На правах рукописи



АЛЕКСЕЕВА Дарья Сергеевна

МОДЕЛЬ, АЛГОРИТМ И ПРОГРАММНОЕ СРЕДСТВО КВАЛИФИКАЦИИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ СБОРА ИНФОРМАЦИИ

Специальность 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

АВТОРЕФЕРАТ
Диссертации на соискание ученой степени кандидата технических наук

Уфа - 2026

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Уфимский университет науки и технологий (УУНиТ)».

Научный руководитель: доктор технических наук, доцент
Антонов Вячеслав Викторович

Официальные оппоненты:
Болодурина Ирина Павловна, доктор технических наук, профессор, ФГБОУ ВО «Оренбургский государственный университет», заведующий кафедрой прикладной математики
Бельтюков Анатолий Петрович, доктор физико-математических наук, профессор, ФГБОУ ВО «Удмуртский государственный университет», заведующий кафедрой вычислительных интеллектуальных систем

Ведущая организация: ФГАОУ ВО «Пермский национальный исследовательский политехнический университет», г. Пермь

Защита диссертации состоится 25.12.2026 в 10^{00} часов на заседании диссертационного совета 24.2.479.07, созданного на базе ФГБОУ ВО «Уфимский университет науки и технологий», по адресу: 450008, Республика Башкортостан, г. Уфа, ул. Карла Маркса, д.12

С диссертацией можно ознакомиться в библиотеке федерального государственного бюджетного образовательного учреждения высшего образования «Уфимский университет науки и технологий» и на сайте https://uust.ru/

Автореферат разослан «___» _____ 2026 г.

Ученый секретарь
диссертационного совета,
доктор технических наук А.М. Вульфин

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования.

В Стратегии развития информационного общества в Российской Федерации на 2017-2030 годы, где зафиксированы национальные цели и стратегические задачи развития страны на период до 2030 года, выделены специализированные целевые установки и задачи. Положения стратегии служат важным нормативным основанием для внедрения интеллектуальных информационно-управленческих технологий, а также технологий обработки больших данных в различных сферах деятельности на территории Российской Федерации.

В частности, в Указе Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» содержится положение, касающееся проектов, обеспечивающих достижение целевых ориентиров и плановых показателей деятельности федеральных органов исполнительной власти. На основании данного положения была сформулирована задача по созданию инструментария, позволяющего осуществлять анализ потоков неструктурированных данных. Для этого необходимо разработать программный прототип, представляющий из себя подключаемый модуль к системе автоматизированной обработки данных, служащий мостом между парсинговыми модулями и аналитическими модулями, на основании результатов которых принимаются решения.

В связи с изложенным **актуальность** диссертационного исследования продиктована необходимостью разработки гибридных методов анализа неструктурированных данных в потоковом режиме, обеспечивающих классификацию и содержательную квалификацию релевантной информации, что служит цели повышения эффективности информационной поддержки процессов принятия решений.

Степень разработанности темы.

Проблема квалификации неструктурированных текстовых данных исследована в работах по обработке естественного языка, машинному обучению и онтологическому моделированию. Теоретический фундамент составляют исследования формальных языков и автоматов (Н. Хомский, С. Клини), представления знаний (М. Минский, А. Вежбицкая), дистрибутивного анализа (З. Харрис), машинного обучения (наивный байесовский классификатор, метод опорных векторов), нейровероятностных языковых моделей (Й. Бенжио), векторных представлений слов (Word2Vec, GloVe) и трансформерных архитектур (А. Вашвани, BERT, GPT). Дополнительно развиты исследования в области стилометрии (Д. Берроуз), субсловных единиц (Р. Сенрих), атрибуции авторства и анализа социальных медиа.

Вместе с тем в большинстве анализируемых работ недостаточно проработаны вопросы семантической глубины, операционной эффективности и адаптивности к изменению структуры данных в условиях сохранения интерпретируемости. Следовательно, ключевая проблема заключается в

необходимости достижения равновесия между эффективностью, адаптивностью и семантической содержательностью.

Объектом исследования являются неструктурированные текстовые данные, извлекаемые из открытых интернет-источников, и процессы их автоматизированной обработки.

Предмет исследования: модели, методы и алгоритмы семантической квалификации контента, основанные на комбинации нейросетевых подходов и онтологического моделирования, направленные на повышение точности обработки данных в системах сбора информации.

Целью диссертационной работы является повышение точности классификации (квалификации) неструктурированных текстовых данных за счет разработки гибридного метода, интегрирующего контекстуальные векторные представления, тональный и стилометрический анализ, а также его программная реализация в виде модуля для систем сбора информации.

Задачи исследования:

Для достижения поставленной в диссертационной работе цели необходимо решить следующие задачи:

1. Разработать формальную модель квалификации неструктурированных данных (п. 1 паспорта специальности).
2. Разработать алгоритм квалификации контента интернет-источников на базе предложенной модели (п. 3 паспорта специальности).
3. Спроектировать программное решение системы квалификации и реализовать её ключевые модули (п. 3 паспорта специальности).
4. Провести внедрение предлагаемой модели и алгоритма квалификации в действующую систему автоматизированного сбора данных и экспериментальную оценку их эффективности с формированием рекомендаций по дальнейшему развитию (п. 10 паспорта специальности).

