R}^{1024}$ ,  $b_{\alpha} \in \mathbb{R}$  – обучаемые параметры;
- $\sigma(x) = \frac{1}{1+e^{-x}}$  – сигмоидная функция, гарантирующая  $\alpha \in (0,1)$  [224];
- $z_{multimodal} \in \mathbb{R}^{512}$  – итоговое мультимодальное представление.

Интерпретация коэффициента  $\alpha$ :

- $\alpha \approx 1$  – модель доверяет тексту (визуал неинформативен);
- $\alpha \approx 0$  – модель доверяет визуалу (текст неинформативен);
- $\alpha \approx 0,5$  – обе модальности равно информативны.

На тестовой выборке ( $N = 258$ ) коэффициент  $\alpha$  принимает значения в диапазоне:

$$\alpha \in [0,12; 0,82], \bar{\alpha} = 0,54, \text{median} = 0,52.$$

В среднем обе модальности вносят сопоставимый вклад, однако для отдельных пользователей наблюдается выраженное доминирование одной из них.

Таблица Д.12 – Зависимость  $\alpha$  от типа признаков

| Категория пользователей                           | Среднее $\alpha$ | Интерпретация                     |
|---------------------------------------------------|------------------|-----------------------------------|
| Неинформативный никнейм ("user123") и личное фото | 0,15             | доминирует визуальная модальность |
| Литературный никнейм и аватар-пейзаж              | 0,78             | доминирует текстовая модальность  |
| Метафорический никнейм и селфи с улыбкой          | 0,52             | сбалансированный вклад            |

#### Д.5.6 Компонент 5: Классификационная голова

Заключительный компонент модели выполняет преобразование мультимодального представления  $z_{multimodal} \in \mathbb{R}^{512}$  в вектор предсказаний психологических факторов [224].

Линейное преобразование задаётся следующим образом [224]:

$$\hat{y} = W_{out} z_{multimodal} + b_{out},$$

где:

- $W_{out} \in \mathbb{R}^{26 \times 512}$  – выходная матрица весов;
- $b_{out} \in \mathbb{R}^{26}$  – вектор смещения;
- $\hat{y} \in \mathbb{R}^{26}$  – предсказанные значения 26 факторов (непрерывные значения, диапазон 1–10 стенов) [93, 94].

#### Д.5.7 Функция потерь

Для обучения модели используется комбинированная функция потерь, учитывающая как точность предсказаний, так и сохранение корреляционной структуры факторов:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{Corr}},$$

где  $\lambda = 0,1$  – вес корреляционного штрафа.

1. Mean Squared Error (MSE) [224]:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N \cdot F} \sum_{i=1}^N \sum_{j=1}^F (\hat{y}_{ij} - y_{ij})^2,$$

где  $F = 26$  – количество факторов,  $N$  – размер батча.

2. Correlation Loss

Стандартная функция MSE оптимизирует каждую компоненту независимо и не учитывает взаимосвязи между факторами. Для сохранения структуры зависимостей вводится дополнительный штраф:

$$\mathcal{L}_{\text{Corr}} = 1 - \frac{1}{F \cdot (F - 1)} \sum_{i=1}^F \sum_{k \neq j} \text{corr}(\hat{y}_j, y_j) \cdot \text{corr}(\hat{y}_k, y_k),$$

где  $\text{corr}(\cdot, \cdot)$  – корреляция Пирсона между вектором предсказаний и истинных значений для фактора [224].

Корреляция  $\text{corr}(\hat{y}_j, y_j)$  оценивается на уровне пользователей (across users), а не на уровне батчей, тем самым обеспечивается устойчивая оценка корреляционной структуры. Если модель хорошо предсказывает корреляционную структуру личности (например, Фактор Н и Фактор Е обычно положительно коррелированы [93, 94]), то  $\mathcal{L}_{\text{Corr}}$  будет низким.

#### Д.5.8 Обучение модели

Стратификация гарантирует, что распределения демографических и категориальных переменных одинаковы во всех трёх выборках [197].

Результаты обучения мультимодальной модели (MAE, Pearson  $r$ , F1) и их сравнение с базовыми моделями приведены в Приложении Е и Приложении В.



Таблица Д.13 – Гиперпараметры обучения

| Гиперпараметр                          | Значение    | Обоснование                                            |
|----------------------------------------|-------------|--------------------------------------------------------|
| Оптимизатор                            | AdamW [219] | Стандарт для Transformer-based моделей [219]           |
| Скорость обучения                      | 1e-4        | Компромисс между скоростью и стабильностью [219]       |
| Размер пакета                          | 16          | Ограничено памятью GPU (модель: 158M параметров) [217] |
| Количество эпох обучения               | 30          | Ранняя остановка после 25 эпох [40]                    |
| Коэффициент регуляризации              | 0,01        | L2-регуляризация [219]                                 |
| Число эпох разогрева                   | 3           | Постепенное увеличение lr [219]                        |
| Ограничение нормы градиента            | 1,0         | Предотвращение взрывного роста градиентов [219]        |
| Обучение со смешанной точностью (FP16) | Да          | Ускорение обучения на GPU с Tensor Cores [217]         |

Таблица Д.14 – Разбиение данных ( $N = 1\,722$ )

| Выборка    | Размер | Доля | Стратификация                                           |
|------------|--------|------|---------------------------------------------------------|
| Train      | 1206   | 70%  | По полу (57,6% мужчины), категориям никнеймов [212,215] |
| Validation | 258    | 15%  | Те же критерии                                          |
| Test       | 258    | 15%  | Те же критерии                                          |

## Д.6 Метрики качества

Для оценки качества мультимодальной модели используются три группы метрик, соответствующие регрессионной постановке задачи (26 факторов) и её классификационной интерпретации [224]. Детальные метрики модели приведены в Приложении В и Приложении Е.

### Д.6.1 Средняя абсолютная ошибка (MAE)

Для оценки качества используется средняя абсолютная ошибка (Mean Absolute Error, MAE) [224]:

$$\text{MAE} = \frac{1}{N \cdot F} \sum_{i=1}^N \sum_{j=1}^F |\hat{y}_{ij} - y_{ij}|,$$

где:

- $N$  – размер выборки;
- $F = 26$  – количество факторов;
- $\hat{y}_{ij}$  – предсказанное значение фактора  $j$  для пользователя  $i$ ;
- $y_{ij}$  – истинное значение.

Единица измерения: стены (диапазон 1–10), MAE = 3,1 означает среднюю ошибку 3,1 стена [93, 94].

MAE выбрана в качестве основной метрики регрессии, поскольку она интерпретируема в единицах измерения целевых переменных и устойчива к выбросам [224].

### Д.6.2 Коэффициент корреляции Пирсона ( $r$ )

Для оценки согласованности предсказанных и истинных значений используется коэффициент корреляции Пирсона для каждого фактора [224]:

$$r_j = \frac{\sum_{i=1}^N (\hat{y}_{ij} - \bar{\hat{y}}_j)(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^N (\hat{y}_{ij} - \bar{\hat{y}}_j)^2} \sqrt{\sum_{i=1}^N (y_{ij} - \bar{y}_j)^2}},$$

где:

- $r_j$  – корреляция между предсказанными и истинными значениями для фактора  $j$ ;
- $\bar{\hat{y}}_j, \bar{y}_j$  – средние значения предсказанных и истинных значений соответственно [224].

Среднее  $r$  по всем 26 факторам:

$$\bar{r} = \frac{1}{F} \sum_{j=1}^F r_j.$$

Результат модели:  $\bar{r} = 0,72$  (сильная корреляция) [224].

Коэффициент корреляции Пирсона позволяет оценить, насколько хорошо модель воспроизводит относительный порядок значений факторов, что важно для психометрических шкал [93, 94].

### Д.6.3 F1-score (для дискретизированных факторов)

Для оценки классификационной производительности психологические факторы дискретизированы в 3 категории [40, 93, 94]:

- Низкий: 1–3 стена.
- Средний: 4–7 стенов.
- Высокий: 8–10 стенов.

F1-score (macro-average) [40]:

$$F1_{macro} = \frac{1}{3} \sum_{c=1}^3 F1_c,$$

где  $F1_c$  – F1-score для категории  $c$  (низкий/средний/высокий) [40].

Полученное значение  $F1 = 0,81$  говорит о высокой точности классификации при переходе от регрессионной постановки к дискретной интерпретации факторов [40, 224].

Дискретизация факторов по стеновым значениям соответствует стандартной психометрической практике интерпретации результатов 16PF [93, 94].

## ПРИЛОЖЕНИЕ Е Результаты исследования вклада компонентов модели и таблицы корреляций по 26 факторам личности

### Е.1 Оценка вклада архитектурных компонентов модели

Цель исследования вклада компонентов модели (ablation study) – количественно оценить вклад отдельных компонентов архитектуры гибридного слияния (Hybrid Fusion) в итоговую производительность модели. Для этого проводится поочерёдное исключение компонентов с последующим измерением изменения качества.

Методика исследования вклада компонентов модели соответствует стандартным подходам в области интерпретируемого машинного обучения [176]. Детальное описание архитектуры гибридного слияния и гиперпараметров обучения приведено в Приложении Д (параграф Д.5) и Приложении Б (параграф Б.6).

Все 26 психологических факторов представлены в шкале стенов (диапазон 1–10, стандарт для 16PF Кэттелла), поэтому значения ошибки MAE интерпретируются в тех же единицах (например, MAE = 3,1 соответствует средней ошибке 3,1 стена). В качестве базовой используется полная модель гибридного слияния, включающая:

- кросс-модальное внимание (Cross-Modal Attention);
- адаптивное слияние модальностей (Adaptive Fusion);
- остаточные связи (residual connections);
- корреляционный член функции потерь (Correlation Loss).

Рассматриваются модифицированные версии модели, в которых последовательно исключается один из компонентов:

1. Без кросс-модального внимания – используется простая конкатенация  $Z_{text}$  и  $Z_{visual}$  без механизма внимания.
2. Без адаптивного слияния модальностей – фиксированное объединение модальностей:  $\alpha = 0,5$ .
3. Без остаточных связей в кросс-модальном внимании – используются только  $Z_{text \rightarrow visual}$  и  $Z_{visual \rightarrow text}$  без добавления исходных представлений.
4. Без корреляционного члена функции потерь – используется только
$$\mathcal{L} = \mathcal{L}_{MSE}.$$
5. Базовая модель слияния – конкатенация эмбеддингов на входе без механизмов внимания и адаптивного взвешивания.
6. Базовая модель позднего слияния – независимые предсказания по модальностям с последующим усреднением.

Для сравнения моделей используются следующие метрики:

- MAE (Mean Absolute Error) – средняя абсолютная ошибка (в стенах);
- Pearson correlation  $r$  – согласованность предсказанных и истинных значений;
- F1-score (macro) – вычисляется после дискретизации факторов на три уровня (низкий / средний / высокий) по 33-му и 67-му перцентилям;
- Макроусреднение обеспечивает равный вклад всех классов независимо от их частоты.

Все варианты моделей обучаются: с одинаковыми гиперпараметрами (см. Приложение Д.5) и на одинаковых разбиениях данных:  $N = 1722$  (1206/258/258).

Таким образом обеспечивается корректность сравнения и исключается влияние внешних факторов.

#### Е.1.1 Результаты исследования вклада компонентов модели

- MAE,  $\Delta r$ ,  $\Delta F1$  – относительное изменение по сравнению с полной моделью (Full Model).
- Положительное  $\Delta MAE$  и отрицательные  $\Delta r$ ,  $\Delta F1$  указывают на ухудшение качества.
- Значения « $\pm$ » соответствуют стандартному отклонению по 5-фолдовой кросс-валидации.

Таблица Е.1 – Вклад каждого компонента архитектуры

| №        | Модель<br>(компонент удалён)                    | MAE           | $\Delta$ MAE | Pearson $r$     | $\Delta r$ | F1<br>(macro) | $\Delta F1$ | Параметры |
|----------|-------------------------------------------------|---------------|--------------|-----------------|------------|---------------|-------------|-----------|
| 0        | Full Model<br>(Hybrid Fusion)                   | 3,1 $\pm$ 0,3 | –            | 0,72 $\pm$ 0,04 | –          | 0,81          | –           | 158M      |
| 1        | – Cross-Modal<br>Attention                      | 3,5 $\pm$ 0,3 | +13%         | 0,67 $\pm$ 0,04 | -7%        | 0,76          | -6%         | 156M      |
| 2        | – Adaptive<br>Fusion (fixed<br>$\alpha = 0,5$ ) | 3,4 $\pm$ 0,3 | +10%         | 0,69 $\pm$ 0,04 | -4%        | 0,77          | -5%         | 158M      |
| 3        | – Residual<br>connections                       | 3,3 $\pm$ 0,3 | +6%          | 0,70 $\pm$ 0,04 | -3%        | 0,79          | -2%         | 158M      |
| 4        | – Correlation<br>Loss                           | 3,2 $\pm$ 0,3 | +3%          | 0,70 $\pm$ 0,04 | -3%        | 0,80          | -1%         | 158M      |
| 5        | Early Fusion<br>(конкатенация)                  | 3,6 $\pm$ 0,4 | +14%         | 0,65 $\pm$ 0,05 | -10%       | 0,74          | -9%         | 150M      |
| 6        | Late Fusion<br>(усреднение<br>$\alpha = 0,5$ )  | 3,5 $\pm$ 0,3 | +11%         | 0,67 $\pm$ 0,04 | -7%        | 0,76          | -6%         | 150M      |
| Baseline | Text-only<br>(RoBERTa)                          | 3,8 $\pm$ 0,4 | +18%         | 0,62 $\pm$ 0,05 | -14%       | 0,71          | -12%        | 125M      |
| Baseline | Visual-only<br>(ResNet50)                       | 4,2 $\pm$ 0,5 | +26%         | 0,55 $\pm$ 0,06 | -24%       | 0,68          | -16%        | 25M       |

### Е.1.2. Интерпретация результатов

Критичные компоненты (ухудшение >10%)

1. Кросс-модальное внимание (+13% MAE, -7% Pearson  $r$ ) – основной компонент архитектуры. Его удаление приводит к замене механизма внимания на простую конкатенацию признаков без моделирования семантических соответствий между модальностями:

$$z_{concat} = [z_{text}; z_{visual}] \in \mathbb{R}^{1024} \Rightarrow z_{multimodal} = W_{proj} z_{concat} + b.$$

В этом случае модель теряет способность выборочно усиливать релевантные кросс-модальные связи (например, соответствие «Wolf» в никнейме и изображение волка в аватаре), что приводит к ухудшению качества.

Снижение корреляции особенно выражено для факторов:

- I (Чувствительность): 0,78 до 0,68 (–13%),
- M (Мечтательность): 0,74 до 0,66 (–11%).

Эти факторы наиболее чувствительны к согласованности текстовых и визуальных признаков. Таким образом, кросс-модальное внимание является основным механизмом, обеспечивающим преимущество гибридного слияния над ранним и поздним слиянием.

2. Адаптивное слияние модальностей (+10% MAE против фиксированного  $\alpha = 0,5$ )

При фиксированном значении  $\alpha = 0,5$  вклад модальностей становится одинаковым для всех пользователей:

$$z = 0,5 \cdot z_{text} + 0,5 \cdot z_{visual}.$$

Однако информативность модальностей существенно варьирует:

- для неинформативного текста («user123») должна доминировать визуальная модальность ( $\alpha \rightarrow 0$ );
- для аватара-пейзажа – текст ( $\alpha \rightarrow 1$ ).

Эмпирический анализ ( $N = 52$  пользователей с неинформативным текстом) показал:

- Адаптивное слияние модальностей: MAE = 3,6,  $\alpha = 0,18$ ;
- фиксированное  $\alpha = 0,5$ : MAE = 4,8 (+33%).

Таким образом, адаптивное слияние модальностей является важным компонентом для обеспечения робастности модели при неполных или шумных данных.

### 3. Раннее и гибридное слияние (+14% MAE, -10% Pearson r)

В раннем слиянии объединение происходит на раннем этапе:

$$z_{early} = [h_{text}; h_{visual}] \in \mathbb{R}^{2816},$$

с последующей проекцией в общее пространство.

Данный подход имеет три фундаментальных ограничения:

- несогласованность пространств признаков (768 и 2048 размерностей), которая приводит к доминированию визуальной модальности;
- отсутствие механизма внимания, ограничивающее способность выявлять кросс-модальные соответствия;
- отсутствие адаптивного взвешивания, снижающего устойчивость модели.

Несмотря на меньшее число параметров (150 млн вместо 158 млн), раннее слияние показывает худшие результаты (MAE 3,6 против 3,1). Следовательно, качество определяется архитектурой, а не только размером модели.

Гибридное слияние обеспечивает выигрыш за счёт унификации представлений, моделирования взаимодействий и адаптивного управления модальностями.

### 4. Остаточные связи (+6% MAE) реализуются как:

$$z^{cross} = z + z_{attention}$$

и сохраняют исходную информацию, наряду с результатами внимания.

Их удаление приводит к:

$$z^{cross} = z_{attention},$$

что ухудшает стабильность обучения и усиливает риск потери информации.

Наблюдаемое ухудшение (MAE 3,3 вместо 3,1) подтверждает, что остаточные связи играют важную роль в стабилизации обучения, но не являются определяющим фактором качества.

### 5. Корреляционный член функции потерь (+3% MAE)

Комбинированная функция потерь:

$$\mathcal{L} = \mathcal{L}_{MSE} + 0,1 \cdot \mathcal{L}_{Corr}.$$

Корреляционный член функции потерь учитывает взаимосвязи между психологическими факторами, тогда как MSE оптимизирует их независимо.

Минимальное ухудшение при удалении (+3%) показывает, что MSE уже обеспечивает хорошую точность, однако корреляционный член функции потерь улучшает структурную согласованность предсказаний.

Это особенно важно с точки зрения психологической интерпретируемости, поскольку факторы личности статистически связаны (например,  $r(H, E) \approx 0,4$ ).

## Е.2. Таблица корреляций мультимодальной модели по 26 факторам

Для каждого из 26 психологических факторов вычислена корреляция Пирсона между предсказанными и истинными значениями на тестовой выборке ( $N = 258$ ).

Сравниваются три модели:

- Text-only (RoBERTa) – текстовая модальность (никнейм, статус);
- Visual-only (ResNet50) – визуальная модальность (аватар);
- Multimodal (Hybrid Fusion) – объединённая модель.

Метрика улучшения определяется как:

$$\Delta r = r_{multimodal} - \max(r_{text}, r_{visual}).$$

Таблица Е.2 – Корреляции (Pearson  $r$ ) между предсказанными и истинными значениями психологических факторов для различных модальностей ( $N = 258$ )

| № | Психологический фактор                | Text-only $r$ | Visual-only $r$ | Multimodal $r$ | $\Delta r$ (улучшение) | Интерпретация                                                  |
|---|---------------------------------------|---------------|-----------------|----------------|------------------------|----------------------------------------------------------------|
| 1 | Фактор А (Общительность)              | 0,58          | 0,52            | 0,64           | +0,06                  | Текст и визуал (селфи и пейзаж) комплементарны                 |
| 2 | Фактор В (Интеллект)                  | 0,64          | 0,41            | 0,68           | +0,04                  | Текст доминирует (литературные референсы)                      |
| 3 | Фактор С (Эмоциональная стабильность) | 0,48          | 0,50            | 0,52           | +0,02                  | Слабый эффект (статические профили не показывают стабильность) |
| 4 | Фактор Е (Доминантность)              | 0,61          | 0,68            | 0,74           | +0,06                  | Визуал чуть важнее (псовые зоонимы в аватарах)                 |
| 5 | Фактор F (Беспечность)                | 0,66          | 0,54            | 0,71           | +0,05                  | Текст доминирует (юмор, оптимистичные статусы)                 |
| 6 | Фактор G (Ответственность)            | 0,52          | 0,48            | 0,55           | +0,03                  | Слабый эффект (сложно проявляется в профилях)                  |
| 7 | Фактор Н (Активность)                 | 0,63          | 0,72            | 0,81           | +0,09                  | Сильный эффект: визуал (групповые фото), текст (знаменитости)  |
| 8 | Фактор I (Чувствительность)           | 0,67          | 0,72            | 0,84           | +0,12                  | Сильный эффект: зоонимы (текст, визуал), эмоции на фото        |
| 9 | Фактор L (Честность)                  | 0,64          | 0,49            | 0,69           | +0,05                  | Текст доминирует (неименованные никнеймы)                      |

Продолжение таблицы Е.2

|                             |                                                   |               |               |           |             |                                                                          |
|-----------------------------|---------------------------------------------------|---------------|---------------|-----------|-------------|--------------------------------------------------------------------------|
| 10                          | Фактор М<br>(Мечтательность)                      | 0,68          | 0,60          | 0,78      | +0,10       | Сильный эффект:<br>литературные ники<br>и символические<br>аватары       |
| 11                          | Фактор N<br>(Дипломатичность)                     | 0,59          | 0,51          | 0,62      | +0,03       | Текст доминирует<br>(тон статусов)                                       |
| 12                          | Фактор O<br>(Неуверенность в<br>себе)             | 0,61          | 0,66          | 0,71      | +0,05       | Визуал чуть<br>важнее (грустная<br>эмоция на фото)                       |
| 13                          | Фактор Q1<br>(Радикализм)                         | 0,62          | 0,54          | 0,68      | +0,06       | Текст доминирует<br>(метафоры)                                           |
| 14                          | Фактор Q2<br>(Самостоятельность)                  | 0,56          | 0,58          | 0,61      | +0,03       | Баланс (одиночные<br>фото и социальные<br>статусы)                       |
| 15                          | Фактор Q3 (Психол.<br>проницательность)           | 0,49          | 0,46          | 0,51      | +0,02       | Слабый эффект<br>(внутренний<br>контроль не<br>проявляется<br>визуально) |
| 16                          | Фактор Q4<br>(Напряжённость)                      | 0,58          | 0,61          | 0,65      | +0,04       | Визуал чуть<br>важнее<br>(напряжённые<br>эмоции)                         |
| 17                          | Фактор F1<br>(Тревожность)                        | 0,65          | 0,59          | 0,74      | +0,09       | Сильный эффект:<br>тревожные статусы<br>и эмоции на фото                 |
| 18                          | Фактор F2<br>(Интроверсия)                        | 0,69          | 0,71          | 0,78      | +0,07       | Визуал чуть<br>важнее (одиночные<br>и групповые фото)                    |
| 19                          | Фактор F3 (Эмоц.<br>зрелость)                     | 0,68          | 0,64          | 0,73      | +0,05       | Текст доминирует                                                         |
| 20                          | Фактор F4 (Волевые<br>качества)                   | 0,60          | 0,62          | 0,66      | +0,04       | Баланс                                                                   |
| 21                          | 1. Невротичность                                  | 0,51          | 0,48          | 0,53      | +0,02       | Слабый эффект                                                            |
| 22                          | 2. Рефлексивность                                 | 0,57          | 0,60          | 0,63      | +0,03       | Визуал чуть<br>важнее                                                    |
| 23                          | 3–6. Другие<br>психометрические<br>характеристики | 0,53–<br>0,62 | 0,48–<br>0,58 | 0,56–0,67 | +0,02–+0,06 | Умеренные<br>эффекты                                                     |
| Среднее<br>по 26<br>фактор. | 0,60                                              | 0,56          | 0,67          | +0,07     | —           |                                                                          |

Среднее значение коэффициента корреляции для мультимодальной модели составляет  $r = 0,67$ . Для текстовой модели получено значение  $r = 0,60$ , для визуальной –  $r = 0,56$ .

Положительное значение  $\Delta r$  наблюдается для всех факторов, что указывает на устойчивый вклад мультимодальной интеграции. Наиболее выраженный эффект ( $\Delta r \geq 0,09$ ) наблюдается для факторов:

- Н (Активность),
- I (Чувствительность),



- М (Мечтательность),
- F1 (Тревожность).

Данные факторы в наибольшей степени зависят от согласованности текстовых и визуальных признаков, а не от одной модальности.

#### Е.2.1 Анализ результатов

Факторы с выраженным мультимодальным эффектом ( $\Delta r \geq +0,09$ ):

1. Фактор I (Чувствительность),  $\Delta r = +0,12$  (максимальный эффект).

Чувствительность последовательно проявляется в обеих модальностях:

- текст: зоонимы (кошачьи, кролики),  $r = 0,67$ ;
- визуал: эмоциональные выражения (грусть, мягкость) и изображения животных,  $r = 0,72$ ;
- мультимодальность: согласованные сигналы усиливают друг друга (например, никнейм «Kitty» и фото с кошкой),  $r = 0,84$ .

Примеры:

- «TenderRabbit» и изображение кролика:  $I = 9,2$  (gt 9,1);
- «SoftHeart» и грустное селфи:  $I = 8,7$  (gt 8,9).

2. Фактор М (Мечтательность),  $\Delta r = +0,10$

Текстовые признаки (литературные референсы,  $r = 0,68$ ) и визуальные (символические аватары,  $r = 0,60$ ) дополняют друг друга, обеспечивая итоговую корреляцию  $r = 0,78$ .

3. Фактор Н (Активность),  $\Delta r = +0,09$

Текст (никнеймы, связанные с публичными фигурами,  $r = 0,63$ ) и визуальная модальность (групповые фотографии,  $r = 0,72$ ) вносят взаимодополняющий вклад, что приводит к увеличению  $r$  до 0,81.

4. Фактор F1 (Тревожность),  $\Delta r = +0,09$

Текстовые индикаторы (тревожные формулировки в статусах,  $r = 0,65$ ) и визуальные признаки (негативные эмоции,  $r = 0,59$ ) совместно повышают точность предсказания до  $r = 0,74$ .

В целом, наибольший прирост наблюдается для факторов, в которых признаки одновременно представлены в тексте и визуальной модальности. Следовательно, кросс-модальное согласование играет главную роль в повышении точности модели.

Факторы С (эмоциональная стабильность), G (ответственность) и Q3 (самоконтроль) демонстрируют минимальный прирост качества ( $\Delta r = +0,02 \dots +0,03$ ).

Данные характеристики относятся к внутренним регуляторным чертам и слабо отражаются в статических элементах профиля (текст и аватар). Их надёжное выявление требует учёта динамических поведенческих паттернов, таких как частота публикаций, реакции на события и регулярность активности.

Следовательно, для повышения точности по данным факторам необходимо расширение модели за счёт поведенческой модальности, которая в текущей архитектуре не используется.

### Е.3. SHAP для вклада модальностей по 26 факторам

Для интерпретации вклада модальностей в предсказание психологических факторов используется метод SHAP (SHapley Additive exPlanations), основанный на теории кооперативных игр. Метод количественно оценивает вклад каждой модальности (текстовой и визуальной) в итоговое предсказание модели.

Значения SHAP вычисляются с использованием алгоритма Kernel SHAP, который является модельно-агностическим и корректно применяется к нелинейным моделям. Он аппроксимирует точные значения Шепли посредством взвешенной линейной регрессии по различным комбинациям входных признаков (в данном случае – модальностей).

Вклад модальности  $m \in M$  определяется как:

$$\phi_m(f, x) = \sum_{S \subseteq M \setminus \{m\}} \frac{|S|! \cdot (|M| - |S| - 1)!}{|M|!} [f(S \cup \{m\}) - f(S)],$$

где:

- $M = \{\text{text}, \text{visual}\}$  – множество модальностей;
- $f(S)$  – предсказание модели при использовании только модальностей из множества  $S$ ;
- $\phi_m$  – вклад модальности  $m$  (SHAP value).

Если  $\phi_{\text{text}} > \phi_{\text{visual}}$ , то текстовая модальность вносит больший вклад в предсказание данного фактора, и наоборот.

Для сопоставимости вкладов используется нормированная величина:

$$\text{SHAP}_{\text{text}} = \frac{|\phi_{\text{text}}|}{|\phi_{\text{text}}| + |\phi_{\text{visual}}|} \times 100\%.$$

Аналогично определяется вклад визуальной модальности.

Использование SHAP даёт возможность оценивать вклад отдельных модальностей в предсказание каждого фактора личности.

Таблица Е.3 – SHAP для вклада текстовой и визуальной модальностей в предсказание 26 психологических факторов ( $N = 258$ , тестовая выборка)

| №                | Психологический фактор                      | SHAP text | SHAP visual | Доминирующ. модальность | Интерпретация                                         |
|------------------|---------------------------------------------|-----------|-------------|-------------------------|-------------------------------------------------------|
| 1                | Фактор А (Общительность)                    | 57%       | 43%         | Текст                   | Статусы о дружбе и социальных взаимодействиях         |
| 2                | Фактор В (Интеллект)                        | 72%       | 28%         | Текст                   | Литературные референсы, сложная лексика               |
| 3                | Фактор С (Эмоц. стабильность)               | 51%       | 49%         | Баланс                  | Обе модальности малоинформативны                      |
| 4                | Фактор Е (Доминантность)                    | 44%       | 56%         | Визуал                  | Псовые зоонимы в аватарах (волк)                      |
| 5                | Фактор F (Беспечность)                      | 62%       | 38%         | Текст                   | Юмор, оптимистичные статусы                           |
| 6                | Фактор G (Ответственность)                  | 53%       | 47%         | Баланс                  | Обе модальности малоинформативны                      |
| 7                | Фактор H (Активность)                       | 42%       | 58%         | Визуал                  | Групповые фото, улыбки                                |
| 8                | Фактор I (Чувствительность)                 | 48%       | 52%         | Визуал                  | Эмоции на фото, кошачьи зоонимы                       |
| 9                | Фактор L (Честолюбие)                       | 68%       | 32%         | Текст                   | Неименованные никнеймы                                |
| 10               | Фактор M (Мечтательность)                   | 65%       | 35%         | Текст                   | Литературные ники, философские статусы                |
| 11               | Фактор N (Дипломатичность)                  | 61%       | 39%         | Текст                   | Тон статусов (прямолинейность против вежливости)      |
| 12               | Фактор O (Неуверенность в себе)             | 46%       | 54%         | Визуал                  | Грустная эмоция на фото                               |
| 13               | Фактор Q1 (Радикализм)                      | 59%       | 41%         | Текст                   | Метафоры, креативные неологизмы                       |
| 14               | Фактор Q2 (Самостоятельность)               | 50%       | 50%         | Баланс                  | Одиночные фото против социальные статусы              |
| 15               | Фактор Q3 (Психол. проницательность)        | 52%       | 48%         | Баланс                  | Обе модальности малоинформативны                      |
| 16               | Фактор Q4 (Напряжённость)                   | 47%       | 53%         | Визуал                  | Напряжённые эмоции на фото                            |
| 17               | Фактор F1 (Тревожность)                     | 55%       | 45%         | Текст                   | Тревожные статусы и эмоции на фото (баланс)           |
| 18               | Фактор F2 (Интроверсия)                     | 49%       | 51%         | Визуал                  | Количество людей на фото (одиночные против групповые) |
| 19               | Фактор F3 (Эмоц. зрелость)                  | 58%       | 42%         | Текст                   | Зрелость тона статусов                                |
| 20               | Фактор F4 (Волевые качества)                | 51%       | 49%         | Баланс                  | —                                                     |
| 21               | 1. Невротичность                            | 54%       | 46%         | Текст                   | Слабые сигналы в обеих модальностях                   |
| 22               | 2. Рефлексивность                           | 46%       | 54%         | Визуал                  | Эмоции на фото                                        |
| 23–26            | 3–6. Другие психометрические характеристики | 48–58%    | 42–52%      | Варьируется             | —                                                     |
| Среднее по факт. | 54%                                         | 46%       | Текст       | Текст                   | Текст чуть важнее, но визуал критичен                 |

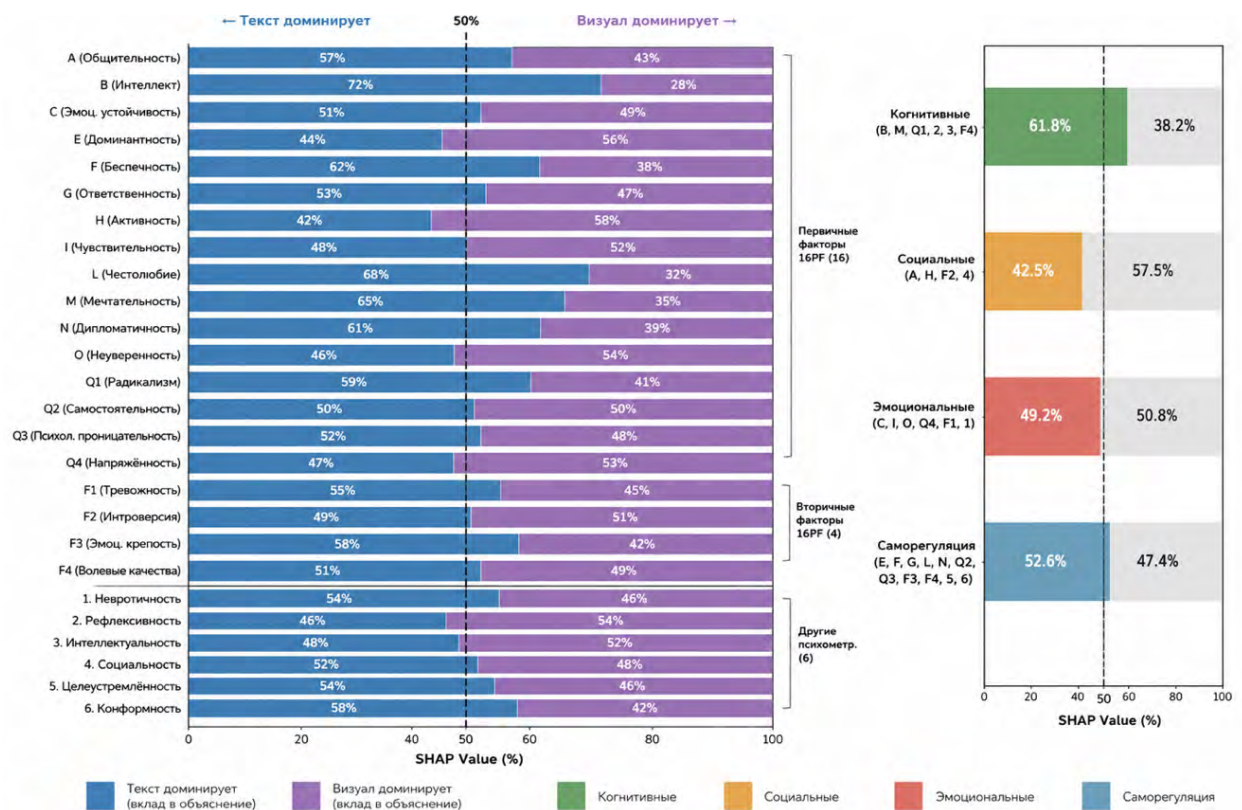


Рисунок Е.1 – SHAP для вклада текстовой и визуальной модальностей

#### Основные выводы:

1. В среднем текстовая модальность вносит несколько больший вклад (54% вместо 46%), однако ни одна из модальностей не является доминирующей для всех факторов.

2. Когнитивные факторы (B, M, Q1) лучше предсказываются на основе текста (SHAP 55–72%), так как вербальная самопрезентация (лексика, стилистика, культурные референсы) напрямую отражает когнитивные характеристики пользователя.

3. Социально-эмоциональные факторы (H, I, E, F1) в большей степени зависят от визуальной модальности (SHAP 52–58%), поскольку невербальные сигналы (эмоции на лице, тип изображения, присутствие других людей, зоонимы) являются более информативными индикаторами данных характеристик.

4. Факторы внутреннего контроля (C, G, Q3) характеризуются близкими значениями SHAP (50/50). Следовательно, низкая информативность текстовой и визуальной модальностей. Это согласуется с результатами корреляционного анализа и доказывает, что данные черты слабо проявляются в статических цифровых следах.

5. Полученные результаты подтверждают корректность использования мультимодальной архитектуры:

- текст обеспечивает более стабильные и интерпретируемые сигналы;

- визуальная модальность добавляет критически важную информацию для социально-эмоциональных факторов;
- их объединение позволяет достичь более высокой точности по всем группам факторов.

Таким образом, SHAP-анализ показывает, что выигрыш мультимодальной модели обусловлен не доминированием одной модальности, а их взаимным дополнением в зависимости от типа предсказываемого фактора.

## ПРИЛОЖЕНИЕ Ж Математические детали SHAP для интерпретации мультимодальной модели

### Ж.1 Теоретические основы из теории игр Shapley Values

Метод SHAP (SHapley Additive exPlanations) основан на значениях Шепли из кооперативной теории игр и используется для количественной оценки вклада отдельных компонентов модели в итоговое предсказание [16].

В рамках кооперативной игры задано множество игроков  $N = \{1, 2, \dots, M\}$ , которые совместно формируют выигрыш, определяемый характеристической функцией  $v(N)$ . Задача состоит в справедливом распределении общего выигрыша между игроками с учётом их индивидуального вклада.

Значения Шепли представляют собой единственное распределение, удовлетворяющее следующим аксиомам:

1. Эффективность – сумма вкладов всех игроков равна общему выигрышу:

$$\sum_{i=1}^M \phi_i = v(N).$$

2. Симметричность – если два игрока вносят одинаковый вклад во все коалиции, их значения Шепли равны.

3. Нулевой игрок – если игрок не влияет на выигрыш ни одной коалиции, его значение Шепли равно 0.

4. Аддитивность – для двух игр с функциями выигрыша  $v_1$  и  $v_2$ , значение Шепли для объединённой игры  $v_1 + v_2$  равно сумме значений Шепли для каждой игры.

Значение Шепли для игрока  $i$  определяется как средний маргинальный вклад игрока во все возможные коалиции:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)],$$

где:

- $N$  – множество всех игроков (размер  $|N| = M$ ),
- $S$  – коалиция (подмножество игроков, не включающее  $i$ ),
- $v(S)$  – выигрыш коалиции  $S$  (функция характеристики коалиции),
- $v(S \cup \{i\}) - v(S)$  – маргинальный вклад игрока  $i$  в коалицию  $S$ ,
- $\frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!}$  – весовой коэффициент, учитывающий все возможные

порядки присоединения игроков к коалиции.

Таким образом, значение Шепли представляет собой усреднённый вклад игрока по всем возможным комбинациям взаимодействия.

В данной работе игроками выступают модальности:  $M = \{\text{text}, \text{visual}\}$ .

Характеристическая функция  $v(S)$  соответствует предсказанию модели Hybrid Fusion при использовании только подмножества модальностей  $S$ .

Тогда значение Шепли  $\phi_m$  отражает вклад модальности  $m \in M$  в итоговое предсказание психологического фактора:

- если  $\phi_{\text{text}} > \phi_{\text{visual}}$ , доминирует текстовая модальность;
- если  $\phi_{\text{visual}} > \phi_{\text{text}}$ , доминирует визуальная модальность.

Таким образом будем интерпретировать модель на уровне факторов и на уровне источников информации.

Точное вычисление значений Шепли требует перебора всех подмножеств  $S \subseteq N$ , что приводит к экспоненциальной сложности:  $O(2^M)$ .

Для практического применения используется аппроксимация, реализуемая в методе Kernel SHAP, который оценивает значения Шепли через взвешенную линейную регрессию по случайным подмножествам признаков. В данной работе Kernel SHAP применяется для оценки вклада модальностей (текст, визуал) в предсказание каждого из 26 психологических факторов.

Рассмотрим пример кооперативной игры с тремя игроками:  $N = \{1, 2, 3\}$  и функцией выигрыша.

Таблица Ж.1 – Значения функции выигрыша для всех коалиций в иллюстративном примере

| Коалиция $S$  | Выигрыш $v(S)$ |
|---------------|----------------|
| $\emptyset$   | 0              |
| $\{1\}$       | 5              |
| $\{2\}$       | 3              |
| $\{3\}$       | 2              |
| $\{1, 2\}$    | 10             |
| $\{1, 3\}$    | 8              |
| $\{2, 3\}$    | 6              |
| $\{1, 2, 3\}$ | 15             |

Значение Шепли для игрока 1:

$$\phi_1 = \frac{1}{6} [(5 - 0) + (10 - 3) + (8 - 2) + (15 - 6)] = \frac{1}{6} (5 + 7 + 6 + 9) = 4,5.$$

Аналогично:

$$\phi_2 = \frac{19}{6} \approx 3,17, \phi_3 = \frac{13}{6} \approx 2,17.$$

Таким образом реализуется принцип вычисления значений Шепли как усреднённого маргинального вклада.

Метод SHAP используется для анализа вклада текстовой и визуальной модальностей в предсказание психологических факторов. Результаты позволяют определить доминирующую модальность для каждого фактора и оценить её влияние на итоговое предсказание модели Hybrid Fusion.

Метод создает теоретически обоснованную и практически применимую интерпретацию предсказаний сложной нейросетевой архитектуры.

## Ж.2. Адаптация Shapley Values для машинного обучения

В задачах машинного обучения значения Шепли используются для объяснения предсказания модели  $f(x)$  для конкретного объекта  $x$  путём оценки вклада отдельных признаков.

В отличие от классической кооперативной игры, где функция выигрыша задаётся явно, в машинном обучении она определяется через поведение модели. Для подмножества признаков  $S \subseteq M$  вводится функция:

$$v(S) = f_S(x_S),$$

где  $f_S(x_S)$  – предсказание модели при использовании только признаков из множества  $S$ .

Поскольку модель обучается на полном пространстве признаков, необходимо определить её поведение при отсутствии части входных данных. Для этого применяется маргинализация отсутствующих признаков:

$$f_S(x_S) = \mathbb{E}_{x_{\bar{S}} \sim p(x_{\bar{S}} | x_S)} [f(x_S, x_{\bar{S}})],$$

где:

- $x_S$  – наблюдаемые признаки,
- $x_{\bar{S}}$  – отсутствующие признаки,
- $p(x_{\bar{S}} | x_S)$  – условное распределение.

Данный подход соответствует методологии маргинализации отсутствующих признаков в задачах интерпретации машинного обучения [224].

На практике точное вычисление математического ожидания затруднено, поэтому используется приближение через усреднение по обучающей выборке:

$$f_S(x_S) \approx \frac{1}{N_{train}} \sum_{k=1}^{N_{train}} f(x_S, x_{\bar{S}}^{(k)}).$$

Данный подход лежит в основе алгоритма Kernel SHAP и позволяет корректно учитывать влияние отсутствующих признаков. [16]



Вклад признака  $i$  в предсказание модели определяется как средний маргинальный вклад по всем возможным подмножествам признаков:

$$\phi_i(f, x) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|! \cdot (|M| - |S| - 1)!}{|M|!} [f_{S \cup \{i\}}(x) - f_S(x)],$$

где:

- $M = \{1, 2, \dots, M\}$  – множество признаков,
- $f_S(x_S)$  – предсказание модели с учётом только признаков из множества  $S$ .

Свойство аддитивности позволяет представить предсказание модели в виде суммы вкладов:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i(f, x),$$

где:

- $\phi_0 = \mathbb{E}[f(x)]$  – базовое значение (среднее предсказание модели);
- $\phi_i$  – вклад признака в отклонение от базового значения.

Значения Шепли обладают прямой интерпретацией:

- $\phi_i > 0$  – признак увеличивает значение предсказания;
- $\phi_i < 0$  – признак уменьшает предсказание;
- $\phi_i \approx 0$  – влияние признака незначительно.

Таким образом, SHAP обеспечивают аддитивное и интерпретируемое разложение предсказания модели.

В классической формулировке SHAP применяется к отдельным признакам. Однако в данной работе признаки агрегируются в более крупные группы – модальности (текстовая и визуальная), что требует адаптации подхода к оценке вкладов.

### Ж.3 SHAP для мультимодальной модели

В мультимодальной модели признаки группируются по модальностям:

- текстовая модальность  $M_{text}$  – эмбединг никнейма и статуса  $h_{text} \in \mathbb{R}^{768}$ ,
- визуальная модальность  $M_{visual}$  – эмбединг аватара  $h_{visual} \in \mathbb{R}^{2048}$ .

Цель – оценить вклад каждой модальности, а не отдельных признаков в предсказание психологического фактора.

Для  $M = 2$  модальностей (текст и визуал) формула Shapley Value упрощается:

$$\phi_{text} = \frac{1}{2} [f(\text{text}) - f(\emptyset)] + \frac{1}{2} [f(\text{text}, \text{visual}) - f(\text{visual})],$$

$$\Phi_{visual} = \frac{1}{2}[f(\text{visual}) - f(\emptyset)] + \frac{1}{2}[f(\text{text,visual}) - f(\text{text})],$$

где:

- $f(\emptyset)$  – базовое предсказание (среднее значение фактора на обучающей выборке);
- $f(\text{text})$  – предсказание модели при использовании только текстовой модальности;
- $f(\text{visual})$  – предсказание модели при использовании только визуальной модальности;
- $f(\text{text,visual})$  – предсказание полной мультимодальной модели.

Упрощение через разность:

$$\Phi_{text} = \frac{1}{2}[f(\text{text}) + f(\text{text,visual}) - f(\text{visual}) - f(\emptyset)],$$

$$\Phi_{visual} = \frac{1}{2}[f(\text{visual}) + f(\text{text,visual}) - f(\text{text}) - f(\emptyset)].$$

Проверка эффективности:

$$\Phi_{text} + \Phi_{visual} = f(\text{text,visual}) - f(\emptyset),$$

что соответствует аксиоме эффективности: сумма вкладов модальностей равна отклонению предсказания от базового значения.

Упрощённая формула для 2 модальностей является частным случаем Shapley-Taylor interaction index для двух коалиций. Этот индекс обобщает Shapley values на случай взаимодействий высших порядков, но для  $M = 2$  сводится к приведённой формуле.

Проблема вычисления  $f(\text{text})$  и  $f(\text{visual})$  для проектируемой модели в том, что полная мультимодальная модель обучена с обеими модальностями. Задача заключается в предсказании при использовании только текста  $f(\text{text})$ .

Решение 1 – маргинализация визуальной модальности. Заменить  $h_{visual}$  на среднее значение визуального эмбединга по обучающей выборке:

$$f(\text{text}) = f(h_{text}, \bar{h}_{visual}),$$

где  $\bar{h}_{visual} = \frac{1}{N_{train}} \sum_{k=1}^{N_{train}} h_{visual}^{(k)}$ .

Замена визуала на среднее значение эквивалентна отсутствию информации из визуальной модальности (визуал становится «нейтральным»).

Решение 2 – обнуление визуальной модальности. Заменить  $h_{visual}$  на нулевой вектор:

$$f(\text{text}) = f(h_{\text{text}} 0).$$

Однако это менее корректно, так как нулевой вектор может быть вне области определения модели, модель не обучалась на нулевых эмбедингах.

В нашей реализации используется Решение 1.

Алгоритм вычисления SHAP для мультимодальной модели

Вход:

- Обученная мультимодальная модель  $f$ .
- Тестовый пользователь  $x$  с текстовым эмбедингом  $h_{\text{text}}(x)$  и визуальным эмбедингом  $h_{\text{visual}}(x)$ .

- Обучающая выборка  $\{x^{(k)}\}_{k=1}^{N_{\text{train}}}$  для вычисления средних эмбедингов.

Алгоритм:

1. Вычислить среднее визуального эмбединга по обучающей выборке:

$$\bar{h}_{\text{visual}} = \frac{1}{N_{\text{train}}} \sum_{k=1}^{N_{\text{train}}} h_{\text{visual}}^{(k)}.$$

2. Вычислить среднее текстового эмбединга:

$$\bar{h}_{\text{text}} = \frac{1}{N_{\text{train}}} \sum_{k=1}^{N_{\text{train}}} h_{\text{text}}^{(k)}.$$

3. Вычислить 4 предсказания:

- $f(\emptyset) = f(\bar{h}_{\text{text}} \bar{h}_{\text{visual}})$  – базовое предсказание;
- $f(\text{text}) = f(h_{\text{text}}(x), \bar{h}_{\text{visual}})$  – предсказание только по тексту;
- $f(\text{visual}) = f(\bar{h}_{\text{text}} h_{\text{visual}}(x))$  – предсказание только по визуалу;
- $f(\text{text}, \text{visual}) = f(h_{\text{text}}(x), h_{\text{visual}}(x))$  – полное предсказание.

4. Вычислить SHAP values:

$$\begin{aligned} \Phi_{\text{text}} &= \frac{1}{2} [f(\text{text}) + f(\text{text}, \text{visual}) - f(\text{visual}) - f(\emptyset)], \\ \Phi_{\text{visual}} &= \frac{1}{2} [f(\text{visual}) + f(\text{text}, \text{visual}) - f(\text{text}) - f(\emptyset)]. \end{aligned}$$

5. Нормализация (для интерпретации в процентах):

$$\begin{aligned} \text{SHAP}_{\text{text}} &= \frac{|\Phi_{\text{text}}|}{|\Phi_{\text{text}}| + |\Phi_{\text{visual}}|} \times 100\%, \\ \text{SHAP}_{\text{visual}} &= \frac{|\Phi_{\text{visual}}|}{|\Phi_{\text{text}}| + |\Phi_{\text{visual}}|} \times 100\%. \end{aligned}$$

Вычислительная сложность:  $O(1)$  – фиксированное количество прямых проходов (4 предсказания модели), независимо от размера выборки.

#### Ж.4 Пример вычисления SHAP для конкретного пользователя

Пользователь: User ID 127

- Никнейм: «PhilosopherStone».
- Статус: «Ищу философский камень в библиотеках».
- Аватар: Фото за книжной полкой.
- Психологический фактор (целевой): Фактор М (Мечтательность).
- Истинное значение:  $y_{true} = 9,0$  стен (высокое абстрактное мышление).

После прохождения через модальные энкодеры (RoBERTa, ResNet50):

- $h_{text}(\text{User 127}) \in \mathbb{R}^{768}$  – текстовый эмбединг (вектор 768 чисел, среднее абсолютное значение  $\approx 0,42$ ),
- $h_{visual}(\text{User 127}) \in \mathbb{R}^{2048}$  – визуальный эмбединг (вектор 2048 чисел, среднее абсолютное значение  $\approx 0,31$ ).

Средние эмбединги по обучающей выборке ( $N = 1206$ ):

- $\bar{h}_{text} \in \mathbb{R}^{768}$  (среднее абсолютное значение  $\approx 0,38$ ),
- $\bar{h}_{visual} \in \mathbb{R}^{2048}$  (среднее абсолютное значение  $\approx 0,29$ ).

Нормализация эмбедингов (LayerNorm в RoBERTa и BatchNorm в ResNet50) сохраняет диапазон значений и обеспечивает сопоставимость масштабов, поэтому усреднение по выборке  $\bar{h} = \frac{1}{N} \sum_{k=1}^N h^{(k)}$  остаётся математически корректным и семантически интерпретируемым как «типичное» представление модальности [216, 7].

Подаём различные комбинации эмбедингов в полную мультимодальную модель:

1. Базовое предсказание (обе модальности усреднены):

$$f(\emptyset) = f(\bar{h}_{text}, \bar{h}_{visual}) = 5,8 \text{ стен.}$$

Интерпретация: среднее значение фактора М (Мечтательность) в обучающей выборке составляет 5,8 стенов, что соответствует нормальному уровню по шкале 16PF.

2. Предсказание только по тексту (визуал усреднён):

$$f(\text{text}) = f(h_{text}(\text{User 127}), \bar{h}_{visual}) = 8,1 \text{ стен.}$$

Интерпретация: модель предсказывает повышенную мечтательность (8,1 стена), что обусловлено сочетанием литературного никнейма и философской тематики статуса.

3. Предсказание только по визуалу (текст усреднён):

$$f(\text{visual}) = f(\bar{h}_{\text{text}} h_{\text{visual}}(\text{User 127})) = 6,8 \text{ стен.}$$

Интерпретация: модель предсказывает умеренную мечтательность (6,8 стена) на основе аватара с книгами, который выступает маркером интеллектуальной самопрезентации.

4. Полное предсказание (обе модальности реальные):

$$f(\text{text,visual}) = f(h_{\text{text}}(\text{User 127}), h_{\text{visual}}(\text{User 127})) = 8,9 \text{ стен.}$$

Интерпретация: синергия модальностей, литературный никнейм вместе с аватаром с книгами, усиливает предсказание мечтательности (8,9), делая его практически совпадающим с истинным значением (9,0).

Вычисление SHAP values:

$$\phi_{\text{text}} = \frac{1}{2} [f(\text{text}) + f(\text{text,visual}) - f(\text{visual}) - f(\emptyset)],$$

$$\phi_{\text{text}} = \frac{1}{2} [8,1 + 8,9 - 6,8 - 5,8] = \frac{1}{2} \cdot 4,4 = 2,2,$$

$$\phi_{\text{visual}} = \frac{1}{2} [f(\text{visual}) + f(\text{text,visual}) - f(\text{text}) - f(\emptyset)],$$

$$\phi_{\text{visual}} = \frac{1}{2} [6,8 + 8,9 - 8,1 - 5,8] = \frac{1}{2} \cdot 1,8 = 0,9.$$

Проверка эффективности:

$$\phi_{\text{text}} + \phi_{\text{visual}} = 2,2 + 0,9 = 3,1,$$

$$f(\text{text,visual}) - f(\emptyset) = 8,9 - 5,8 = 3,1.$$

Эффективность выполняется.

Нормализация (для интерпретации в процентах):

$$\text{SHAP}_{\text{text}} = \frac{|2,2|}{|2,2| + |0,9|} \times 100\% = \frac{2,2}{3,1} \times 100\% = 71\%,$$

$$\text{SHAP}_{\text{visual}} = \frac{|0,9|}{|2,2| + |0,9|} \times 100\% = \frac{0,9}{3,1} \times 100\% = 29\%.$$

Для пользователя User 127 предсказание Фактора М (Мечтательность) на 71% обусловлено текстовой модальностью (литературный никнейм «PhilosopherStone» и философский статус) и на 29% визуальной модальностью (аватар с книгами).

Обе модальности согласованы и указывают на высокую мечтательность, что приводит к синергии и более уверенному предсказанию (8,9 вместо 8,1 text-only и 6,8 visual-only).

### Ж.5 Усреднение SHAP values по тестовой выборке

Для каждого из 26 психологических факторов SHAP values вычисляются для всех  $N_{test} = 258$  тестовых пользователей, затем усредняются:

$$\overline{\text{SHAP}}_{text}^{(j)} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \text{SHAP}_{text}^{(i,j)},$$

где:

- $j$  – индекс психологического фактора (1...26),
- $i$  – индекс тестового пользователя (1...258).

Высокое стандартное отклонение (18–22%) показывает, что вклад модальностей сильно варьирует между пользователями. Следовательно, возникает необходимость применения Adaptive Fusion (динамического взвешивания модальностей).

Таблица Ж.2 – Средние SHAP values по тестовой выборке ( $N = 258$ )

| Психологический фактор      | Средний SHAP text (%) | Средний SHAP visual (%) | Стандартное отклонение SHAP text* |
|-----------------------------|-----------------------|-------------------------|-----------------------------------|
| Фактор М (Мечтательность)   | 65%                   | 35%                     | 18 п.п.                           |
| Фактор I (Чувствительность) | 48%                   | 52%                     | 22 п.п.                           |
| Фактор H (Активность)       | 42%                   | 58%                     | 19 п.п.                           |
| Фактор F1 (Тревожность)     | 55%                   | 45%                     | 20 п.п.                           |
| Среднее по 26 факторам      | 54%                   | 46%                     | 19 п.п.                           |

\*Стандартное отклонение вычислено для нормализованных долей SHAP (в процентах), а не для raw SHAP values (в стенах).

### Ж.6 Вычислительная сложность и оптимизация

Для каждого тестового пользователя требуется 4 forward pass через полную модель (158 млн параметров):

1.  $f(\emptyset)$ .
2.  $f(\text{text})$ .
3.  $f(\text{visual})$ .
4.  $f(\text{text}, \text{visual})$ .

Сложность для  $N_{test} = 258$  пользователей и 26 факторов:

$$4 \cdot 258 \cdot 26 = 26\,832 \text{ forward passes.}$$

Время: 2 часа на NVIDIA RTX 3090 (24GB VRAM) (при batch\_size=1).

Для вычисления SHAP модель используется в режиме оценки (eval()) без расчёта градиентов (torch.no\_grad()). Поскольку обучение модели на данном этапе

не выполняется, вычисляются только значения предсказаний, необходимые для формирования объяснений.

Предсказания  $f(\emptyset)$ ,  $f(\text{text})$ ,  $f(\text{visual})$ ,  $f(\text{text}, \text{visual})$  для разных пользователей независимы и можно обрабатывать батчами.

Оптимизированная реализация:

1. Вычислить средние эмбединги  $\bar{h}_{\text{text}}$ ,  $\bar{h}_{\text{visual}}$  один раз для всей обучающей выборки.
2. Для батча из  $B$  тестовых пользователей выполнить 4 батчевых forward pass:
  - Batch  $[h_{\text{text}}^{(1)}, \dots, h_{\text{text}}^{(B)}]$  + копия  $\bar{h}_{\text{visual}}$   $B$  раз,
  - Batch  $[h_{\text{visual}}^{(1)}, \dots, h_{\text{visual}}^{(B)}]$  + копия  $\bar{h}_{\text{text}}$   $B$  раз,
  - и т.д.

Ускорение: С  $\text{batch\_size} = 32$  время сокращается до 15 минут (ускорение в 8 раз).

#### Ж.7 Ограничения метода SHAP для модели

1. SHAP предполагает, что модальности независимы, т.е.  $p(\text{text}, \text{visual}) = p(\text{text}) \cdot p(\text{visual})$ . В реальности модальности могут быть слабо коррелированы (например, пользователи с литературными никнеймами чаще выбирают символические аватары).

2. Замена  $h_{\text{visual}}$  на  $\bar{h}_{\text{visual}}$  может быть вне области определения модели, если модель не обучалась на такой комбинации. Однако эмпирически это даёт разумные результаты.

3. SHAP оценивает глобальный вклад модальности, но не объясняет, какие конкретные элементы модальности важны (например, какие слова в никнейме или области изображения). Для этого требуется дополнительная визуализация (attention heatmaps, как в параграфе 3.5 Главы 3).

4. Стандартные SHAP измеряют аддитивные вклады модальностей, но не моделируют синергию (нелинейные эффекты взаимодействия). Например, если никнейм «Wolf» усиливает влияние аватара с волком, этот синергетический эффект частично распределяется между модальностями, но не выделяется явно. Для точной оценки взаимодействий требуется Shapley-Taylor interaction index, который имеет ещё большую вычислительную сложность ( $O(3^M)$  вместо  $O(2^M)$ ). В данной работе синергия модальностей частично фиксируется через Cross-Modal Attention (Глава 3), но SHAP её количественно недооценивает.

## ПРИЛОЖЕНИЕ И Визуализации системы анализа цифрового следа

### И.1 Трёхуровневая архитектура системы

На рисунке И.1 представлена архитектура системы. Она реализована в виде трёхуровневой структуры, включающей: уровень представления (клиентский уровень), сервисный уровень (уровень приложений и обработки) и уровень данных (хранение данных и моделей).

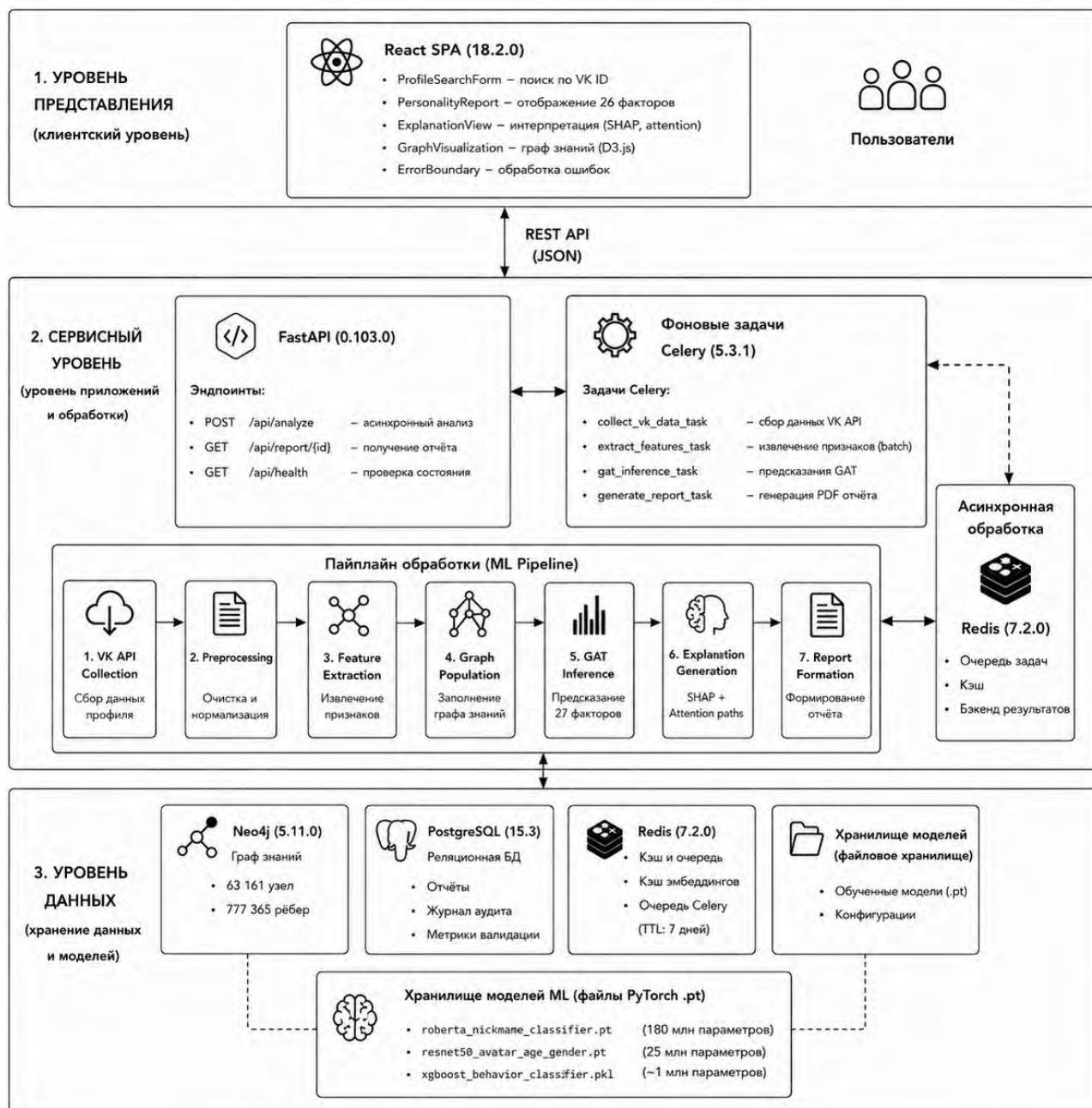


Рисунок И.1 – Трёхуровневая архитектура системы анализа личности по цифровому следу

Клиентская часть реализована в виде одностраничного веб-приложения (React SPA), обеспечивающего взаимодействие пользователя с системой [233].



Основные функции интерфейса:

- ввод идентификатора пользователя (VK ID) и запуск анализа;
- отображение предсказанных психологических факторов;
- визуализация интерпретаций модели (SHAP, механизма внимания);
- отображение графа знаний (D3.js);
- обработка ошибок на стороне интерфейса (Error Boundary).

Взаимодействие с серверной частью осуществляется через REST API (формат JSON).

Сервисный уровень реализован на базе FastAPI и отвечает за обработку запросов, выполнение ML-пайплайна и координацию компонентов системы [235].

Основные компоненты:

- REST API (FastAPI), обеспечивающий: запуск анализа (POST /api/analyze); получение результатов (GET /api/report/{id}); проверку состояния системы (GET /api/health);
- асинхронная обработка (Celery), выполняющая: сбор данных из VK API; извлечение признаков; инференс (получение предсказаний модели) модели; генерацию отчёта;
- очередь задач и кэш (Redis): управление очередью задач; кэширование промежуточных результатов; хранение эмбеддингов;
- ML-конвейер обработки данных.

Обработка данных организована в виде последовательного конвейера:

- сбор данных профиля (VK API);
- предобработка (очистка и нормализация);
- извлечение признаков;
- построение и заполнение графа знаний;
- инференс модели (GAT, предсказание психологических факторов);
- генерация интерпретаций (SHAP, механизм внимания);
- формирование итогового отчёта.

Для масштабирования обработки и выполнения пакетных вычислений конвейер реализуется в асинхронном режиме с использованием Celery.

Уровень данных обеспечивает хранение структурированной информации, графа знаний и обученных моделей.

Используемые компоненты:

- Neo4j – графовая база знаний [236];
- PostgreSQL – хранение отчётов, логов и метрик;
- Redis – кэширование и очередь задач;
- файловое хранилище моделей – хранение обученных моделей (PyTorch, XGBoost).

Основные свойства архитектуры:

- логика разделена между клиентом, сервисным уровнем и хранилищем данных;
- асинхронная обработка и очереди задач позволяет обрабатывать до ~200 профилей в минуту;
- реализованы проверки состояния сервисов и механизмы повторного выполнения задач;
- используется GPU-инференс и кэширование эмбеддингов.

## И.2 Алгоритмический конвейер обработки

На рисунке И.2 представлена последовательность обработки данных пользователя, реализованная в системе. Конвейер включает семь основных этапов, обеспечивающих переход от сырых данных профиля к интерпретируемым результатам анализа.

Оценка времени выполнения каждого этапа приведена для одного пользовательского профиля при стандартной конфигурации системы (GPU-инференс, асинхронная обработка):

- Этап 1. Сбор данных (VK API) – 2–3 секунд (ограничено внешним API, rate limit до 20 запросов в секунду);
- Этап 2. Предобработка – 0,1–0,2 секунды (обнаружение и нормализация лиц, YOLOv5x на CPU);
- Этап 3. Извлечение признаков – 0,3–0,5 секунды (текстовые и визуальные эмбеддинги: RoBERTa и ResNet50, GPU);
- Этап 4. Построение графа знаний – 0,15 секунды (заполнение структуры графа, Neo4j, Cypher-запросы);
- Этап 5. Инференс модели (GAT) – 0,1 секунды (предсказание психологических факторов, batch-обработка);

- Этап 6. Генерация интерпретаций – 0,05 секунды (вычисление SHAP и путей внимания);
- Этап 7. Формирование отчёта – около 0,1 секунды (агрегация результатов и сериализация в JSON).

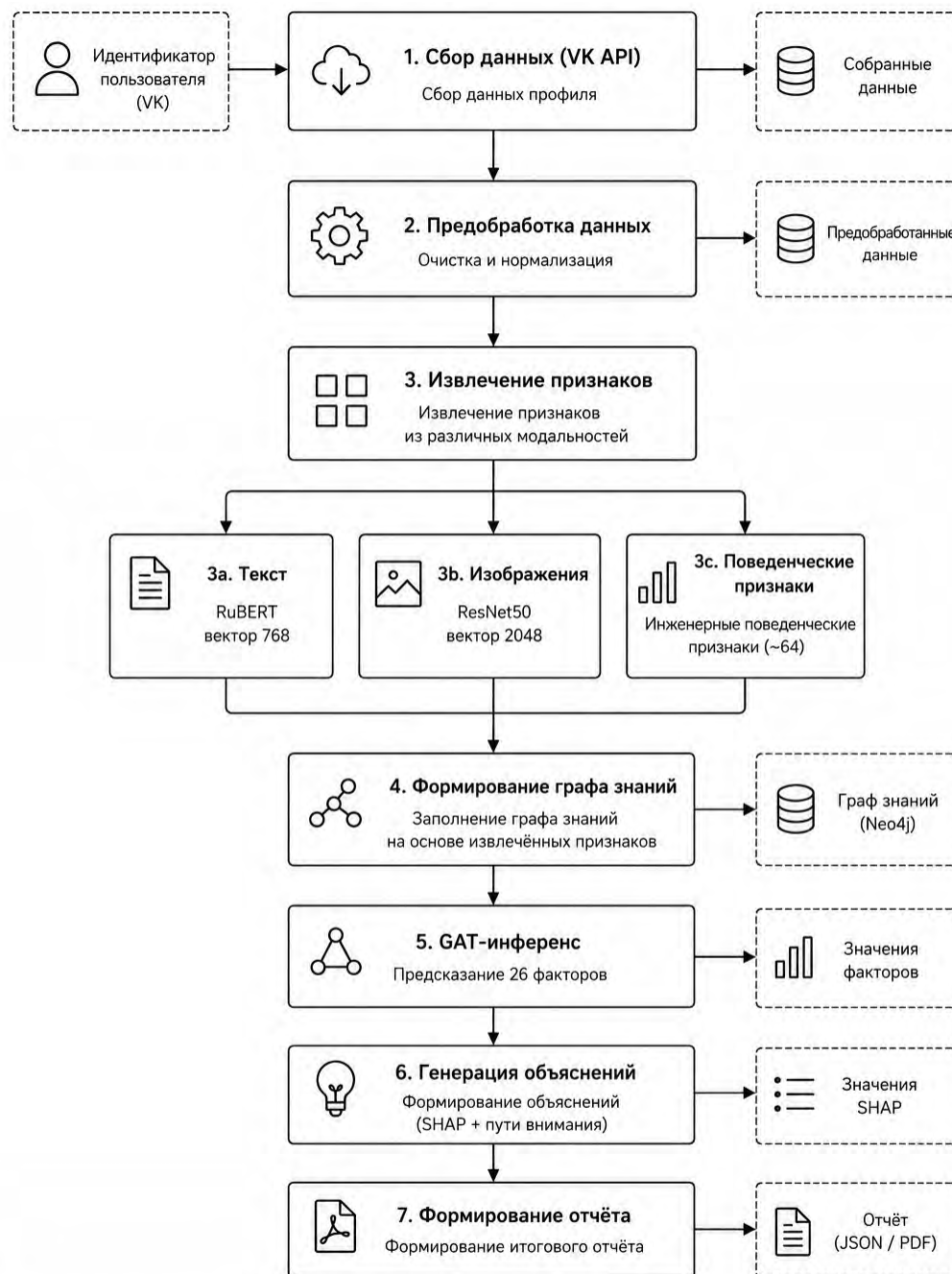


Рисунок И.2 – Алгоритмический конвейер обработки пользовательского профиля

Базовое время обработки одного профиля составляет 3–4 секунды на GPU. При повторных запросах с использованием кэширования эмбедингов время обработки сокращается примерно до 1 секунды. При параллельной пакетной обработке производительность системы превышает 200 профилей в минуту.

### И.3 Граф знаний: структура узлов и связей

На рисунке И.3 представлена структура графа знаний, используемого для моделирования взаимосвязей между пользователем, извлечёнными признаками и психологическими факторами.

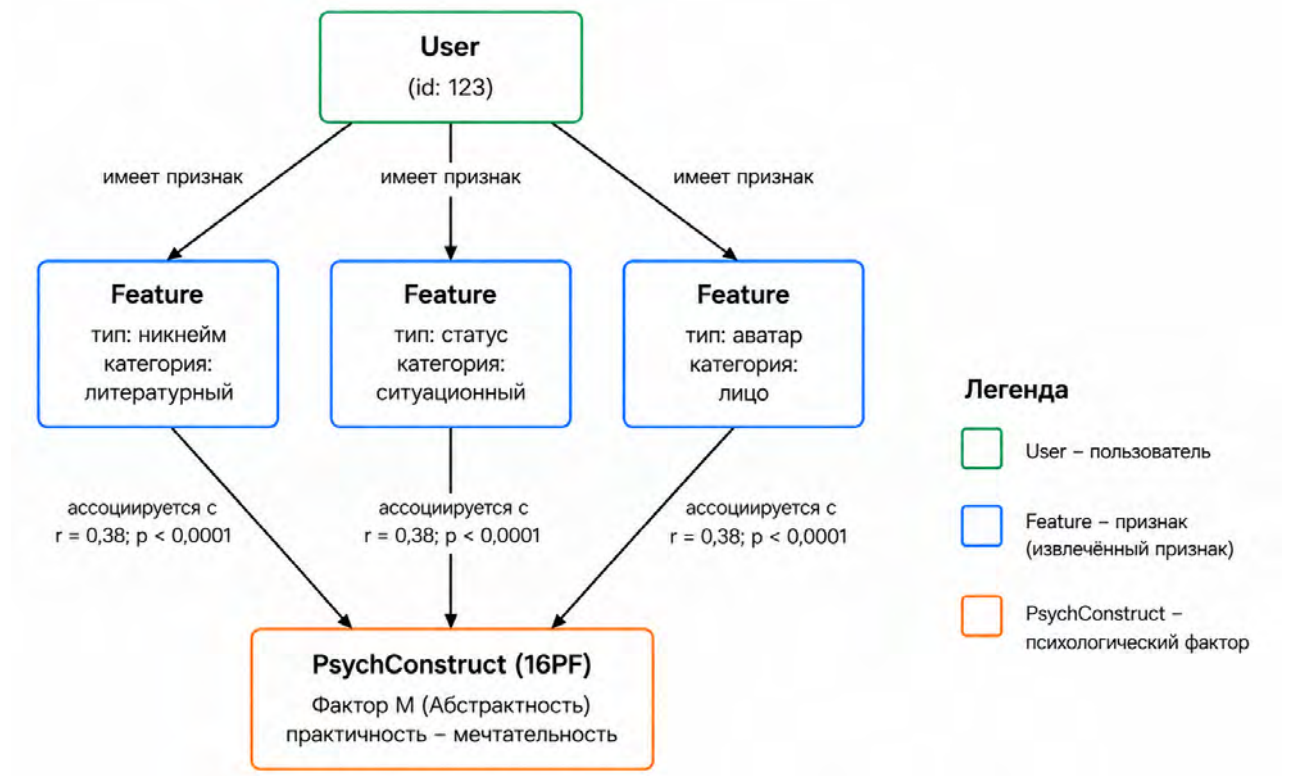


Рисунок И.3 – Структура графа знаний

Граф имеет иерархическую структуру и включает три основных типа сущностей: пользователь (User), признаки (Feature) и психологические конструкции (PsychConstruct), соответствующие 26 целевым характеристикам.

Связи между узлами отражают как принадлежность признаков пользователю, так и статистические зависимости между признаками и психологическими характеристиками.

Характеристики графа:

- узлы: 63 161, включая: User (19 690), Feature (категории признаков), PsychConstruct (психологические конструкторы), а также экземпляры признаков;
- – связи: 777 365, включая: HAS\_FEATURE (пользователь → признак), ASSOCIATED\_WITH (признак → психологический конструктор), CORRELATES\_WITH (психологический конструктор ↔ психологический конструктор), а также дополнительные семантические и структурные связи;
- – типы признаков (Feature): nickname, status, avatar, temporal, network;
- – психологические конструкторы: 26 целевых характеристик;

- вес рёбер: коэффициенты корреляции Пирсона ( $r \in [-1; 1]$ ) с учётом статистической значимости.

## И.4 Скриншоты веб-интерфейса

### И.4.1 Главная страница

На главной странице реализована форма ввода идентификатора пользователя социальной сети ВКонтакте (VK ID). Интерфейс обеспечивает взаимодействие пользователя с системой анализа и инициирует запуск полного пайплайна обработки мультимодальных данных.

Справа отображается информационная панель с характеристиками модели (тип модели, количество параметров, используемые признаки, экспертная оценка), а также перечень используемых источников данных (текст, изображения, поведенческие признаки).

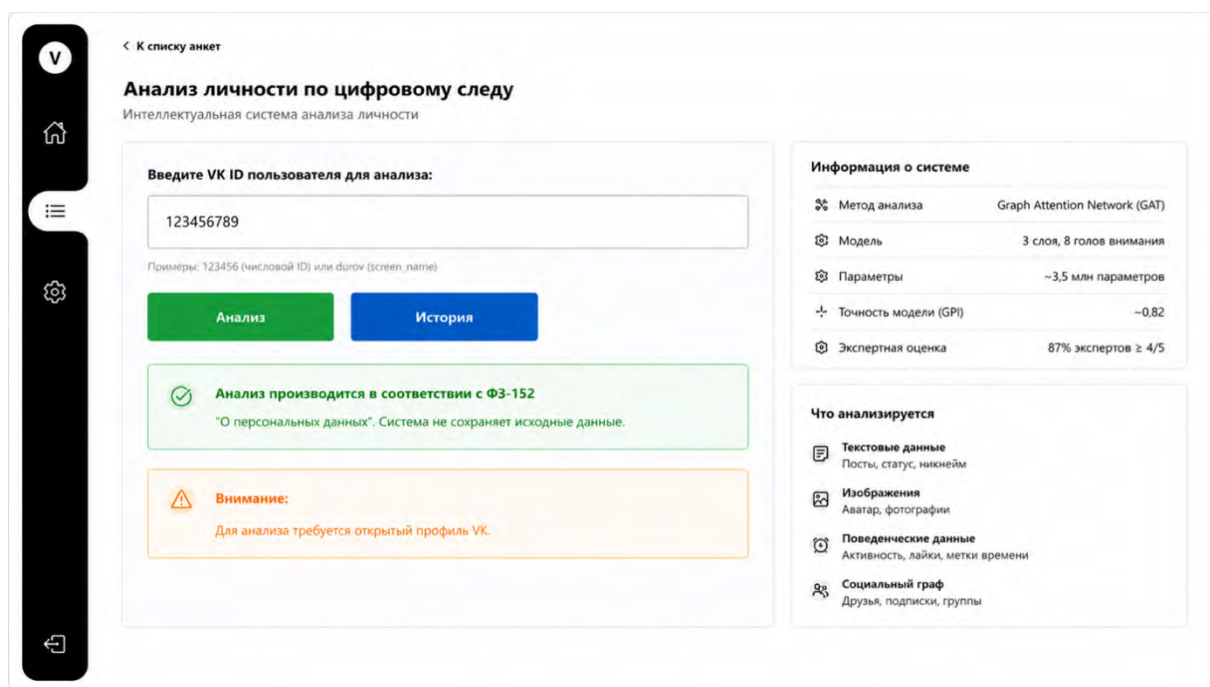


Рисунок И.4 – Главная страница веб-интерфейса системы анализа личности по цифровому следу

Функционал системы:

- поддерживаются два формата VK ID: числовой идентификатор (регулярное выражение  $^{\wedge}\backslash d+\$$ ) и символьный screen\_name (регулярное выражение  $^{\wedge}[a-zA-Z0-9\_]+\$$ );
- символьный screen\_name автоматически преобразуется в числовой ID пользователя через VK API;
- при отправке запроса запускается асинхронный конвейер обработки данных;

- отображается визуальный индикатор загрузки в процессе выполнения запроса;
- предусмотрена обработка типовых исключений: закрытый профиль, удалённый или несуществующий пользователь, превышение лимитов VK API.

#### И.4.2 Отчёт «Профиль личности»

Веб-интерфейс отчёта представляет результаты анализа личности пользователя, сформированные на основе мультимодальной обработки цифрового следа. Отображаются оценки психологических факторов в шкале стенов (1–10), соответствующей методике 16PF Кеттелла.



Рисунок И.5 – Профиль личности пользователя по факторам 16PF с визуализацией стеновых значений

Основная область интерфейса содержит список факторов с их численными значениями и качественной интерпретацией (низкий, средний, высокий уровень). Визуальное представление реализовано в виде горизонтальных прогресс-баров, обеспечивающих наглядное сравнение факторов между собой.

Дополнительно отображается сводная информация о пользователе (возраст, пол, никнейм, источник данных), а также статус анализа.

Значения факторов представлены в шкале стенов (1–10), что соответствует психометрическим стандартам. Используется цветовая кодировка уровней: зелёный – высокий уровень (7–10), жёлтый – средний (4–6), красный – низкий (1–3).



При наведении на элемент интерфейса отображается всплывающая подсказка, содержащая интерпретацию фактора и оценку уверенности модели.

Интерфейс отчёта обеспечивает интерпретируемое представление результатов работы модели, позволяя пользователю быстро оценить профиль личности и выявить доминирующие психологические характеристики.

#### И.4.3 Интерпретация результатов

Интерфейс интерпретации предназначен для объяснения предсказаний мультимодальной модели за счёт визуализации вклада признаков и механизмов внимания.

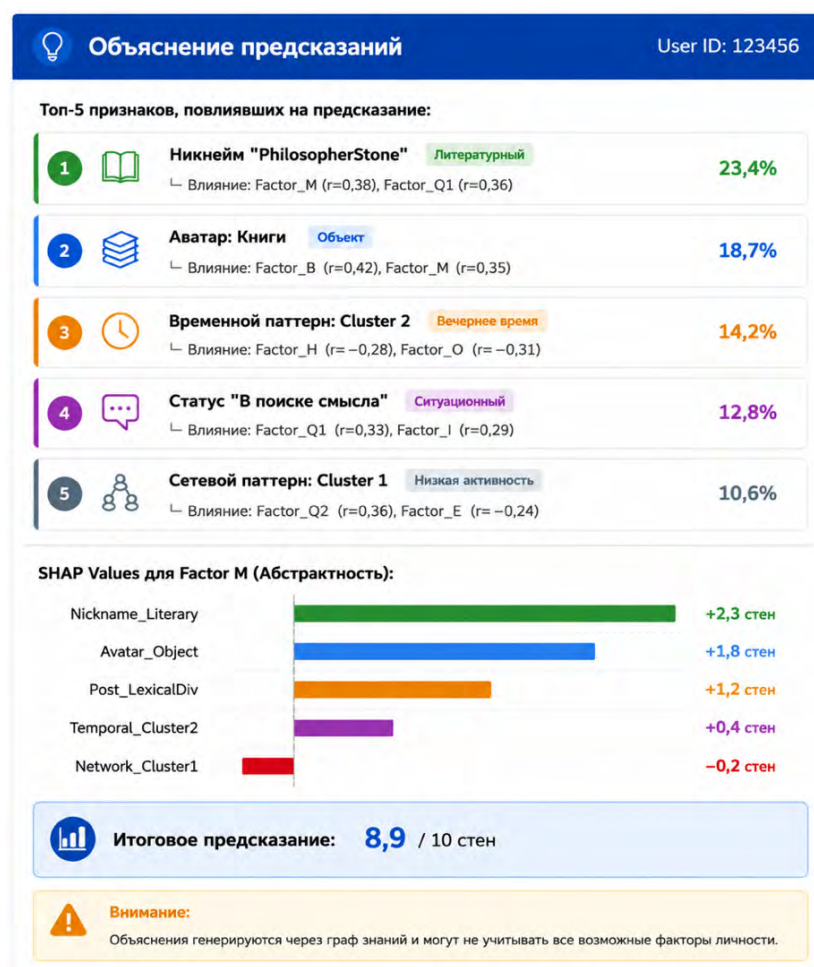


Рисунок И.6 – Визуализация SHAP и путей внимания в графе знаний

Результаты интерпретации отображаются для каждого психологического фактора и включают как количественную оценку вклада модальностей, так и структурное объяснение через граф знаний.

Функциональные возможности:

- вычисление и отображение SHAP для оценки вклада признаков в предсказание каждого психологического фактора;

- визуализация весов внимания, полученных из GAT, в виде путей распространения влияния: User → Feature → PsychConstruct;
- агрегированная оценка важности признаков на основе комбинации SHAP и механизма внимания;
- интерактивная визуализация графа знаний (D3.js) для исследования связей между пользователем, признаками и психологическими факторами [234].

SHAP отражают количественный вклад каждой модальности в итоговое предсказание, тогда как механизм внимания выявляет структурные зависимости между элементами графа знаний.

Совместное использование этих методов формирует более полное объяснение работы модели: SHAP даёт оценку «насколько важно», а механизм внимания – «через какие связи это происходит».

Данный компонент обеспечивает прозрачность и интерпретируемость мультимодальной модели, что особенно важно при анализе психологических характеристик и повышает доверие пользователя к результатам системы.

#### И.4.4 Визуализация графа знаний

Интерфейс визуализации графа знаний предназначен для отображения структурных связей между пользователем, извлечёнными признаками и психологическими факторами. Визуализация реализована с использованием библиотеки D3.js в виде силовой модели графа, обеспечивающей наглядное представление мультимодальных зависимостей.

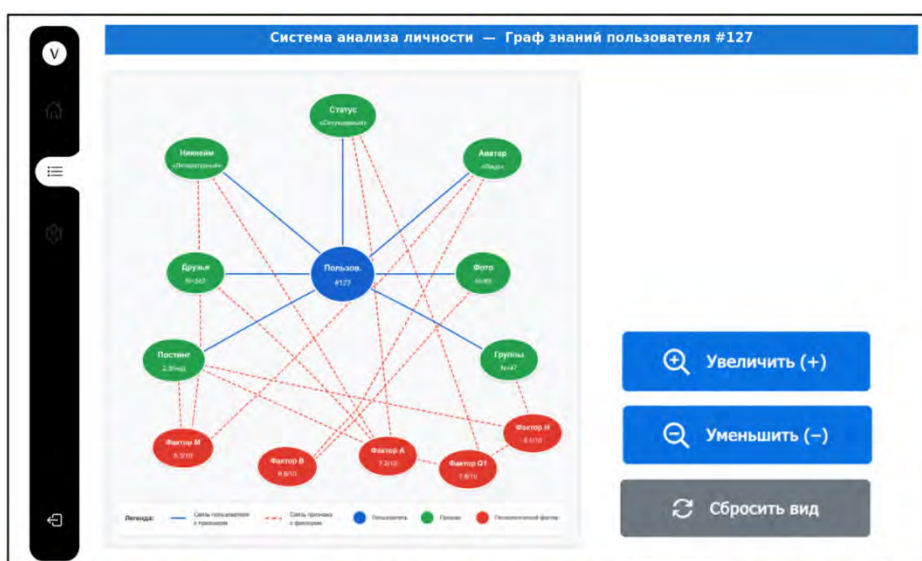


Рисунок И.7 – Интерактивное отображение связей между пользователем, признаками и психологическими факторами



Центральным элементом является узел пользователя, от которого исходят связи к узлам признаков (Feature) и далее к психологическим конструкциям (PsychConstruct). Цветовая кодировка узлов отражает их тип и семантическую роль в модели.

Интерфейс предоставляет инструменты управления отображением графа (масштабирование, сброс вида), а также поддерживает интерактивное исследование структуры связей.

Интерактивные возможности:

- перемещение узлов для локальной реорганизации структуры графа;
- отображение дополнительной информации при наведении: свойства узла и его связи;
- выбор узла с фокусировкой на локальном подграфе (соседи на расстоянии 1–2 hop);
- масштабирование;
- автоматическая укладка графа с использованием алгоритма силового размещения графа (d3-force).

Параметры визуализации: коэффициент отталкивания:  $\text{charge} = -300$ , расстояние между связанными узлами:  $\text{linkDistance} = 100$

## И.5 Диаграммы компонентов и взаимодействий

### И.5.1 Диаграмма последовательности: Анализ профиля

Диаграмма последовательности иллюстрирует процесс анализа профиля пользователя в системе, включая взаимодействие между клиентским интерфейсом, серверной частью и компонентами обработки данных. Процесс реализован в асинхронном режиме и включает этапы извлечения, обработки и интерпретации мультимодальных данных.

Основные этапы обработки:

- Асинхронный сбор данных пользователя через VK API с учётом ограничений: до 20 запросов/с.
- Предобработка данных, включая детекцию лиц на изображениях с использованием модели YOLOv5x.
- Параллельное извлечение признаков из различных модальностей (текст, изображения, поведенческие данные).
- Формирование и обновление графа знаний в базе Neo4j.

- Выполнение инференса графовой нейронной сети (GAT) на подграфе пользователя.

- Генерация объяснений предсказаний.
- Кэширование промежуточных и итоговых результатов.
- Сохранение сформированного отчёта в базе данных.
- Возврат результата клиентскому приложению в формате JSON.

Для графовой нейронной сети (GAT) вычислительная сложность определяется структурой подграфа пользователя и составляет:

$$O(|V| + |E|),$$

где  $|V|$  – количество узлов,  $|E|$  – количество рёбер подграфа.

В среднем для одного пользователя:  $|V| \approx 150$  узлов,  $|E| \approx 500$  рёбер.

Общее время обработки одного профиля составляет 3–4 секунды, что соответствует требованиям интерактивного анализа.

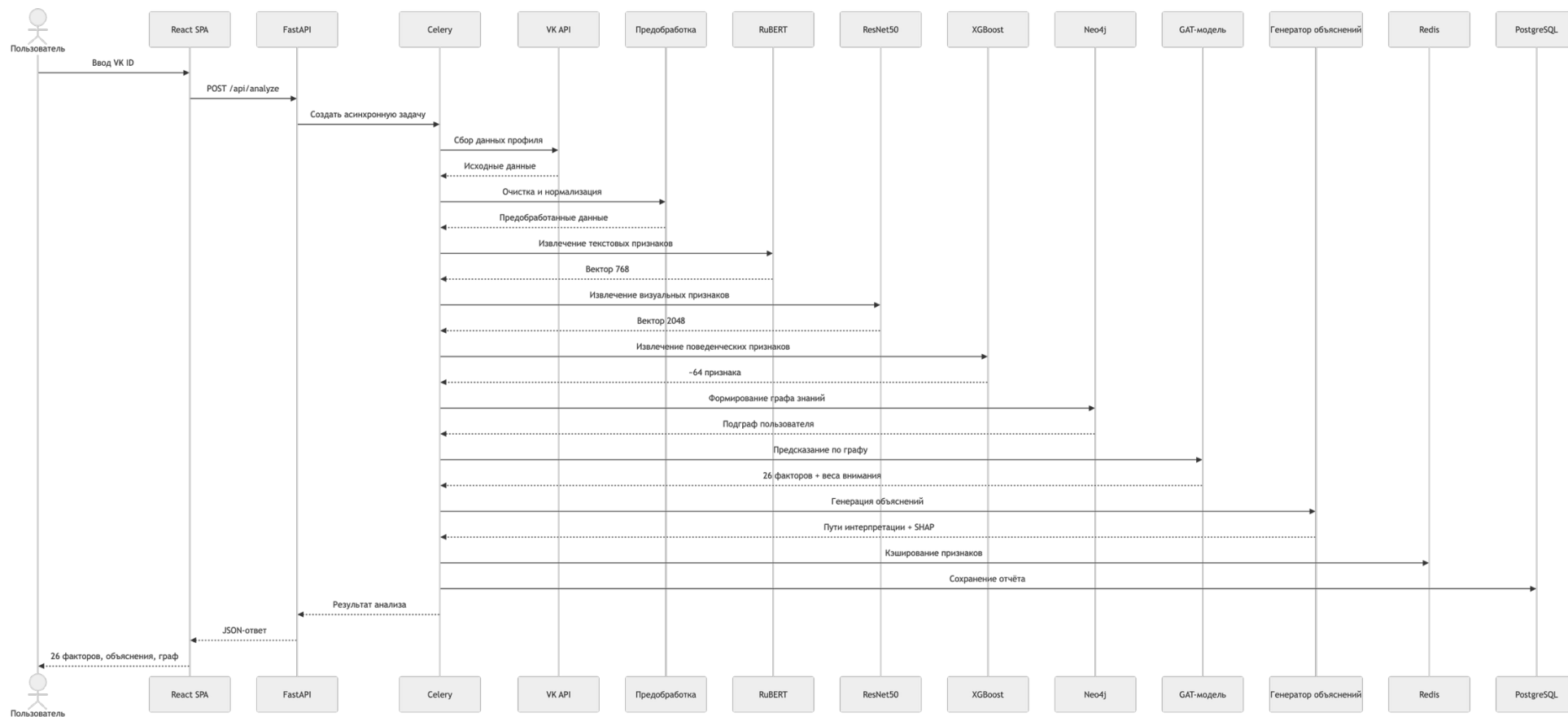


Рисунок И.8 – Диаграмма последовательности процесса анализа профиля пользователя в системе

## И.5.2 Компонентная диаграмма: ML-конвейер

Компонентная диаграмма ML-конвейер отражает структуру и взаимодействие моделей машинного обучения, используемых для извлечения мультимодальных признаков и предсказания психологических характеристик пользователя.

Конвейер включает специализированные модели для обработки текстовых, визуальных и поведенческих данных, а также интеграционный модуль на основе графовой нейронной сети.

Основные компоненты и модели:

- Текстовая модальность: RoBERTa-base используется для кодирования никнеймов и статусов пользователей, формируя эмбединги размерности 768. Модель содержит порядка 125 млн параметров и обеспечивает извлечение семантических и стилистических признаков текста.

- Визуальная модальность: ResNet50 применяется для извлечения признаков из изображений (аватары), формируя эмбединги размерности 2048. Модель содержит около 25 млн параметров и используется для задач определения возраста и типа аватара. Для определения пола используется модель VGG16.

- Данные цифровой активности: Модель XGBoost применяется для анализа признаков цифровой активности (временные паттерны, сетевые характеристики). Используется ансамбль из 100 деревьев с максимальной глубиной 6 и скоростью обучения 0,1.

- Кластеризация признаков выполняется с использованием алгоритма k-means:  $k = 5$  для временных паттернов,  $k = 4$  для сетевых характеристик.

- Интеграция модальностей: GAT объединяет признаки всех модальностей в рамках графа знаний. Архитектура включает 3 слоя с 8 головами внимания и скрытым представлением размерности 512. Общее число параметров составляет около 3,5 млн.

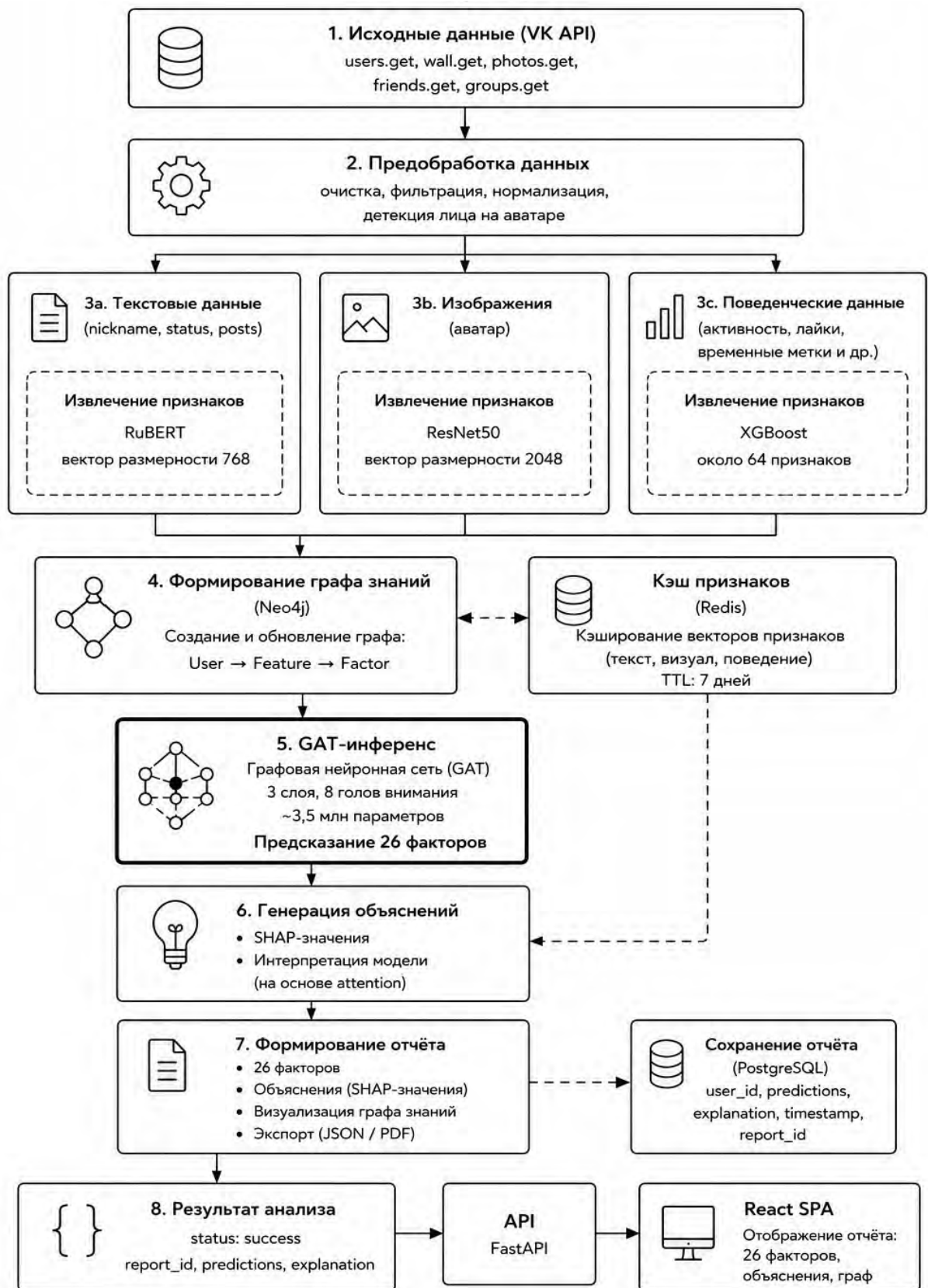


Рисунок И.9 – Компонентная диаграмма ML-конвейер мультимодальной системы анализа личности

Обучение и инференс моделей выполняются на следующей вычислительной инфраструктуре:

- GPU: NVIDIA RTX 3090 (24 GB VRAM);
- CPU: 16 ядер (Intel Xeon);
- оперативная память: 64 GB;
- дисковое хранилище: 2 TB NVMe SSD (для хранения моделей и кэша эмбедингов).

## И.6 Результаты валидации и апробации

### И.6.1 Апробация системы на задаче анализа цифровых профилей пользователей

Целью эксперимента являлась оценка применимости модели для предварительного анализа личностных характеристик соискателей на основе цифрового следа.

Эксперимент проведён на выборке из 50 кандидатов, для которых были доступны публичные профили социальной сети. Для каждого пользователя система выполняла полный цикл обработки: сбор данных, извлечение признаков, построение подграфа и предсказание психологических факторов.

Результаты анализа показали, что модель демонстрирует устойчивую работу в условиях реальных данных, однако ряд факторов приводит к снижению точности предсказаний.

Выявленные источники ошибок можно разделить на несколько категорий.

- Закрытые профили ( $\approx 9\%$  случаев). При отсутствии доступа к данным пользователя (закрытые аккаунты) система ограничена минимальным набором признаков. В этом случае используется fallback-режим на основе поведенческих признаков (модель XGBoost). Заметно сильное снижение качества предсказаний.
- Минимально заполненные профили ( $\approx 5\%$ ). Профили с недостаточным количеством информации (менее 10 постов, отсутствие аватара) существенно ограничивают информативность входных данных. В таких случаях средняя ошибка возрастает до  $MAE \approx 4,2$ , что значительно выше среднего значения по выборке.
- Эффект социальной желательности ( $\approx 12\%$ ). Часть пользователей демонстрирует в цифровом профиле поведение, отличающееся от реального. В этом случае происходит расхождение между предсказанными характеристиками и экспертной оценкой.
- Культурно-языковая специфика ( $\approx 7\%$ ). Наличие иностранных никнеймов и публикаций на других языках снижает точность текстовой модальности. Несмотря на использование моделей семейства BERT (RuBERT и RoBERTa), обученных

преимущественно на русскоязычных и близких по распределению данных, качество извлечения признаков ухудшается при обработке мультязычного контента.

Таким образом, основными источниками ошибок являются ограничения входных данных и доменные особенности пользователей, а не архитектурные недостатки модели.

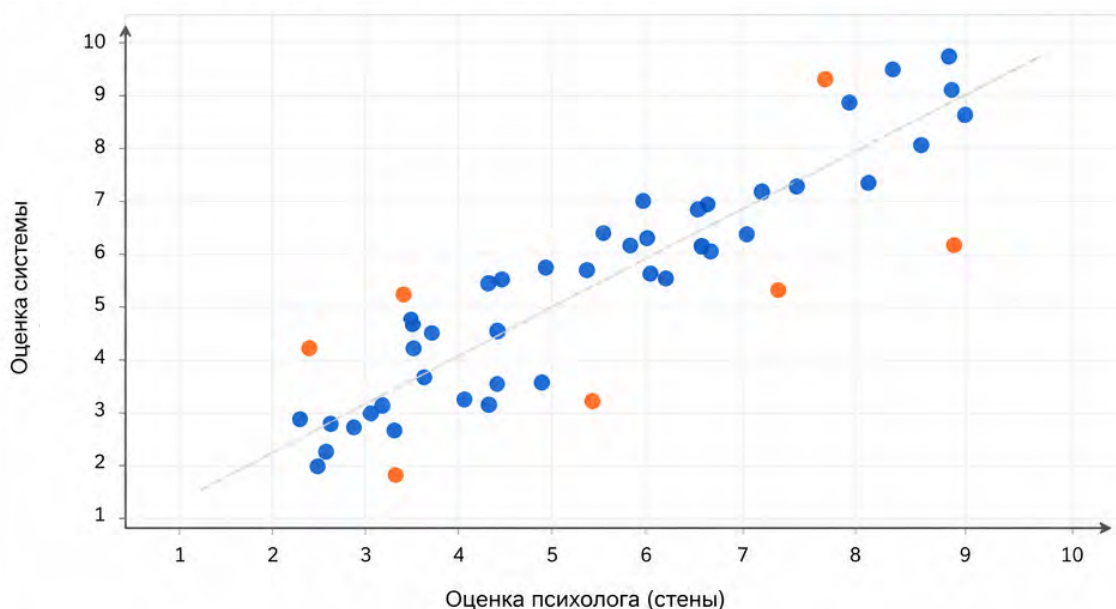


Рисунок И.10 – Диаграмма рассеяния оценок психолога-эксперта и предсказаний модели (N = 50)

Сравнение предсказаний модели с экспертной оценкой психолога проводилось на выборке из 50 пользователей. Средняя корреляция Пирсона по всем факторам составила:  $r = 0,68$  – высокая согласованность результатов модели с экспертной оценкой.

Средняя ошибка предсказания составила:

- MAE = 2,9 стенов,
- RMSE = 3,7 стенов.

Для дополнительной оценки была проведена бинарная классификация (порог  $\geq 7$  стенов):

- Accuracy = 78%,
- Cohen's  $\kappa = 0,61$  (существенная согласованность).

Ошибки распределились следующим образом:

- False Positives: 17%,
- False Negatives: 20%.

Система показывает значительное преимущество по времени анализа: эксперт:  $\approx 25$  часов, система:  $\approx 2$  часа.

Таким образом, использование системы позволяет сократить временные затраты на 92%.

#### И.7 Примечания к визуализациям

В данном параграфе приведена информация об используемых инструментах визуализации, форматах представления данных и обеспечении воспроизводимости результатов.

Для построения визуализаций применялись следующие программные средства:

- архитектурные диаграммы: draw.io, Lucidchart;
- UML-диаграммы последовательностей: PlantUML [237], Visual Paradigm;
- интерактивные визуализации: React и библиотека D3.js (версия 7);
- статистические графики и тепловые карты: Matplotlib и Seaborn;
- визуализация графов знаний: Neo4j Browser и Gephi.

Визуализации представлены в следующих форматах:

- векторные изображения (SVG), обеспечивающие масштабируемость без потери качества;
- растровые изображения (PNG, 300 DPI), используемые для печатной версии работы;
- интерактивные визуализации (HTML/JavaScript), применяемые в веб-интерфейсе системы.

Все представленные визуализации формируются автоматически на основе выходных данных системы анализа цифрового следа.

Используемые подходы позволяют воспроизводить результаты при повторной обработке входных данных, отвечая требованиям экспериментальной проверяемости.

Визуализации используются как средство интерпретации результатов и не влияют на процесс обучения моделей.



## ПРИЛОЖЕНИЕ К Кросс-доменный перенос метода: дерматоскопическая диагностика новообразований кожи

В настоящем приложении приведены результаты экспериментальной проверки переносимости предложенных в диссертации онтологического паттерна проектирования «объект анализа – признак – целевая характеристика» и метрики интерпретируемости путей в графе GPI (Graph Path Interpretability) на иную предметную область – дерматоскопическую диагностику новообразований кожи. Перенос на область, не связанную с исходной задачей профилирования пользователей социальных сетей, подтверждает обобщаемость разработанных модели, метода и метрики интерпретируемости.

### К.1 Постановка задачи переноса

Цель эксперимента: показать, что инвариантное ядро онтологического паттерна (классы концептов, предикаты, правила инстанцирования онтологии в граф знаний и подключения графовой нейронной сети) и определённая над ним метрика GPI применимы без изменения структуры метода при смене доменно-специфичной онтологии признаков. Исходная прикладная задача (предсказание психологических характеристик по цифровому следу) и целевая задача (классификация дерматоскопических изображений) различаются модальностью данных, природой признаков и целевых характеристик, что делает их пригодными для проверки доменной независимости.

Реализация паттерна в дерматоскопии:

- объект анализа – дерматоскопическое изображение пигментного поражения кожи;
- признак – интерпретируемые дерматоскопические признаки правила ABCD (асимметрия, нерегулярность границы, вариативность цвета, диаметр) и структурная гетерогенность;
- целевая характеристика – диагностическая категория новообразования (семь классов).

### К.2 Набор данных и опорная сеть классификации

Использован набор HAM10000 [4]: 10 015 дерматоскопических изображений семи классов: меланоцитарные невусы (nv), меланома (mel), доброкачественные кератозоподобные поражения (bkl), базальноклеточная карцинома (bcc), актинические кератозы / внутриэпителиальная карцинома (akiec), сосудистые поражения (vasc) и дерматофиброма (df). Во избежание утечки данные разбиты по

идентификатору поражения (lesion\_id) на обучающую, валидационную и тестовую части (7008/1515/1492 изображений); зерно генератора зафиксировано (seed = 42).

Опорная сеть EfficientNet-B0, предобученная на ImageNet, дообучена на семь классов (вход 224×224, обратно-частотное взвешивание классов, оптимизатор AdamW, 30 эпох). На тесте достигнуты точность 0,807, сбалансированная точность 0,677, макро-F1 0,668 и макро-AUC 0,955 (рисунок К.1), что соответствует уровню одно-сетевых моделей на HAM10000 и служит базовым классификатором для оценки интерпретируемости.

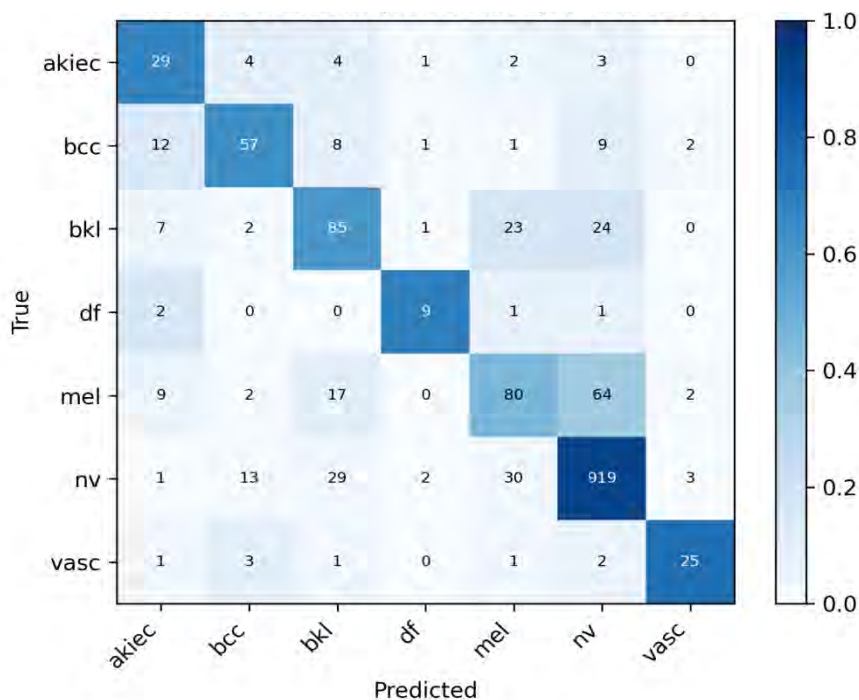


Рисунок К.1 – Нормированная матрица ошибок на тесте HAM10000 (n = 1492)

### К.3 Реализация онтологического паттерна: дерматоскопические признаки

Для каждого изображения вычислены пять интерпретируемых признаков, заземлённых в правиле ABCD, нормированных к  $[0, 1]$  и зависящих только от изображения и маски поражения (но не от модели, что сохраняет модельную независимость метрики):

- Асимметрия (A) – несовпадение маски поражения с её отражениями относительно главных осей через центрoид;
- Граница (B) – нерегулярность границы как  $1 - \text{круговость}$  ( $\text{круговость} = 4\pi \cdot \text{площадь} / \text{периметр}^2$ );
- Цвет (C) – вариативность цвета как нормированное число различных цветовых кластеров (k-means в пространстве CIE-Lab);

- Диаметр (D) – эквивалентный диаметр поражения относительно стороны кадра;
- Структура (S) – структурная гетерогенность как нормированная энтропия Шеннона гистограммы яркости поражения.

Маски поражений взяты из набора сегментаций HAM10000, при их отсутствии получены пороговой обработкой Оцу. Эти пять признаков образуют доменно-специфичную онтологию признаков, подставляемую в инвариантное ядро паттерна вместо текстовых, визуальных и поведенческих признаков исходной задачи.

#### К.4 Граф знаний «признак – диагноз»

Реализация графа знаний выполнена по правилам паттерна: построен двудольный граф, рёбра которого связывают признак с диагнозом  $k$  и кодируют силу и значимость их ассоциации, оценённые только на обучающей выборке. Вес ребра – точно-бисериальная корреляция  $r$  между значением признака и индикатором класса (один против всех);  $p$ -значения 35 пар скорректированы на множественность процедурой Бенджамини–Хохберга, ребро значимо при  $q < 0,05$  и  $r > 0$ . Сохранены только положительные ассоциации (пусть GPI кодирует присутствие признака, поддерживающего диагноз), что соответствует определению объясняющего ребра ASSOCIATED\_WITH в исходной формализации метрики.

Из 35 пар значимыми оказались 25 (11 положительных рёбер графа и 14 отрицательных, исключаемых правилом положительного ребра). Полная матрица корреляций приведена в таблице К.1, граф из 11 значимых рёбер – на рисунке К.2. Структура клинически непротиворечива: структурная гетерогенность сильнее всего связана с меланомой ( $r = 0,22$ ), к меланоме также ведут рёбра границы, диаметра и асимметрии (диагноз-хаб с четырьмя входящими рёбрами).

Таблица К.1 – Точечно-бисериальная корреляция  $r$  «признак – диагноз» (один против всех) на обучающей выборке HAM10000 ( $n = 7008$ )

| Признак    | akiec | bcc   | bkl   | df    | mel  | nv    | vasc  |
|------------|-------|-------|-------|-------|------|-------|-------|
| Асимметрия | 0,02  | –0,01 | 0,00  | –0,06 | 0,03 | –0,01 | –0,04 |
| Граница    | 0,12  | –0,01 | 0,15  | –0,06 | 0,13 | –0,19 | –0,10 |
| Цвет       | –0,05 | –0,19 | –0,06 | –0,01 | 0,02 | 0,14  | –0,01 |
| Диаметр    | 0,13  | –0,01 | 0,10  | 0,03  | 0,10 | –0,17 | –0,09 |
| Структура  | –0,09 | –0,13 | –0,14 | –0,10 | 0,22 | 0,07  | –0,01 |



Рисунок К.2 – Граф знаний из 11 значимых положительных ассоциаций  
«признак – диагноз»

#### К.5 Рассуждающие модули и применение метрики GPI

Над вектором из пяти признаков (концептуальное «бутылочное горлышко», при котором диагноз получается только из признаков) обучены три рассуждающих модуля, различающиеся лишь способом агрегации: GAT (графовое внимание, веса признаков  $w_i$ ), GCN (равномерное среднее) и MLP («чёрный ящик» без путей). Схема удерживает все модули на одной диагностической информации, поэтому различие интерпретируемости не сводится к точности.

Кандидат на объяснение предсказания  $n$  – путь из двух рёбер: изображение  $\rightarrow$  (HAS\_FEATURE)  $\rightarrow$  признак  $i$   $\rightarrow$  (ASSOCIATED\_WITH)  $\rightarrow$  предсказанный класс. Первое ребро требует присутствия признака ( $c_i > 0,2$ ), второе – значимости ассоциации ( $q < 0,05$ ). Метрика GPI применена в виде системы (формальное определение – Приложение П):

- GPI\_cov (покрытие путями) – доля предсказаний, для которых найден хотя бы один значимый путь;
- GPI\_w (взвешенная согласованность путей) – доля веса решения модели, приходящаяся на клинически валидные пути;
- проверка верности методом удаления концептов (обнуление признаков в порядке убывания важности).

## К.6 Результаты

Покрытие и сравнение архитектур. При равной точности ( $\approx 0,67$ ) все модули показали  $GPI\_cov \approx 1,0$ : граф признаков ABCD допускает значимый объясняющий путь почти для каждого решения. Основное различие – по взвешенной согласованности: графовые модули направляют существенную долю веса решения через клинически валидные пути ( $GPI\_w = 0,416$  для GAT,  $0,559$  для GCN), тогда как «чёрный ящик» – нет ( $GPI\_w = 0,000$ ). Результаты сведены в таблицу К.2 и на рисунке К.3.

Таблица К.2 – Точность и GPI по рассуждающим модулям (полный НАМ10000)

| Модуль | Тестовая точность | $GPI\_cov$ | $GPI\_w$ | AUC удаления     |
|--------|-------------------|------------|----------|------------------|
| GAT    | 0,670             | 1,000      | 0,416    | 0,885            |
| GCN    | 0,668             | 1,000      | 0,559    | 0,953            |
| MLP    | 0,676             | 0,999      | 0,000    | 0,691 (контроль) |

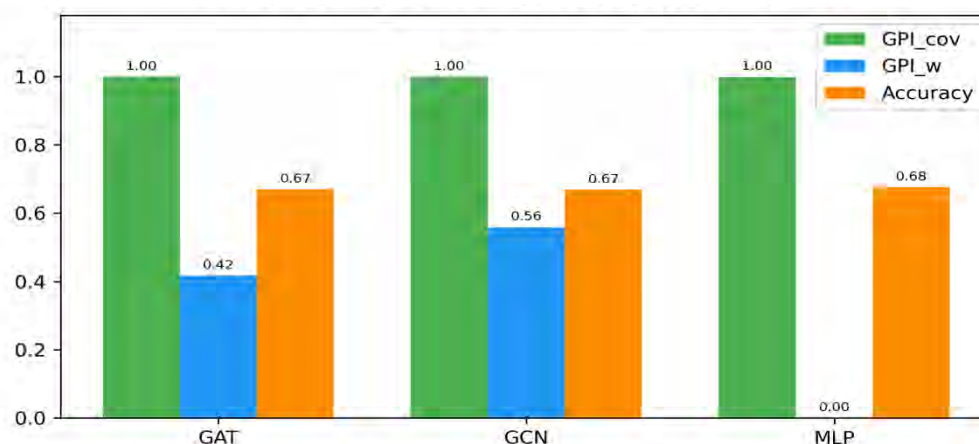


Рисунок К.3 – Сопоставление GPI и точности по рассуждающим модулям  
Кривые удаления концептов приведены на рисунке К.4.

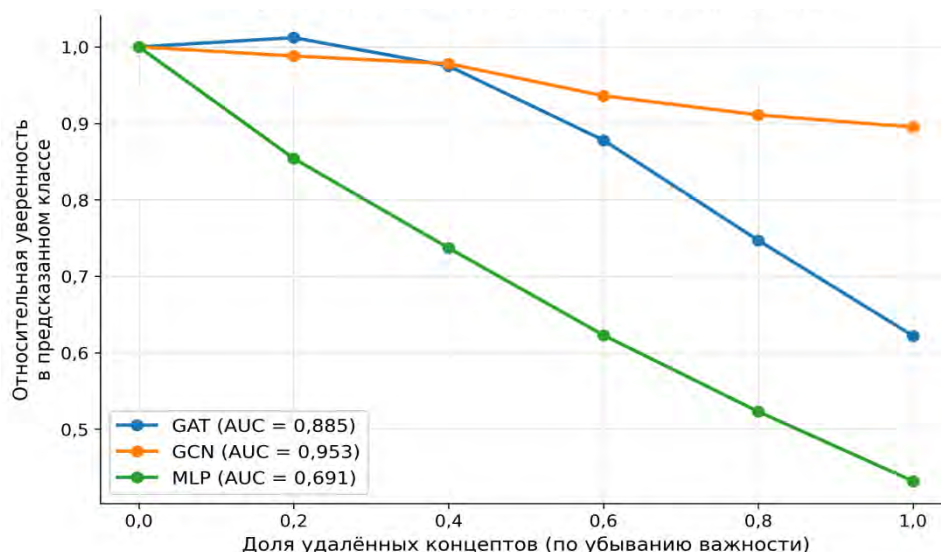


Рисунок К.4 – Кривые верности при удалении концептов

Видно, что выдаваемые модулем веса действительно определяют решения: удаление признаков в порядке важности по GAT снижает уверенность круче ( $AUC = 0,885$ ), чем по равномерному GCN ( $0,953$ ). Пример объясняющего пути «изображение  $\rightarrow$  признак  $\rightarrow$  диагноз» с весами приведён на рисунке К.5.



Рисунок К.5 – Пример объясняющего пути от изображения к признаку и диагнозу

Для проверки обобщения за пределами одного архива модель, обученная на HAM10000, оценена на независимом наборе ISIC 2018 Task 3 (1511 изображений из других источников, те же семь классов). Переносимые компоненты сохранили показатели: опорный классификатор – точность  $0,784$ , макро- $AUC$   $0,939$  (рисунок К.6); граф знаний воспроизвёл те же значимые рёбра ABCD (13 рёбер; структура – меланома,  $r = 0,22$ ); основная метрика GPI\_cov осталась  $\approx 1,0$ .

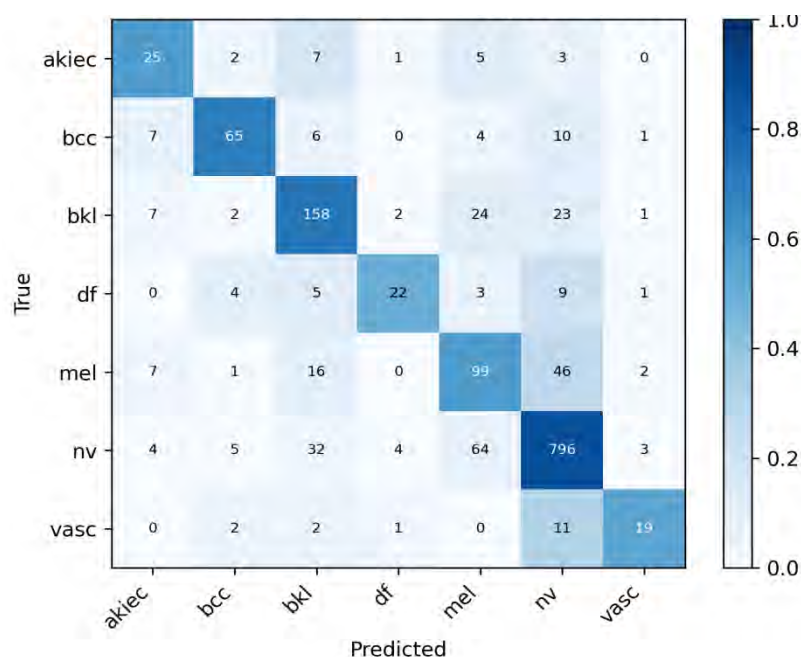


Рисунок К.6 – Матрица ошибок классификатора (обучен на HAM10000) на независимом наборе ISIC 2018 ( $n = 1511$ )

## К.7 Выводы по приложению

Эксперимент подтверждает переносимость разработанного метода между предметными областями: инвариантное ядро онтологического паттерна и метрика интерпретируемости путей в графе применены к дерматоскопической диагностике без изменения структуры метода – заменена лишь доменно-специфичная онтология

признаков (текстовые/визуальные/поведенческие → дерматоскопические ABCD) и целевые характеристики (психологические факторы → диагностические классы). Граф признаков объясняет почти все решения классификатора ( $GPI_{cov} \approx 1,0$ ), метрика GPI при равной точности отделяет графовое рассуждение с концептами от «чёрного ящика», а результаты воспроизводятся на независимом наборе ISIC 2018. Тем самым подтверждается обобщаемость предложенных онтологического паттерна и метрики GPI за пределами исходной прикладной задачи. В сценарии поддержки принятия решения врач-дерматолог (лицо, принимающее решение) использует прогноз диагноза, объясняющие пути в графе знаний (значимые визуальные признаки) и оценку их качества метрикой GPI для подтверждения либо корректировки диагностического решения, что недоступно при использовании модели типа «чёрный ящик».



## ПРИЛОЖЕНИЕ Л Кросс-доменная проверка метода на задачах интерпретируемого анализа текста

Настоящее приложение посвящено экспериментальной проверке переносимости предложенных онтологического паттерна и метрики интерпретируемости путей в графе GPI (Graph Path Interpretability) на новые предметные области анализа текста. Основная проверка выполнена на задаче интерпретируемой классификации токсичных комментариев, а дополнительная – на задаче эмоциональной классификации, требующей использования полного онтологического паттерна с предикатом `коррелирует_с`.

Целью экспериментов является подтверждение того, что инвариантное ядро метода сохраняет свою структуру при переходе к различным предметным областям, а изменяется только доменная онтология признаков.

В качестве основной области проверки выбрана классификация токсичных комментариев как один из наиболее распространённых бенчмарков анализа текста, для которого принципиальное значение имеет прозрачность процесса принятия решения «токсично / не токсично». Дополнительным преимуществом данной задачи является наличие набора данных, содержащего вместе с разметкой классов выполненные человеком обоснования решений (*rationales*), который дает возможность, наряду с оценкой качества классификации, объективно оценить согласованность автоматически формируемых объяснений с человеческими обоснованиями.

Дополнительная проверка выполнена на задаче эмоциональной классификации русскоязычных текстов, где одновременно прогнозируются несколько взаимосвязанных целевых характеристик. При её выполнении используется полный онтологический паттерн, включающий предикат `коррелирует_с`, что подтверждает его применимость в задачах с несколькими целевыми характеристиками.

### Л.1 Постановка задачи переноса

Целью эксперимента является экспериментальное подтверждение того, что инвариантное ядро разработанного метода, включающее онтологический паттерн «объект анализа – признак – целевая характеристика», граф знаний, графовую сеть внимания и метрику интерпретируемости GPI, может применяться без изменения архитектуры при переходе от психологического профилирования и дерматоскопии к задаче интерпретируемой классификации токсичных комментариев.



При переходе к новой предметной области сохраняются структура онтологического паттерна, механизм построения графа знаний, архитектура графовой нейронной сети и метрика интерпретируемости GPI. Изменяется только предметно-ориентированная онтология признаков, соответствующая новой задаче анализа.

Реализация онтологического паттерна имеет следующий вид:

- Объект анализа – комментарий (текст).
- Признаки – интерпретируемые лексические признаки токсичности: обценная лексика, оскорбление, угроза, упоминание защищённой группы, адресность (обращение во втором лице), токсичные термины.
- Целевая характеристика – бинарная токсичность (toxic / non-toxic).

## Л.2 Набор данных

Использован набор данных HateXplain (2021) [238], содержащий 20 148 комментариев из Twitter и Gab с разметкой по трём классам (hate speech, offensive, normal), агрегированным в бинарную метку токсичности по результатам мажоритарного голосования трёх аннотаторов. Принципиальной особенностью набора является наличие эталонных обоснований (rationales): для токсичных комментариев аннотаторы выделили на уровне токенов слова, послужившие основанием для принятия решения. Наличие эталонных обоснований создаёт возможность объективной проверки соответствия автоматически формируемых объяснений человеческим обоснованиям решений.

Набор разделён на обучающую выборку объёмом 15 383 комментария и тестовую выборку объёмом 1 924 комментария; доля токсичных комментариев в тестовой выборке составляет 0,59.

HateXplain выбран в качестве основного набора данных, поскольку является единственным широко используемым бенчмарком классификации токсичности, содержащим человеческие обоснования на уровне токенов. В отличие от него, русскоязычные наборы данных ((Russian Language Toxic Comments) [239] и другие) содержат только метки классов и не позволяют объективно оценить соответствие автоматически формируемых объяснений человеческим обоснованиям. Дополнительно использование англоязычного набора данных для проверки метода, разработанного на русскоязычном материале, позволяет подтвердить его языковую независимость.

### Л.3 Реализация паттерна

Компонент «Признак» онтологического паттерна реализован в виде шести интерпретируемых категорий, используемых при оценке токсичности комментариев. Выбор признаков основан на общепринятой таксономии токсичности, включающей основные подтипы набора Jigsaw Toxic Comment (оскорбление, обценная лексика, угроза, атака на идентичность), и дополнен прагматическим признаком адресности.

Таким образом, признаки выполняют ту же функцию, что и дескрипторы ABCD в дерматоскопической предметной области (Приложение К): формируют интерпретируемый словарь предметной области, используемый при построении графа знаний и формировании объяснений.

Каждый признак вычисляется непосредственно по тексту комментария, а не по внутренним представлениям модели классификации. Признаки представлены в бинарном виде (1 – признак присутствует в комментарии, 0 – отсутствует). Для каждого сработавшего признака дополнительно фиксируются позиции соответствующих токенов, используемые при сопоставлении автоматически сформированных объяснений с человеческими обоснованиями.

Определение каждого признака и источник его формирования приведены в таблице Л.1.

Таблица Л.1 – Интерпретируемые признаки токсичности: определение и источник

| Признак               | Определение (срабатывает, если в комментарии присутствует...)                                                                           | Источник                                                                                                                                           |
|-----------------------|-----------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| Оскорбление           | уничижительная лексика (idiot, stupid, loser, trash и др.)                                                                              | лексикон оскорблений                                                                                                                               |
| Обценная лексика      | обценная (бранная) лексика                                                                                                              | лексикон обценной лексики                                                                                                                          |
| Угроза                | лексика, выражающая угрозу (kill, hurt, destroy, attack и др.)                                                                          | лексикон угроз                                                                                                                                     |
| Упоминание группы     | нейтральные наименования защищённых социальных групп (muslim, women, black и др.)                                                       | список названий групп (используется для анализа возможного смещения модели)                                                                        |
| Адресность (2-е лицо) | прямое обращение к адресату посредством местоимений второго лица (you, your); в английском языке обращение выражается единой формой you | правило (закрытый список местоимений)                                                                                                              |
| Токсичный термин      | слово, входящее в автоматически сформированный словарь токсичной лексики                                                                | сформирован по обучающей выборке на основе топ-500 токенов, наиболее часто отмечавшихся аннотаторами HateXplain в качестве обоснования токсичности |

В эксперименте использованы компактные встроенные лексиконы. В прикладных системах они могут быть заменены стандартными многоязычными ресурсами (например, HurtLex, Hatebase) без изменения структуры метода.

Словарь токсичных терминов формируется исключительно по обучающей выборке на основе человеческих обоснований и затем применяется к тестовой выборке. Таким образом, проверка согласованности автоматически сформированных объяснений с человеческими обоснованиями выполняется без использования информации из тестовой разметки.

Таким образом, указанные шесть признаков образуют доменную онтологию признаков, используемую вместо онтологии психологической и дерматоскопической предметных областей при сохранении неизменной структуры онтологического паттерна, графа знаний, графовой нейронной сети и метрики интерпретируемости GPI.

#### Л.4 Формирование графа знаний

Граф знаний формируется в соответствии с предложенным онтологическим паттерном. Рёбра типа «ассоциируется\_с» связывают узлы признаков с целевой характеристикой «Токсичность». Вес каждого ребра определяется значением точечно-бисериальной корреляции  $r$ , вычисленной только по обучающей выборке. Статистическая значимость ассоциаций оценивается с использованием поправки Бенджамини–Хохберга. В граф знаний включаются только статистически значимые положительные ассоциации ( $q < 0,05$ ;  $r > 0$ ). Отрицательные ассоциации не используются при построении графа объяснений, поскольку они не являются подтверждающими признаками целевой характеристики и не могут служить основанием для вывода о токсичности.

Результаты расчёта приведены в таблице Л.2.

Таблица Л.2 – Ассоциация признаков с токсичностью

| Признак               | $r$    | $q$                  |
|-----------------------|--------|----------------------|
| Оскорбление           | +0,110 | $5,9 \cdot 10^{-42}$ |
| Обценная лексика      | +0,087 | $6,9 \cdot 10^{-27}$ |
| Токсичный термин      | +0,075 | $3,1 \cdot 10^{-20}$ |
| Адресность (2-е лицо) | +0,040 | $9,2 \cdot 10^{-7}$  |
| Угроза                | +0,015 | $6,4 \cdot 10^{-2}$  |
| Упоминание группы     | -0,137 | $1,0 \cdot 10^{-64}$ |

В соответствии с указанным правилом в граф знаний включаются признаки «Оскорбление», «Обценная лексика», «Токсичный термин» и «Адресность (2-е лицо)», имеющие статистически значимую положительную ассоциацию с

токсичностью. Признак «Угроза» не включается в граф вследствие отсутствия статистической значимости ( $q \geq 0,05$ ), а признак «Упоминание группы» исключается, поскольку характеризуется статистически значимой отрицательной ассоциацией ( $r < 0$ ).

Структура сформированного графа знаний представлена на рисунке Л.1.

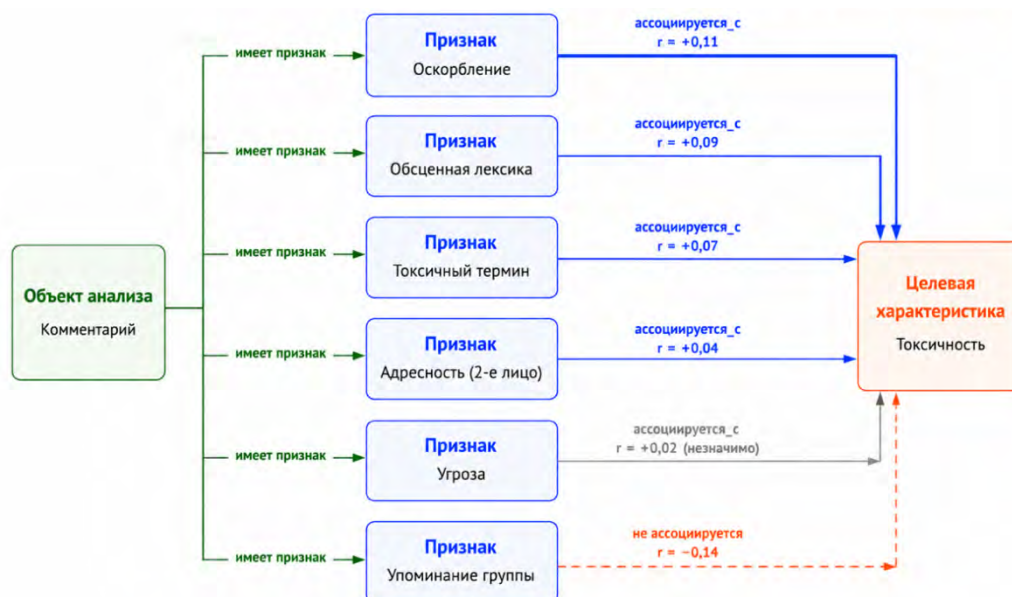


Рисунок Л.1 – Граф знаний, реализующий онтологический паттерн для задачи классификации токсичных комментариев

Наиболее сильные положительные связи наблюдаются для признаков «Оскорбление», «Обценная лексика» и «Токсичный термин», тогда как признак «Адресность (2-е лицо)» характеризуется более слабой, но статистически значимой ассоциацией. Отрицательная связь признака «Упоминание группы» не используется при формировании графа объяснений в соответствии с принятым правилом построения.

#### Л.5 Проверка метода на различных моделях вывода

Для проверки переносимости предложенного метода над шестимерным вектором интерпретируемых признаков обучены три модели: графовая сеть внимания (GAT), графовая сверточная сеть (GCN) и многослойный перцептрон (MLP). Во всех случаях входом служит один и тот же вектор признаков, поэтому различия в результатах обусловлены исключительно особенностями используемых моделей вывода.

Во всех моделях единственным входом служит шестимерный вектор интерпретируемых признаков. Поэтому итоговое решение формируется исключительно на основе признаков предметной области, без использования

скрытых представлений исходного текста. В качестве базовой модели использован классификатор TF-IDF и логистическая регрессия, обученный на полном тексте комментариев. Он достигает точности 0,77 и F1-меры 0,81 и используется как ориентир максимально достижимого качества без ограничений на интерпретируемость.

Для оценки интерпретируемости предложенного метода использована метрика GPI, включающая два взаимодополняющих компонента:

- GPI\_cov – доля объектов, для которых построен хотя бы один объясняющий путь;
- GPI\_w – взвешенная согласованность объясняющих путей с вкладом используемых признаков.

Дополнительно проведена проверка корректности объяснений путём последовательного удаления концептов графа знаний.

Для дополнительной оценки качества формируемых объяснений вычисляются метрики token-F1 и AUPRC, характеризующие степень их соответствия человеческим обоснованиям. Для этого каждому токenu присваивается оценка важности на основе весов внимания значимых признаков, после чего полученная карта важности сравнивается с разметкой токенов, выполненной аннотаторами.

## Л.6 Результаты

Все три модели демонстрируют близкое качество классификации (точность около 0,61), что обеспечивает корректность их сравнения с точки зрения интерпретируемости. Для всех комментариев тестовой выборки удаётся построить хотя бы один объясняющий путь ( $GPI\_cov = 1,0$ ), поэтому различия между моделями определяются не наличием объяснений, а их качеством.

Наибольшую взвешенную согласованность объясняющих путей демонстрирует графовая сеть внимания GAT ( $GPI\_w = 0,333$ ), далее следует GCN (0,254), тогда как для многослойного перцептрона (MLP) этот показатель равен нулю вследствие отсутствия графового механизма вывода.

Полученные результаты подтверждаются тестом удаления концептов (Deletion Test). Модель GAT характеризуется наименьшей площадью под кривой удаления ( $AUC = 0,860$  против 0,920 у GCN и 0,914 у MLP), что указывает на более высокую верность формируемых объяснений. Удаление признаков, которым модель

GAT присвоила наибольшую важность, приводит к более резкому снижению уверенности классификатора, чем для GCN и MLP (рис. Л.3).

Дополнительно выполнена оценка согласованности автоматически сформированных объяснений с человеческими обоснованиями. По показателям token-F1 и token-AUPRC наиболее высокие результаты получены для GAT, что согласуется с выводами, сделанными по метрике GPI и результатам теста удаления концептов.

Результаты сравнения моделей приведены в таблице Л.3, а графическая интерпретация показателей интерпретируемости GPI\_w и согласованности автоматически сформированных объяснений с человеческими обоснованиями (token-AUPRC) – на рисунке Л.2.

Таблица Л.3 – Точность классификации, показатели интерпретируемости GPI и согласованности с человеческими обоснованиями (1 924 комментария)

| Модель | Точность | GPI cov | GPI w | AUC удаления | token-F1 | token-AUPRC |
|--------|----------|---------|-------|--------------|----------|-------------|
| GAT    | 0,611    | 1,000   | 0,333 | 0,860        | 0,456    | 0,391       |
| GCN    | 0,612    | 1,000   | 0,254 | 0,920        | 0,456    | 0,386       |
| MLP    | 0,609    | 1,000   | 0,000 | —            | —        | —           |

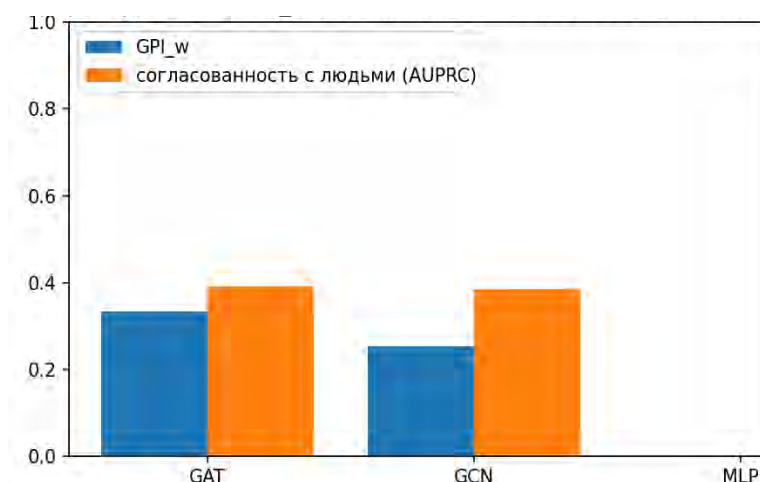


Рисунок Л.2 – Сравнение моделей по показателям интерпретируемости GPI\_w и согласованности объяснений с человеческими обоснованиями

Признаки последовательно удаляются в порядке убывания их важности, определённой каждой моделью. Более раннее и резкое снижение уверенности классификатора (меньшая площадь под кривой) свидетельствует о более высокой верности формируемых объяснений.

#### Л.7 Согласованность объяснений с человеческими обоснованиями

Дополнительная проверка качества объяснений выполнена путём их сопоставления с человеческими обоснованиями. По результатам оценки достигнуто значение token-F1=0,456 и token-AUPRC=0,391, что подтверждает согласованность

автоматически сформированных объяснений с токенами, выделенными аннотаторами [242]. Для модели MLP такая оценка не выполнялась, поскольку отсутствие графового механизма вывода не позволяет сформировать объясняющие пути.

Примеры автоматически сформированных объяснений приведены в таблице Л.4 (токсичные термины частично цензурированы).

Выделенные моделью токены совпадают с токенами, отмеченными аннотаторами в качестве обоснования токсичности. Для комментариев, не содержащих статистически значимых токсичных признаков, объясняющий путь не формируется, что соответствует решению «не токсично».

Приведённые примеры иллюстрируют, что автоматически сформированные объясняющие пути опираются на те же ключевые токены, которые используются человеком при обосновании решения о токсичности.

Таблица Л.4 – Примеры решений и объясняющих путей

| Комментарий (фрагмент)                   | Вердикт     | Объясняющий путь                                            | Модель отметила | Человек отметил |
|------------------------------------------|-------------|-------------------------------------------------------------|-----------------|-----------------|
| «...that ghetto mess for everybody else» | токсично    | комментарий → имеет_признак[токсичный термин] → токсичность | ghetto          | ghetto          |
| «...i'm a little r****ded»               | токсично    | комментарий → имеет_признак[токсичный термин] → токсичность | r**ded          | r**ded          |
| «these nr bs look like...»               | токсично    | комментарий → имеет_признак[токсичный термин] → токсичность | n**r            | nr, bs          |
| «the m***ie loving natsoc...»            | токсично    | комментарий → имеет_признак[токсичный термин] → токсичность | m***ie*         | m***ie*         |
| «I hope you have a wonderful day»        | не токсично | статистически значимый объясняющий путь отсутствует         | —               | —               |

## Л.8 Анализ смещения

В ходе эксперимента установлено, что признак «Упоминание защищённой группы» характеризуется статистически значимой отрицательной ассоциацией с токсичностью ( $r = -0,137$ ,  $q \approx 10^{-64}$ ) и в соответствии с правилом формирования графа знаний не включается в граф объяснений.

Объяснение токсичности формируется только при наличии признаков, действительно связанных с токсичностью. Тем самым при формировании объяснений исключается использование признаков, потенциально приводящих к смещению интерпретации результатов классификации.

Таким образом, прозрачная структура графа знаний и объясняющих путей делает возможным выявление признаков, способных вносить систематическое смещение в процесс принятия решений. Это расширяет возможности предложенного метода при построении интерпретируемых интеллектуальных систем.

#### Л.9 Языково-согласованная проверка на русскоязычном наборе

Для проверки переносимости в пределах языка предметной области диссертации метод дополнительно применён к русскоязычной классификации токсичных комментариев на наборе Russian Language Toxic Comments (223 461 и 24 829 комментариев обучающей и тестовой частей, доля токсичных  $\approx 0,18$ ; использованы подвыборки 50 000 / 5 000). Русскоязычные наборы не содержат человеческих обоснований, поэтому сверка объяснений с людьми здесь не проводится; оцениваются покрытие путями, взвешенная согласованность и верность. Реализация паттерна та же; интерпретируемые признаки заданы русскими лексиконами (обценная лексика – по корням, оскорбление, угроза, упоминание группы), правилом адресности (в русском языке различие 2-го лица «ты/вы» осмысленно) и выученным по логарифму отношения шансов токсичным словарём.

Граф знаний приведен в таблице Л.5 и на рисунке Л.3.

Таблица Л.5 – Ассоциация признаков с токсичностью

| Признак               | $r$    | $q$                   | Ребро графа |
|-----------------------|--------|-----------------------|-------------|
| Обценная лексика      | +0,610 | $\approx 0$           | *           |
| Токсичный термин      | +0,581 | $\approx 0$           | *           |
| Оскорбление           | +0,371 | $\approx 0$           | *           |
| Угроза                | +0,131 | $1,1 \cdot 10^{-189}$ | *           |
| Адресность (2-е лицо) | +0,119 | $1,0 \cdot 10^{-155}$ | *           |
| Упоминание группы     | +0,033 | $2,8 \cdot 10^{-13}$  | *           |

Он имеет более сильные ассоциации, чем англоязычный набор: ведущие признаки: обценная лексика ( $r = +0,61$ ), токсичный термин ( $r = +0,58$ ) и оскорбление ( $r = +0,37$ ); все шесть признаков значимы и положительны. В отличие от англоязычного набора, где упоминание группы было отрицательным, здесь оно слабоположительно ( $r = +0,03$ ), то есть выраженного смещения по упоминанию группы в данном наборе не выявлено.



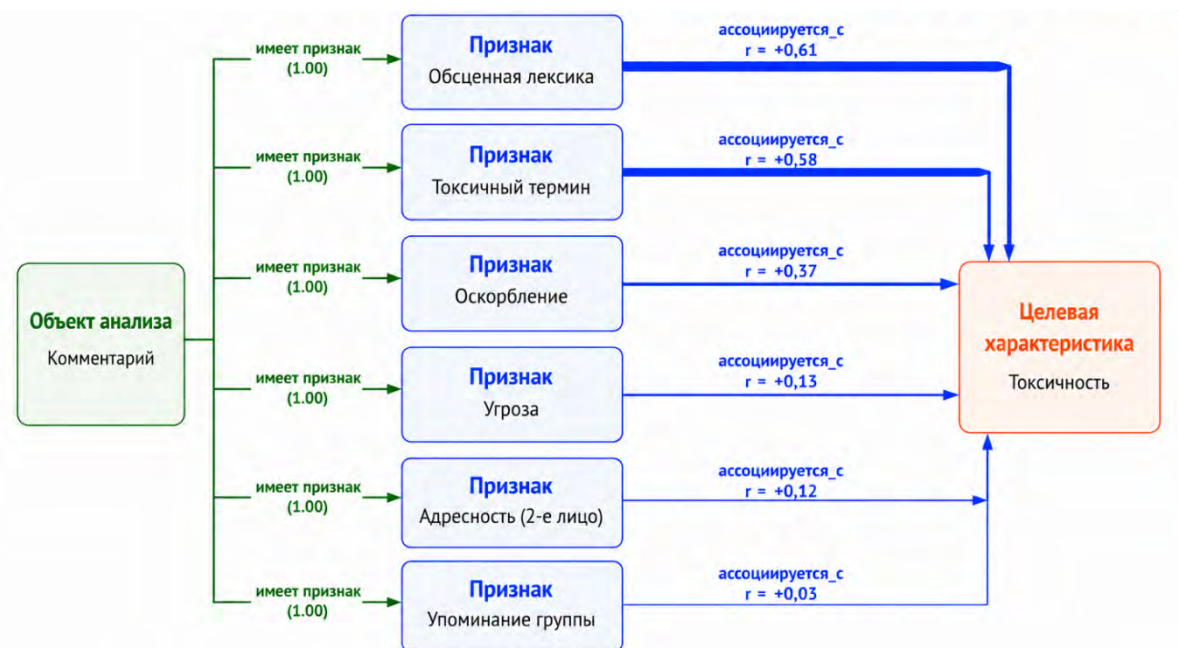


Рисунок Л.3 – Граф знаний для русскоязычного набора

Результаты рассуждающих модулей приведены в таблице Л.6. Они воспроизводят картину англоязычного эксперимента: при близкой точности графовая сеть внимания направляет существенную долю веса решения на валидные пути ( $GPI\_w = 0,259$ ), равномерная агрегация – заметно меньше ( $0,079$ ); объяснения GAT наиболее верны (наименьшая площадь под кривой удаления). Точность здесь выше ( $\approx 0,93$ ), поскольку русская обсценная лексика является сильным сигналом токсичности.

Таблица Л.6 – Точность, GPI и верность по рассуждающим модулям (русскоязычный набор, тест 5 000 комментариев).

| Модуль | Точность | GPI cov | GPI w | AUC удаления |
|--------|----------|---------|-------|--------------|
| GAT    | 0,929    | 1,000   | 0,259 | 0,838        |
| GCN    | 0,926    | 1,000   | 0,079 | 0,924        |
| MLP    | 0,929    | 1,000   | 0,000 | 0,914        |

Примеры объясняющих путей приведены в таблице. Л.7 (токсичные слова частично цензурированы).

Таблица Л.7 – Примеры объясняющих путей для русскоязычных токсичных комментариев

| Комментарий (фрагмент)                       | Вердикт        | Объясняющий путь                                             | Модель отметила    |
|----------------------------------------------|----------------|--------------------------------------------------------------|--------------------|
| «...за это расстрел!<br>...убить...»         | токсично       | комментарий → имеет_признак[токсичный термин] → токсичность  | убить,<br>расстрел |
| «...ты б**, долбо****...»                    | токсично       | комментарий → имеет_признак[обсценная лексика] → токсичность | б**,<br>долбо****  |
| «Спасибо за<br>информацию, очень<br>полезно» | не<br>токсично | статистически значимый объясняющий<br>путь отсутствует       | —                  |

Для комментария «...за это расстрел! ...убить...» формируется объясняющий путь «комментарий → имеет\_признак[токсичный термин] → ассоциируется\_с → токсичность», при этом в качестве ключевых токенов выделяются слова «убить» и «расстрел». Для комментария с обценной лексикой решение формируется через признак «обценная лексика» (б\*\*, долбо\*\*\*\*). Для нетоксичных комментариев статистически значимый объясняющий путь отсутствует. Полученные результаты подтверждают переносимость предложенного метода на задачу интерпретируемой классификации русскоязычных токсичных комментариев.

#### Л.10 Реализация полного паттерна на задаче эмоциональной классификации

Дополнительная проверка выполнена на русскоязычном корпусе CEDR (Corpus for Emotion Detection in Russian) [240], содержащем 7 528 сообщений обучающей и 1 882 сообщения тестовой выборки из LiveJournal и новостной ленты. Так как каждому сообщению сопоставлен набор из пяти эмоций (радость, грусть, удивление, страх и гнев), решается задача мультиклассовой классификации эмоциональной окраски текстов.

Особенность данной проверки заключается в наличии нескольких взаимосвязанных целевых характеристик. В отличие от предыдущих экспериментов, здесь используется полный онтологический паттерн, включающий все три типа отношений: имеет\_признак, ассоциируется\_с и коррелирует\_с. Тем самым воспроизводится структура паттерна, использованная при психологическом профилировании (глава 5). Интерпретируемые признаки представлены пятью эмоциональными лексиконами, автоматически сформированными по статистической значимости слов для соответствующих эмоций, а также признаком наличия восклицательного знака.

Сформированный граф знаний представлен на рисунке Л.4.

Граф отражает структуру предметной области. Каждый эмоциональный лексикон демонстрирует наиболее сильную положительную ассоциацию с соответствующей эмоцией, представленную в таблице Л.8, а предикат коррелирует\_с описывает взаимосвязи между целевыми характеристиками. Наиболее выраженная отрицательная корреляция наблюдается между эмоциями «радость» и «грусть» ( $r = -0,23$ ), что согласуется с характером разметки корпуса, в котором данные эмоции преимущественно представлены как взаимоисключающие.



Рисунок Л.4 – Граф знаний для эмоциональной классификации

Таблица Л.8 – Ассоциации «признак → эмоция и корреляции эмоций» (CEDR)

| Предикат        | Связь                           | г     |
|-----------------|---------------------------------|-------|
| ассоциируется_с | лексика «грусть» → грусть       | +0,54 |
| ассоциируется_с | лексика «страх» → страх         | +0,53 |
| ассоциируется_с | лексика «радость» → радость     | +0,52 |
| ассоциируется_с | лексика «удивление» → удивление | +0,49 |
| ассоциируется_с | лексика «гнев» → гнев           | +0,36 |
| ассоциируется_с | восклицание → радость           | +0,14 |
| коррелирует_с   | радость ↔ грусть                | -0,23 |
| коррелирует_с   | грусть ↔ удивление              | -0,14 |
| коррелирует_с   | радость ↔ страх                 | -0,13 |

Результаты сравнения моделей представлены в таблице Л.9. Несмотря на сопоставимое качество классификации ( $\text{micro-F1} \approx 0,4$ ), графовая сеть внимания обеспечивает значительно более высокую интерпретируемость ( $\text{GPI}_w = 0,962$ ) по сравнению с GCN (0,177), тогда как для MLP этот показатель равен нулю. Таким образом, даже в условиях мультиклассовой классификации с несколькими взаимосвязанными целевыми характеристиками метрика GPI сохраняет способность различать модели по качеству формируемых объяснений.

Таблица Л.9 – Точность и GPI по рассуждающим модулям

| Модуль | micro-F1 | GPI_cov | GPI_w |
|--------|----------|---------|-------|
| GAT    | 0,434    | 1,000   | 0,962 |
| GCN    | 0,362    | 1,000   | 0,177 |
| MLP    | 0,403    | 1,000   | 0,000 |

Полученные результаты подтверждают применимость полного онтологического паттерна к задачам с несколькими целевыми характеристиками и расширяют экспериментальное подтверждение его переносимости на ещё одну предметную область анализа текстовых данных.

Проведённые эксперименты подтверждают переносимость предложенного метода на две новые предметные области: интерпретируемую классификацию токсичных комментариев и анализ эмоциональной окраски текстов, существенно отличающиеся от психологического профилирования и дерматоскопической диагностики (Приложение К). Во всех случаях без изменения структуры использованы инвариантное ядро метода, включающее онтологический паттерн, граф знаний, графовую сеть внимания и метрику GPI; изменялась только доменная онтология признаков.

При сопоставимом качестве классификации метрика GPI устойчиво различает модели с графовым механизмом вывода и модели, не использующие явное представление предметных знаний, демонстрируя более высокую интерпретируемость графовых подходов. Согласованность автоматически формируемых объяснений с человеческими обоснованиями объективно оценена с использованием набора HateXplain.

Дополнительная проверка на русскоязычных наборах данных показала применимость предложенного подхода независимо от языка анализируемых текстов. Эксперимент на корпусе CEDR подтвердил возможность использования полного онтологического паттерна, включая предикат коррелирует\_с, при анализе нескольких взаимосвязанных целевых характеристик.

Таким образом, результаты приложения экспериментально доказывают доменную и языковую независимость предложенного метода, а также универсальность разработанных онтологического паттерна и метрики GPI для интерпретируемого анализа мультимодальных данных в различных предметных областях.

## ПРИЛОЖЕНИЕ М Формализация онтологии и графа знаний предметной области

В настоящем приложении представлены формальные спецификации онтологии цифрового поведения и реализующего её графа знаний для основной предметной области исследования – психологического профилирования пользователей социальной сети. Концептуальные основы онтологии рассмотрены в главе 2, тогда как в настоящем приложении приведены её полная спецификация в формате OWL/RDF, схема реализации графа знаний в СУБД Neo4j и примеры триплетов, не включённые в основной текст диссертации ввиду их значительного объёма. Представленная онтология является предметной реализацией доменно-инвариантного онтологического паттерна «объект анализа – признак – целевая характеристика», описание которого приведено в приложении Н.

### М.1 Концепты и предикаты онтологии

Онтология предметной области представляет собой предметную реализацию доменно-инвариантного онтологического паттерна и специализирует его основные классы, представленные в таблице М.1.

Таблица М.1 – Специализация классов доменно-инвариантного онтологического паттерна для предметной области психологического профилирования пользователей социальной сети

| Класс паттерна                                | Доменная специализация                                                     | Метка в графе       |
|-----------------------------------------------|----------------------------------------------------------------------------|---------------------|
| AnalysisObject (Объект анализа)               | профиль пользователя социальной сети                                       | User                |
| Feature (Признак)                             | категориальный признак (никнейм, тип аватара, поведенческий паттерн и др.) | Feature             |
| TargetCharacteristic (Целевая характеристика) | психологический конструкт (фактор личности)                                | PsychConstruct      |
| TargetMeasurement (Измерение)                 | измеренное значение фактора (методика 16PF и др.)                          | PsychometricMeasure |

Семантические связи между сущностями формализуются бинарными предикатами. Инвариантное ядро онтологии образуют отношения `has_feature` (объект анализа → признак), `associated_with` (признак → целевая характеристика) и `correlates_with` (целевая характеристика ↔ целевая

характеристика). Для отношения `associated_with` вес ребра соответствует силе статистической ассоциации между признаком и целевой характеристикой и уточняется в процессе обучения графовой нейронной сети. Полный перечень используемых типов отношений, включая отношение `influences` (целевая характеристика  $\rightarrow$  признак), приведён в главе 2.

## M.2 Спецификация онтологии в OWL/RDF

Доменная онтология специфицирована средствами языка OWL на основе доменно-инвариантного онтологического паттерна, описание которого приведено в приложении Н (листинг Н.1). Спецификация предметной онтологии в синтаксисе Turtle [204, 205]:

```
@prefix : <http://example.org/dbo#> .      # digital-behavior ontology
@prefix insp: <http://example.org/insp#> .    # инвариантное ядро паттерна
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

# --- доменные классы как подклассы классов паттерна ---
:User          a owl:Class ; rdfs:subClassOf insp:AnalysisObject .
:Feature        a owl:Class ; rdfs:subClassOf insp:Feature .
:PsychConstruct a owl:Class ; rdfs:subClassOf insp:TargetCharacteristic .
:PsychometricMeasure a owl:Class ; rdfs:subClassOf insp:TargetMeasurement .

# --- предикаты (наследуют домены/ко-домены ядра) ---
:has_feature    a owl:ObjectProperty ; rdfs:domain :User ; rdfs:range :Feature .
:associated_with a owl:ObjectProperty ; rdfs:domain :Feature ; rdfs:range :PsychConstruct .
:correlates_with a owl:ObjectProperty , owl:SymmetricProperty ;
    rdfs:domain :PsychConstruct ; rdfs:range :PsychConstruct .
:influences     a owl:ObjectProperty ; rdfs:domain :PsychConstruct ; rdfs:range :Feature .
:has_measurement a owl:ObjectProperty ; rdfs:domain :User ; rdfs:range :PsychometricMeasure .

# --- атрибуты и веса ---
:associationStrength a owl:AnnotationProperty .    # вес ребра associated_with
:featureType        a owl:DatatypeProperty ; rdfs:domain :Feature ; rdfs:range xsd:string .
:factorCode         a owl:DatatypeProperty ; rdfs:domain :PsychConstruct ; rdfs:range xsd:string .

# --- пример экземпляров (ABox) ---
:user_001 a :User ;      :has_feature :nickname_literary ; :has_measurement :m_001 .
:nickname_literary a :Feature ; :featureType "nickname" ;
    :associated_with [ a :PsychConstruct ; :factorCode "M" ;
        :associationStrength 0.41 ] .
```

В листинге представлены определения основных классов предметной онтологии, их соответствие классам доменно-инвариантного паттерна, объектные и атрибутивные свойства, а также пример задания экземпляров

предметной области. Предметная реализация онтологии сохраняет доменно-инвариантную структуру онтологического паттерна, благодаря чему при переходе к другой предметной области изменяется только её предметное наполнение, тогда как структура паттерна остаётся неизменной.

### М.3 Схема графа знаний в Neo4j

Онтология предметной области реализуется в виде графа свойств под управлением СУБД Neo4j [236]. В графе используются следующие типы узлов: User, Feature, PsychConstruct, PsychometricMeasure. Семантические отношения онтологии представлены следующими типами рёбер: HAS\_FEATURE (User → Feature), ASSOCIATED\_WITH (Feature → PsychConstruct) и CORRELATES\_WITH (PsychConstruct ↔ PsychConstruct). Для рёбер типа ASSOCIATED\_WITH задаются свойства weight, отражающее силу ассоциации, и p\_value, соответствующее уровню статистической значимости; для рёбер типа CORRELATES\_WITH задаётся свойство r, соответствующее коэффициенту корреляции. Определение схемы графа, включая ограничения уникальности и индексы, а также пример наполнения графа приведены в листинге.

```
// --- ограничения уникальности (DDL) ---
CREATE CONSTRAINT user_id IF NOT EXISTS
  FOR (u:User)      REQUIRE u.id IS UNIQUE;
CREATE CONSTRAINT feature_id IF NOT EXISTS
  FOR (f:Feature)   REQUIRE f.id IS UNIQUE;
CREATE CONSTRAINT pc_id IF NOT EXISTS
  FOR (p:PsychConstruct) REQUIRE p.id IS UNIQUE;
CREATE INDEX feature_type IF NOT EXISTS FOR (f:Feature) ON (f.type);

// --- пример наполнения графа (триплеты) ---
MERGE (u:User {id:'user_001', age:27, gender:'F', city:'Samara'})
MERGE (f:Feature {id:'nickname_literary', type:'nickname', value:'literary'})
MERGE (p:PsychConstruct {id:'factor_M', code:'M', name:'Мечтательность'})
MERGE (u)-[:HAS_FEATURE]->(f)
MERGE (f)-[:ASSOCIATED_WITH {weight:0.41, p_value:0.003}]->(p);

// --- извлечение объясняющего пути (запрос вывода) ---
MATCH path = (u:User {id:'user_001'})-[:HAS_FEATURE]->(f:Feature)
  -[a:ASSOCIATED_WITH]->(p:PsychConstruct)
WHERE a.p_value < 0.05
RETURN f.value AS признак, p.name AS целевая_характеристика, a.weight AS вес
ORDER BY a.weight DESC;
```

#### М.4 Примеры триплетов и логического вывода

Представление знаний в графе основано на RDF-триплетах вида (субъект, предикат, объект), описывающих семантические отношения между сущностями предметной области. Примеры доменных триплетов приведены ниже:

- (user\_001, has\_feature, nickname\_literary) – пользователь обладает признаком «литературный никнейм»;
- (nickname\_literary, associated\_with, factor\_M) – признак статистически ассоциирован с фактором М (Мечтательность), вес связи составляет 0,41;
- (factor\_M, correlates\_with, factor\_Q1) – между факторами М (Мечтательность) и Q1 (Радикализм) установлена корреляционная связь.

Композиция отношений has\_feature◦associated\_with формирует элементарный семантический путь вывода вида «объект анализа → признак → целевая характеристика», который используется при построении интерпретируемых объяснений и вычислении метрики Graph Path Interpretability (GPI) (приложение П). Так, для предсказания значения фактора М у пользователя user\_001 объясняющий путь имеет вид: user\_001 → has\_feature → nickname\_literary → associated\_with → factor\_M.

Полный перечень статистически значимых ассоциаций «признак → целевая характеристика», а также матрица корреляций целевых характеристик приведены в приложениях Г и Е.



## ПРИЛОЖЕНИЕ Н Онтологический паттерн проектирования «объект анализа – признак – целевая характеристика»

В настоящем приложении приведена полная формализация предметно-независимого онтологического паттерна проектирования интерпретируемых нейросимвольных систем, составляющего основу диссертационной работы. Паттерн задаёт доменно-инвариантное ядро онтологии, в котором прослеживаемость вывода является следствием онтологической структуры, а не отдельно добавленным свойством. Развёрнутое изложение приведено в публикации автора [А3].

### Н.1 Классы концептов

Паттерн содержит четыре класса сущностей, инвариантных относительно предметной области:

- Объект анализа (AnalysisObject) – анализируемая сущность как совокупность наблюдаемых проявлений; описывается характеристиками, извлекаемыми из источника напрямую, и признаками, генерируемыми моделями машинного обучения из исходных модальностей.

- Признак (Feature) – категориальное значение атрибута объекта анализа (состав доменно-специфичен).

- Целевая характеристика (TargetCharacteristic) – теоретическое понятие предметной области, на основе которого строится интерпретация; промежуточный класс, отделяющий теоретический уровень от эмпирического.

- Измерение целевой характеристики (TargetMeasurement) – значения целевых характеристик, полученные верифицированными методиками; эталон для обучения и валидации.

- Разделение классов «Целевая характеристика» и «Измерение целевой характеристики» отделяет теоретические понятия от измеренных значений, допускает добавление новых характеристик без перестройки структуры данных и поддерживает обобщённые правила вида «признак ассоциирован с целевой характеристикой».

### Н.2 Предикаты

Семантические связи формализованы бинарными предикатами; инвариантное ядро образуют:

- has\_feature (имеет\_признак): домен – Объект анализа, ко-домен – Признак.

– `associated_with` (ассоциируется\_с): домен – Признак, ко-домен – Целевая характеристика; кодирует статистическую ассоциацию, вес ребра соответствует силе связи и уточняется при обучении графовой сети.

– `correlates_with` (коррелирует\_с): домен и ко-домен – Целевая характеристика; теоретически ожидаемые связи характеристик, используемые как априорное ограничение.

Двухзвенная композиция `has_feature` ◦ `associated_with` образует элементарный семантический путь вывода «объект анализа – признак – целевая характеристика», лежащий в основе механизма интерпретации (Н.5).

### Н.3 Формализация в OWL и RDF

Базовые классы и объектные свойства специфицируются стандартными средствами OWL; экземпляры сериализуются в RDF, что обеспечивает интероперабельность, проверяемость ограничений и согласованную передачу структуры в графовую нейронную сеть. Спецификация инвариантного ядра в синтаксисе Turtle приведена в листинге Н.1.

```
@prefix :      <http://example.org/insp#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .

:AnalysisObject      a owl:Class .
:Feature              a owl:Class .
:TargetCharacteristic a owl:Class .
:TargetMeasurement   a owl:Class .
[ a owl:AllDisjointClasses ;
  owl:members ( :AnalysisObject :Feature :TargetCharacteristic
:TargetMeasurement ) ] .

:has_feature          a owl:ObjectProperty ;
  rdfs:domain :AnalysisObject ;      rdfs:range :Feature .
:associated_with      a owl:ObjectProperty ;
  rdfs:domain :Feature ;              rdfs:range :TargetCharacteristic .
:correlates_with      a owl:ObjectProperty , owl:SymmetricProperty ;
  rdfs:domain :TargetCharacteristic ; rdfs:range :TargetCharacteristic .
:has_measurement      a owl:ObjectProperty ;
  rdfs:domain :AnalysisObject ;      rdfs:range :TargetMeasurement .

# вес ассоциации как аннотация ребра
:associationStrength a owl:AnnotationProperty .
```

Листинг Н.1. Спецификация инвариантного ядра паттерна в синтаксисе Turtle (OWL/RDF).

Концептуальная схема инвариантного ядра приведена на рисунке Н.1.

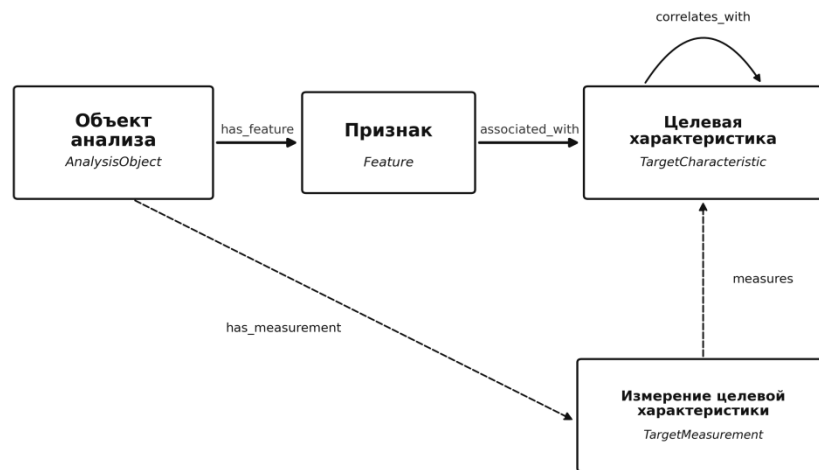


Рисунок Н.1 – Схема онтологического паттерна: классы концептов и предикаты инвариантного ядра

#### Н.4 Аксиомы и ограничения

Паттерн фиксирует набор онтологических аксиом, задающих корректность реализации и определяющих единицу интерпретации:

1. Аксиомы домена и ко-домена. Каждый предикат ограничен по области определения и значениям; это исключает семантически некорректные рёбра.
2. Дизъюнктность классов. Четыре базовых класса попарно не пересекаются, что сохраняет разделение «эмпирическое – теоретическое».
3. Симметричность `correlates_with`. Корреляция целевых характеристик не имеет направления.
4. Аксиома пути интерпретации. Единицей интерпретации служит композиция `has_feature` ◦ `associated_with`: путь «*o* – `has_feature` – *f* – `associated_with` – *t*» допустим тогда и только тогда, когда оба ребра удовлетворяют своим аксиомам; множество допустимых объяснений предопределено онтологией, а не выбирается апостериорно.
5. Аннотация веса. Ребро `associated_with` несёт числовой вес (`associationStrength`); источник веса (статистическая оценка, экспертный приор, обучаемый параметр) задаётся при реализации.

Аксиомы 1–4 инвариантны и переносятся вместе с паттерном; способ получения весов (аксиома 5) относится к доменно-специфичной части.

#### Н.5 Порядок проектирования системы на основе паттерна

Паттерн задаёт структуру и порядок её реализации в работающую интерпретируемую систему. Паттерн представлен на рисунке Н.2.

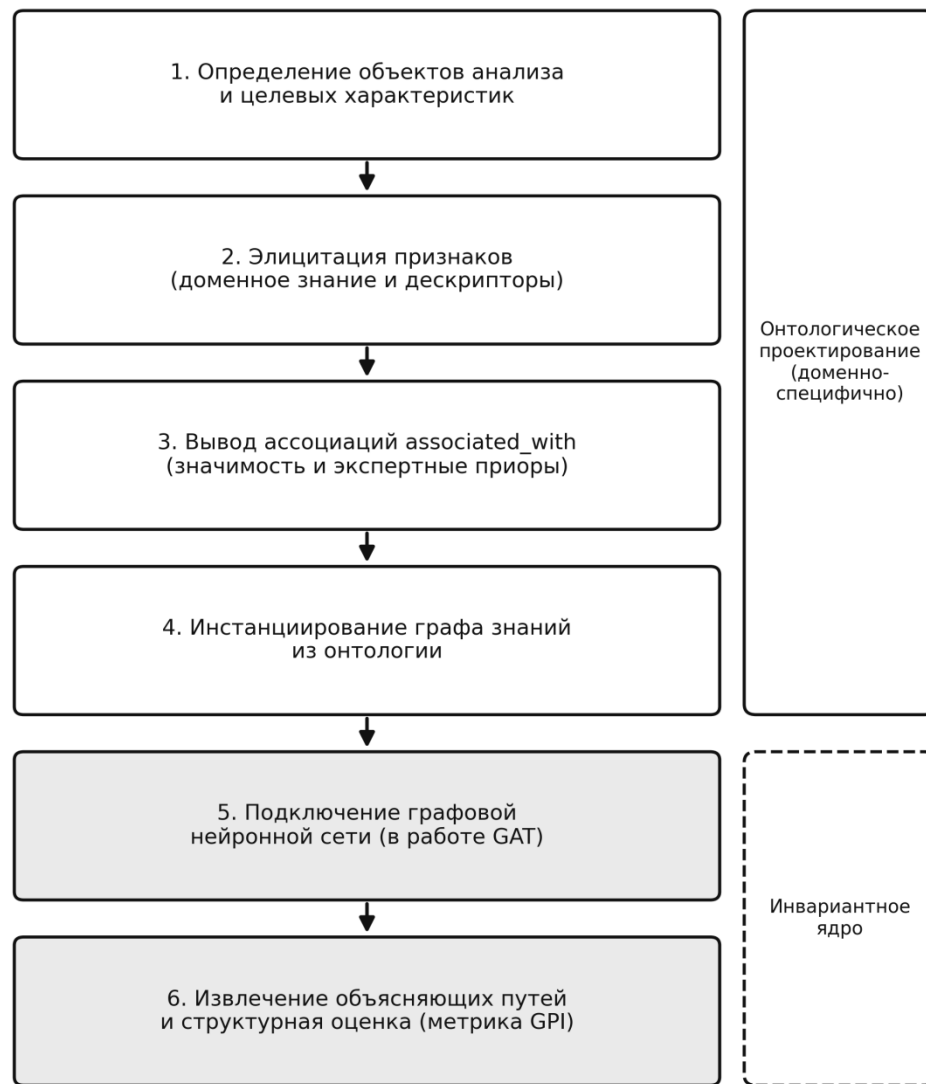


Рисунок Н.2 – Порядок проектирования системы на основе паттерна Н.6 Объяснимость как следствие структуры и переносимость

1. Определение объектов анализа и целевых характеристик – идентификация анализируемой сущности и предсказываемых теоретических понятий.
2. Элицитация признаков – определение состава признаков по предметному знанию и измеримым дескрипторам модальностей.
3. Вывод ассоциаций – формирование рёбер `associated_with` по статистической значимости связи признака с целевой характеристикой (с поправкой на множественность и с учётом приоров).
4. Реализация графа знаний – отображение онтологии в направленный граф с весами.

5. Подключение графовой нейронной сети – обучение поверх графа (в работе – GAT, веса внимания над узлами-признаками интерпретируются как вклад признаков); паттерн не предписывает конкретную архитектуру.

6. Извлечение и оценка объяснений – формирование подграфа вывода, ранжированных путей и их оценка структурной метрикой (GPI, Приложение П).

Шаги 1–4 (онтологическое проектирование) доменно-специфичны по содержанию, но инвариантны по форме; шаги 5-6 доменно-инвариантны – в этом и состоит переиспользуемость метода.

Элементарное объяснение предсказания есть семантический путь из ребра `has_feature` (объект обладает признаком) и ребра `associated_with` (признак ассоциирован с целевой характеристикой предсказания). Путь существует и проверяем именно потому, что онтология заранее фиксирует допустимые типы связей; если модель опирается на признак без значимого ребра к предсказанному классу, путь не строится – сигнал решения без онтологической опоры.

Формальный критерий переносимости. Паттерн как терминологический компонент (TBox) задаётся четвёркой  $P = \langle C, R, A, M \rangle$ :  $C$  – классы (Н.1),  $R$  – типы предикатов (Н.2),  $A$  – аксиомы (Н.4),  $M$  – механизм вывода (семантический путь). Реализация домена пополняет фактологический компонент (ABox): конкретные экземпляры классов Признак и Целевая характеристика и источник весов. Коэффициент переиспользования

$$\rho = |\text{инвариантные элементы}| / |\text{все терминологические элементы}|.$$

Сопоставление двух реализаций приведено в таблице Н.1 и на рисунке Н.3.

Для обеих рассмотренных реализаций (психологическое профилирование и дерматоскопия, Приложение К) все классы, предикаты, аксиомы и механизм вывода перенесены без изменения, доменная адаптация ограничена пополнением ABox; на уровне схемы  $\rho = 1,0$ , а проектные трудозатраты масштабируются только с пополнением фактологического компонента.

Таблица Н.1 – Доменно-инвариантные и доменно-специфичные элементы при реализации паттерна

| Элемент проектирования | Инвариантно (паттерн)               | Домен 1: психология         | Домен 2: дерматоскопия         |
|------------------------|-------------------------------------|-----------------------------|--------------------------------|
| Объект анализа         | класс AnalysisObject                | профиль соцсети             | дерматоскопическое изображение |
| Признаки               | класс Feature, предикат has_feature | поведенческие и визуальные  | дескрипторы ABCD и структуры   |
| Целевая характеристика | класс TargetCharacteristic          | факторы личности            | диагностические категории      |
| Измерение              | класс TargetMeasurement             | психологическая методика    | верифицированный диагноз       |
| Ассоциации             | предикат associated_with            | связь поведения с факторами | связь дескрипторов с диагнозом |
| Обучаемая модель       | графовая нейронная сеть (GAT)       | та же                       | та же                          |
| Механизм объяснения    | семантический путь в графе          | тот же                      | тот же                         |

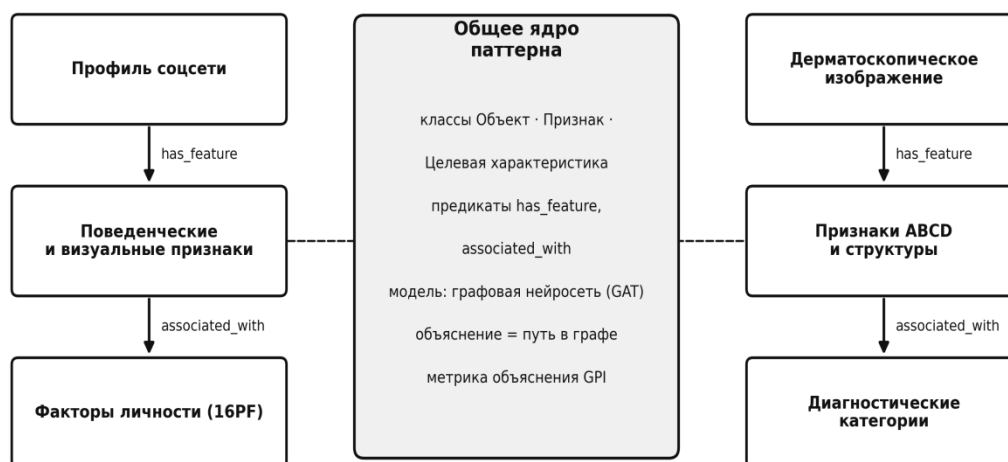


Рисунок Н.3 – Две реализации паттерна (психологическая и дерматоскопическая онтологии) с выделенным общим ядром

Совпадение инвариантного ядра при существенном различии доменов (текстово-поведенческие данные против оптических изображений; психометрические конструкты против диагностических категорий) подтверждает доменную независимость паттерна и переносимость метода.

## ПРИЛОЖЕНИЕ П Метрика интерпретируемости путей в графе GPI

В настоящем приложении приведена полная формализация структурной метрики интерпретируемости решений нейросимвольной модели – метрики интерпретируемости путей в графе (Graph Path Interpretability, GPI), являющейся одним из результатов диссертационной работы. Метрика определяется над онтологической структурой графа знаний и потому переносится вместе с методом между предметными областями (Приложение К, Приложение Л).

### П.1 Идея и определение объясняющего пути

Исходное наблюдение: интерпретируемое объяснение предсказания имеет форму пути в графе знаний – оно связывает объект анализа с целевой характеристикой через один или несколько распознаваемых промежуточных признаков. Если такой путь существует, статистически подтверждён и именно на него опиралась модель, предсказание интерпретируемо в том смысле, который примет специалист предметной области.

Кандидатом на объяснение предсказания  $n$  служит путь из двух рёбер: объект анализа  $o_n$  – has\_feature  $\rightarrow$  признак  $i$  – associated\_with  $\rightarrow$  предсказанная целевая характеристика  $\hat{y}_n$ .

Путь считается значимым (клинически/семантически валидным), если выполнены оба условия:

- признак  $i$  присутствует у объекта ( $c_i^{(n)} > \tau$ , порог присутствия  $\tau = 0,2$ );
- ребро associated\_with от признака  $i$  к  $\hat{y}_n$  статистически значимо (после поправки на множественность,  $q < 0,05$ ).

GPI допускает две согласованные формы представления. В основном тексте диссертации (глава 5) GPI приводится как единый агрегированный показатель – доля решений, объяснение которых опирается на статистически значимый путь в графе знаний. В настоящем приложении и в кросс-доменных приложениях (дерматоскопия, токсичные комментарии) та же метрика рассматривается как система из основной метрики покрытия путями (GPI\_cov), вспомогательной метрики взвешенной согласованности путей (GPI\_w) и проверки верности методом удаления концептов. Единый показатель основного текста отвечает на вопрос о наличии валидного объяснения, а разложение на компоненты дополнительно оценивает степень использования объясняющих путей моделью и применяется для детального сравнения рассуждающих архитектур; при этом метрика покрытия GPI\_cov модельно-независима,

а взвешенная согласованность GPI\_w зависит от конкретной модели (использует её веса признаков).

## П.2 Основная метрика – покрытие путями (GPI\_cov)

Доля предсказаний, для которых найден хотя бы один значимый объясняющий путь. Метрика измеряет, способен ли граф знаний объяснить решения модели:

$$\text{GPI}_{\text{cov}} = (1 / N) \cdot \sum_{n=1} \dots N I_n , \quad (\text{П.1})$$

где  $I_n = 1$ , если существует признак  $i$ , присутствующий у объекта ( $c_i^{(n)} > \tau$ ) и имеющий значимое ребро к предсказанной целевой характеристике  $\hat{y}_n$ , и  $I_n = 0$  иначе. Значение  $\text{GPI}_{\text{cov}} \in [0, 1]$ ; величина вблизи 1 означает, что граф знаний допускает объясняющий путь почти для каждого решения (свойство покрытия словаря признаков), вблизи 0 — что для большинства решений объяснение в терминах онтологии отсутствует.

## П.3 Вспомогательная метрика – взвешенная согласованность путей (GPI\_w)

Степень совпадения собственных весов модели со значимыми путями; осмысленна лишь для моделей, выдающих веса признаков (например, веса внимания GAT), поэтому является вторичным, зависящим от архитектуры показателем:

$$\text{GPI}_w = (1 / N) \cdot \sum_n [ (\sum_i w_i^{(n)} \cdot s_i^{(n)}) / (\sum_i w_i^{(n)}) ], \quad (\text{П.2})$$

где  $w_i^{(n)}$  – важность признака  $i$  для модели (вес внимания для GAT; нормированное присутствие для равномерной агрегации; ноль для модели без путей), а семантическая значимость пути  $s_i^{(n)} = 1$ , если соответствующий путь валиден (признак присутствует и ребро к  $\hat{y}_n$  значимо), и 0 иначе. Тем самым  $\text{GPI}_w$  есть доля веса решения модели, приходящаяся на семантически валидные пути;  $\text{GPI}_w \in [0, 1]$ . Значение вблизи 0 (как у модели «чёрный ящик» без структуры путей) означает, что ни одна часть веса решения не приходится на валидные пути; большее значение – что приходится бóльшая доля.

## П.4 Свойства метрики

– Модельная независимость. GPI определяется над графом знаний и присутствием признаков, а не над внутренними параметрами конкретной сети; применима к любой модели, выдающей предсказание (для  $\text{GPI}_{\text{cov}}$ ) и дополнительно веса признаков (для  $\text{GPI}_w$ ).



- Заземлённость в предметных знаниях. Значимость ребра `associated_with` контролируется статистически (поправка на множественность), поэтому высокое `GPI_cov` означает: «решения объяснимы признаками, статистически связанными с целевой характеристикой» – утверждение, проверяемое специалистом.

- Разделение свойств. `GPI_cov` отвечает на вопрос «существует ли валидное объяснение», `GPI_w` – «насколько модель его использует»; это разные свойства, и метрики дополняют друг друга, а не дублируют.

- Сопоставимость. `GPI` – скалярные величины в  $[0, 1]$ , которые можно сравнивать между моделями, отслеживать в ходе разработки и оптимизировать наряду с точностью.

## П.5 Проверка верности методом удаления концептов

Поскольку объяснение может быть интерпретируемым, но неверным по отношению к модели, `GPI` перекрёстно проверяется тестом удаления концептов: признаки последовательно обнуляются в порядке убывания их важности  $w_i$ , и отслеживается относительная уверенность модели в предсказанном классе. Резкое монотонное падение уверенности (малая площадь под кривой удаления, `AUC`) указывает, что именно высококовзвешенные признаки определяют предсказание, то есть атрибуции `GPI_w` верны. Для модели, не выдающей весов, порядок удаления случаен и служит контролем.

## П.6 Валидация и применение

Метрика `GPI` применена для сравнительного анализа рассуждающих архитектур по критерию объяснимости при контроле точности (концептуальное «бутылочное горлышко», при котором целевая характеристика выводится только из признаков, так что различие интерпретируемости не сводится к точности). Результаты применения `GPI` в основной предметной области (психологическое профилирование) приведены в Главе 5 в виде единого агрегированного показателя; полное разложение метрики на компоненты – покрытие путями `GPI_cov`, взвешенную согласованность `GPI_w` и проверку верности – и независимая проверка переносимости на иные предметные области приведены в Приложениях К (дерматоскопическая диагностика) и Л (классификация токсичных комментариев), где показано, что при равной точности `GPI` отделяет графовое рассуждение с концептами (значительная доля веса решения на валидных путях) от модели «чёрный ящик» (нулевая доля), а тест удаления подтверждает верность атрибуций.

## ПРИЛОЖЕНИЕ Р Акты внедрения

## УТВЕРЖДАЮ

Директор ГБУ ДПО «Региональный социопсихологический центр»

Т.Н. Ключева



2026 г.

## АКТ

### о внедрении результатов диссертационного исследования

Комиссия в составе директора ГБУ ДПО «Региональный социопсихологический центр» Ключевой Татьяны Николаевны, начальника отдела информатизации Дорошенко Алексея Николаевича и начальника лаборатории практической психологии Грачёва Юрия Александровича составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

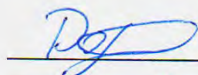
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в деятельность ГБУ ДПО «Региональный социопсихологический центр».

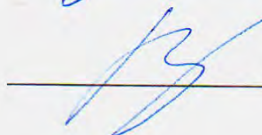
Разработанные результаты используются при анализе информации о рисках девиантного поведения обучающихся по цифровым следам в социальных сетях (VK), материалам различных интернет-источников, уточнении результатов психодиагностических исследований и формировании интерпретируемых рекомендаций для организации профилактической и коррекционной работы.

Использование результатов диссертационного исследования способствует повышению эффективности обработки диагностической информации и обоснованности принимаемых решений при формировании профилактической и воспитательной работы в образовательных организациях.

Начальник отдела информатизации

 / А.Н. Дорошенко

Начальник лаборатории  
практической психологии

 / Ю. А. Грачёв



УТВЕРЖДАЮ

И.О. начальника космодрома-директора

филиала АО «ЦЭНКИ» - КЦ «Восточный»

Д.Н. Люлько

2026 г.

## АКТ

о внедрении результатов диссертационного исследования

Комиссия в составе И.О. начальника космодрома-директора филиала АО «ЦЭНКИ» - КЦ «Восточный» Д.Н. Люлько, начальника управления эксплуатации технологического оборудования И.О.Лункина и начальника отдела информационной безопасности А.Н. Палабугина составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации, внедрены в деятельность ФГБУ «ЦЭНКИ» – Космодром «Восточный».

Разработанные результаты используются при обработке и анализе разнородной информации, интеграции данных из различных источников, формировании интерпретируемых результатов анализа и поддержке принятия решений.

Использование результатов диссертационного исследования способствует повышению эффективности обработки мультимодальной информации и обоснованности принимаемых решений.

Начальник управления эксплуатации  
технологического оборудования

« 20 » 08 /И.О. Лункин  
2026 г.

Начальник отдела  
информационной безопасности

« 20 » 08 / А.Н. Палабугин  
2026 г.

УТВЕРЖДАЮ  
Директор Единой службы заботы о  
клиентах г. Самары  
ПАО «Сбербанк»  
\_\_\_\_\_ /Кузнецова Н.В.  
« 24 » сентября 2026 г.

АКТ  
о внедрении результатов диссертационного исследования

Комиссия в составе директора Единой службы заботы о клиентах г. Самары Кузнецовой Н.В., HR BP Грачева Е.И и руководителя административного отдела Шестатковой Е.В. составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в деятельность подразделений Единой службы заботы о клиентах г. Самара ПАО «Сбербанк».

Разработанные результаты используются при комплексном анализе цифровых профилей кандидатов, интеграции данных из различных источников, формировании интерпретируемых рекомендаций и информационной поддержке принятия кадровых решений.

Использование результатов диссертационного исследования способствует повышению эффективности оценки профессиональных компетенций кандидатов, объективности кадровых решений и совершенствованию процессов подбора персонала.

HR BP

Единой службы о клиентах

\_\_\_\_\_ /Грачева Е.И.  
« 24 » сентября 2026г.

Руководитель административного  
отдела

\_\_\_\_\_ /Шестакова Е.В.  
« 24 » сентября 2026г.



УТВЕРЖДАЮ

Технический директор

ООО НПП «Интеллект Софт»

«23» июля 2026 г. /Кохася С.С.



АКТ

о внедрении результатов

диссертационного исследования в производственную деятельность

ООО НПП «Интеллект Софт»

Комиссия в составе технического директора ООО НПП «Интеллект Софт» Кохася С.С., руководителя отдела технической поддержки Одинцова А.Д. и руководителя отдела аналитики Яковлева А.П. составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации, интегрирующая вычислительную мультимодальную нейронную сеть вывода с интерпретируемым графом знаний, формируемым на основе доменно-инвариантного онтологического паттерна «Объект анализа – Признак – Целевая характеристика»;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации, основанный на совместном использовании графов знаний и графовых нейронных сетей и обеспечивающий формирование интерпретируемых результатов анализа;
- технология обработки и анализа типизированной мультимодальной информации, включающая извлечение признаков из различных модальностей, построение графа знаний, выполнение нейросимвольного вывода и формирование интерпретируемых объяснений;
- информационная система обработки и анализа типизированной мультимодальной информации, реализующая полный цикл обработки данных и предоставления интерпретируемых результатов анализа,

внедрены в практическую деятельность ООО НПП «Интеллект Софт».

Разработанные результаты используются при проектировании и разработке интеллектуальных информационных систем, интеграции разнородных источников данных и интерпретируемом анализе типизированной мультимодальной информации.

Использование результатов диссертационного исследования способствует повышению эффективности разработки интеллектуальных информационных систем и качества интерпретируемого анализа мультимодальной информации.

Руководитель отдела технической поддержки

 /Одинцов А.Д.  
«23» июля 2026 г.

Руководитель отдела аналитики

 /Яковлев А.П.  
«23» июля 2026 г.

УТВЕРЖДАЮ



Ректор Самарского университета

Богатырёв В.Д.

« 22 » июля 2026 г.

АКТ

о внедрении результатов  
диссертационного исследования в учебный процесс

Результаты диссертационного исследования по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем, в учебный процесс на кафедре программных систем внедрены следующие результаты:

1. Гибридная нейросимвольная модель интерпретируемого анализа мультимодальной информации, объединяющая мультимодальную нейронную сеть и граф знаний для совместной обработки разнородных данных и формирования интерпретируемых результатов анализа.
2. Нейросимвольный метод интерпретируемого анализа мультимодальной информации, основанный на совместном использовании графов знаний и графовых нейронных сетей и обеспечивающий получение объяснимых результатов анализа.
3. Информационная система интерпретируемого анализа мультимодальной информации, реализующая полный цикл обработки данных: извлечение признаков из различных модальностей, построение графа знаний, нейросимвольный вывод и формирование интерпретируемых объяснений.

на основании решения кафедры программных систем (протокол №11 от «16» июня 2026 г.).

Указанные результаты включены в курсы «Системы искусственного интеллекта», «Нейронные сети», «ИТ-практикум» образовательной программы «Программирование и информационные технологии» (02.03.02 Фундаментальная информатика и информационные технологии).

И.о. заведующего кафедрой

« 22 » июля 2026 г. /Гаврилов А.В.

Соискатель ученой степени

« 22 » июля 2026 г. /Жданова А.Н.

Начальник методического отдела  
учебно-методического управления  
к.т.н., доцент

« 22 » июля 2026 г. /Вдовин Р.А.

Начальник отдела сопровождения  
научных исследований  
к.т.н., доцент

« 22 » июля 2026 г. /Родионов Д.В.



УТВЕРЖДАЮ

Генеральный директор ООО «Профиль»

/Малюга Е.В.

2026 г.



АКТ

о внедрении результатов диссертационного исследования

Комиссия в составе генерального директора ООО «Профиль» Малюги Е.В., начальника специализированного конструкторско-технологического бюро Соловьева И. С. и Заместителя директора по производству Санникова Я. С. составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

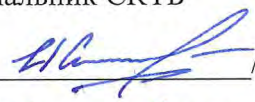
внедрены в инженерную и научно-техническую деятельность ООО «Профиль».

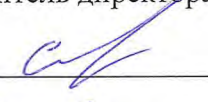
Результаты диссертационного исследования используются ООО «Профиль» при проработке подходов к созданию собственной системы работы с технической документацией и накопленными данными компании. Разработанные подходы применяются при структурировании и анализе технических заданий, дефектных ведомостей, спецификаций, проектной и исполнительной документации, расчетов и других материалов, используемых в инженерной и сметной деятельности. Отдельные элементы разработанного подхода используются при сопоставлении информации из различных документов, выявлении несоответствий и подготовке данных для последующего анализа специалистами.

Использование результатов исследования позволяет ООО «Профиль» систематизировать практический опыт применения искусственного интеллекта при работе с технической документацией и определить требования к дальнейшему развитию соответствующих информационных решений. В перспективе данный подход может быть использован для создания корпоративного инструмента, позволяющего структурировать накопленные документы и использовать опыт ранее выполненных проектов при анализе новых задач.

Начальник СКТБ

Заместитель директора по производству

 /Соловьев И. С.

 /Санников Я.С.

« 28 » августа 2026 г.

« 28 » августа 2026 г.



УТВЕРЖДАЮ

Генеральный директор

ООО «ФОТОН-БИО»

Кукушкин В.И.

«11» сентября 2026 г.



## АКТ

### о внедрении результатов диссертационного исследования

Комиссия в составе генерального директора Кукушкина Владимира Игоревича, руководителя производственного отдела Шапран Виталия Сергеевича и руководителя отдела разработок Кирпичева Вадима Евгеньевича составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в научно-исследовательскую деятельность компании.

Разработанные результаты используются при интеграции и анализе разнородной информации в области спектроскопии комбинационного рассеяния света, обработке спектральных, текстовых и графических данных, проведении научных исследований, а также формировании интерпретируемых аналитических заключений для поддержки разработки и применения спектрометров, SERS-подложек и биосенсоров.

Использование результатов диссертационного исследования способствует повышению эффективности обработки мультимодальных данных спектроскопии и смежных методов молекулярной диагностики, обеспечению интерпретируемости результатов анализа и развитию методов интеллектуальной поддержки научных исследований и прикладных задач в области медицины, биологии и промышленной аналитики.

Руководитель производственного отдела

В.И. Шапран /Шапран В.С.

«11» сентября 2026 г.

Руководитель отдела разработок

В.Е. Кирпичев /Кирпичев В.Е.

«11» сентября 2026 г.

УТВЕРЖДАЮ



Генеральный директор  
АНО «Иннокуб»

/Бураков С.Н.

августа 2026 г.

## АКТ

о внедрении результатов диссертационного исследования

Комиссия в составе генерального директора АНО «Центр развития инновационного потенциала малого и среднего бизнеса «Иннокуб» (АНО «Иннокуб») Буракова Сергея Николаевича, руководителя научно-исследовательского отдела Складовой Екатерины Николаевны и начальника отдела информационных технологий Дудолова Виталия Александровича составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- гибридная нейросимвольная модель интерпретируемого анализа типизированной мультимодальной информации;
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в научно-исследовательскую и инновационную деятельность АНО «Иннокуб».

Разработанные результаты используются при разработке и апробации интеллектуальных информационных технологий, обработке и анализе разнородной мультимодальной информации, интеграции данных из различных источников и формировании интерпретируемых аналитических материалов для анализа проектов в ходе отбора и прохождения акселерационных программ и поддержки исследовательской и инновационной деятельности.

Использование результатов диссертационного исследования способствует повышению эффективности разработки интеллектуальных информационных технологий, совершенствованию методов обработки мультимодальной информации и развитию информационно-аналитического обеспечения инновационной деятельности.

Заместитель начальника  
по эксплуатации

/Складова Е.Н.

« 14 » августа 2026 г.

Начальник отдела  
информационных технологий

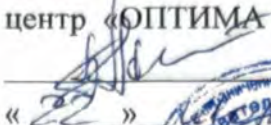
/Дудолов В.А.

« 14 » августа 2026 г.



УТВЕРЖДАЮ

Генеральный директор  
ООО «Международный образовательный  
центр «ОПТИМА СТАДИ»

 / Антипов П.Л.  
« 22 » \_\_\_\_\_ 2026 г.



АКТ

о внедрении результатов диссертационного исследования

Комиссия в составе генерального директора ООО «Международный образовательный центр «ОПТИМА СТАДИ» Антипова П.Л., заместителя директора Гуровой Г.П., руководителя отдела академических программ Худобердиной М.Г. составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

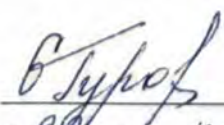
- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации, обеспечивающий формирование интерпретируемых результатов анализа на основе совместного использования графов знаний и графовых нейронных сетей;
- технология обработки и анализа типизированной мультимодальной информации, включающая извлечение признаков, построение графа знаний, нейросимвольный вывод и формирование интерпретируемых объяснений;
- информационная система обработки и анализа типизированной мультимодальной информации, реализующая полный цикл обработки данных и предоставления интерпретируемых результатов анализа,

внедрены в деятельность ООО «Международный образовательный центр «ОПТИМА СТАДИ».

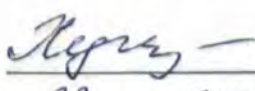
Разработанные результаты используются при анализе образовательных данных, формировании рекомендаций по персонализации образовательного процесса и поддержке принятия решений при сопровождении обучающихся.

Использование результатов диссертационного исследования способствует повышению эффективности анализа образовательных данных и обоснованности принимаемых решений по построению индивидуальных образовательных траекторий.

Заместитель директора

 / Гурова Г.П./  
« 22 » \_\_\_\_\_ июля 2026 г.

Руководитель отдела  
академических программ

 / Худобердина М.Г./  
« 22 » \_\_\_\_\_ июля 2026 г.





ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ УНИТАРНОЕ ПРЕДПРИЯТИЕ  
«РОССИЙСКАЯ ТЕЛЕВИЗИОННАЯ И РАДИОВЕЩАТЕЛЬНАЯ СЕТЬ»  
ФИЛИАЛ «САРАТОВСКИЙ ОБЛАСТНОЙ РАДИОТЕЛЕВИЗИОННЫЙ ПЕРЕДАЮЩИЙ ЦЕНТР»

ул. 2-ая Садовая, д. 7, г. Саратов, Россия, 410004, Тел.: (8452) 20-36-63, факс: (8452) 22-07-39

УТВЕРЖДАЮ

Директор филиала РТРС

Саратовский ОРТПЦ

Филиппов А.А.

«6» августа 2026 г.



АКТ

о внедрении результатов диссертационного исследования

Комиссия в составе директора филиала РТРС Саратовский ОРТПЦ Филиппова Александра Аркадьевича, ведущего специалиста по кадрам Гараниной Н.В. и ведущего инженера отдела информационных технологий Шуршалова Виталия Сергеевича составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в деятельность филиала РТРС Саратовский ОРТПЦ.

Разработанные результаты используются при анализе информации о кандидатах на вакантные должности, обработке данных из различных источников и формировании интерпретируемых рекомендаций для поддержки кадровых решений.

Использование результатов диссертационного исследования способствует повышению эффективности обработки кадровой информации и обоснованности принимаемых решений при подборе персонала.

Ведущий специалист по кадрам

Гаранина Н.В.

Ведущий специалист отдела ИТ

Шуршалов В.С.



Общество с ограниченной ответственностью

## «Самара-Сигнал»

ИНН/КПП 6311061575/631601001

Юридический адрес: 443079, г. Самара, ул. Гагарина, д.6А, литер Б5152Б

Почтовый адрес: 443079, г. Самара, ул. Гагарина, д.6А

Тел.: (846) 300-43-19, (846) 260-05-68; тел/факс: (846) 260-01-11

[www.samara-signal.ru](http://www.samara-signal.ru) E-mail: [post@samara-signal.ru](mailto:post@samara-signal.ru)

УТВЕРЖДАЮ

Управляющий ООО «Самара-Сигнал»

индивидуальный предприниматель

Неудахин С.Б.

« 27 » июля 2026 г.



### АКТ

#### о внедрении результатов диссертационного исследования

Комиссия в составе управляющего ООО «Самара-Сигнал» Неудахина С.Б., директора по персоналу Васильевой Н.Ю. и технического директора Краснова Д.А. составила настоящий акт о том, что результаты диссертационного исследования Ждановой Александры Николаевны по теме «Интерпретируемый анализ мультимодальной информации на основе нейросимвольной модели», выполненного на кафедре программных систем Самарского университета, а именно:

- нейросимвольный метод интерпретируемого анализа типизированной мультимодальной информации;
- технология обработки и анализа типизированной мультимодальной информации;
- информационная система обработки и анализа типизированной мультимодальной информации,

внедрены в деятельность ООО «Самара-Сигнал».

Разработанные результаты используются при анализе информации о кандидатах на вакантные должности, обработке данных из различных источников и формировании интерпретируемых рекомендаций для поддержки кадровых решений.

Использование результатов диссертационного исследования способствует повышению эффективности обработки кадровой информации и обоснованности принимаемых решений при подборе персонала.

Директор по персоналу

\_\_\_\_\_/Васильева Н.Ю.

« 27 » июля 2026г.

Технический директор

\_\_\_\_\_/Краснов Д.А.

« 27 » июля 2026г.