Методологическую основу работы составили методы нейросетевых технологий, а также статистические и онтологические подходы, что позволило обеспечить высокую точность и интерпретируемость результатов.

Достоверность результатов исследований.

Достоверность полученных научных результатов обеспечивается корректным использованием методов системного анализа, математического моделирования, общей теории и проектирования систем, а также методов математического, компьютерного и нейросетевого моделирования при обработке больших данных - на примере неструктурированных текстовых потоков, поступающих из различных интернет - источников. Вводимые допущения о вычислительной нагрузке мотивированы результатами проведённых экспериментальных исследований и практическими данными, подтверждёнными опытным путём.

Положения, выносимые на защиту.

1. Формальная модель квалификации неструктурированных данных, построенная на трехуровневой структуре «признак → категория → сет» и механизме множественной принадлежности семантических единиц. Модель

учитывает контекстную зависимость классификации в условиях полисемии (п. 1 паспорта специальности).

2. Алгоритм гибридной квалификации текстового контента, основанный на совместном применении контекстуализированных векторных представлений, тонального анализа и стилометрической метрики (дельта Берроуза). Такой подход позволяет выявлять аномальный контент и повышает точность категоризации (п. 3 паспорта специальности).

3. Программное решение, реализующее предложенный алгоритм. Программная структура поддерживает гибкие сценарии развертывания и включает механизм динамической маршрутизации потоков данных, обеспечивающий семантическую согласованность и масштабируемость (п. 3 паспорта специальности).

4. Результаты внедрения и экспериментальной апробации предлагаемого решения в составе промышленной информационной аналитической системы. Апробация подтверждает повышение точности квалификации контента и практическую значимость разработанных решений для задач мониторинга информационного пространства.

Научная новизна

1. Разработана формальная модель квалификации с иерархией «признак → категория → сет», реализующая вместо одного уровня принадлежности множественную принадлежность, то есть один объект может одновременно относиться к нескольким категориям. В отличие от классических онтологических моделей с жёсткой привязкой к одной категории, предложенная модель учитывает полисемию и контекстную зависимость смысла коротких зашумлённых сообщений. Модель позволяет формализовать контекстную многозначность и снизить число ошибок категоризации коротких сообщений, где смысл размыт и зависит от контекста.

2. Предложен подход, в котором комплексно объединены контекстуализированные векторные представления, тональный анализ и адаптирована стилометрическая метрика (дельта Берроуза) для коротких текстов социальных медиа. В отличие от существующих подходов, где стилометрические и семантические признаки объединяются на уровне решений, в предложенном алгоритме интеграция выполняется на уровне признакового пространства. Решение позволяет повысить точность квалифкациии неструктурированных данных и обеспечить выявление аномального контента при сохранении интерпретируемости решений.

3. Предложено использование порядковой классификации с обучаемыми порогами, позволяющий оценивать степень серьезности контента независимо от категориального решения. в отличие от традиционных категориальных моделей, которые дают только номинальную метку, предложенный подход раздельно и одновременно упорядоченно оценивает степень значимости контента относительно темы; комбинация категориальной классификации и порядковой регрессии позволяет получать как категорию, так и уровень тяжести, не смешивая эти задачи. Такой подход обеспечивает существенное согласие с

экспертами и позволяет приоритизировать обработку контента без переобучения категориальной модели.

Теоретическая значимость заключается в развитии теоретических основ автоматической классификации неструктурированных текстовых данных:

1. Предложена формальная модель квалификации, которая, в отличие от существующих подходов с плоскостной классификацией, вводит иерархическую структуру и механизм множественной принадлежности семантических единиц. Это развивает представления об учете контекстной зависимости и полисемии при категоризации коротких сообщений в динамичных потоках данных социальных медиа.

2. Разработан гибридный метод семантического анализа, теоретически обосновывающий комбинирование контекстуализированных векторных представлений (на базе архитектуры трансформер), тонального анализа и стилометрической метрики для повышения точности квалификации. Метод расширяет возможности классических статистических и нейросетевых подходов, обеспечивая выявление аномального контента на основе статистических отклонений частотных характеристик.

3. Предложена программная структура реализации с механизмом динамической адаптации маршрутизации потоков данных в зависимости от семантики и аномальности контента. Это вносит вклад в теоретические основы построения масштабируемых систем обработки потоковых данных.

Практическим результатом исследования является программное решение квалификации неструктурированных данных, реализующий разработанные методы и подтверждающий их эффективность. Результаты работы вносят вклад в развитие технологий обработки естественного языка за счет:

- повышения точности квалификации контента;
- ускорения анализа данных за счет распараллеливания;
- учета сложных семантических связей в тексте;
- улучшения контекстного и тонального анализа;
- создания более эффективных систем для понимания естественного языка.

Основные положения и результаты диссертационного исследования докладывались и обсуждались на следующих научных конференциях и семинарах:

– Мавлютовские чтения, Уфа, 2024

– XVII Всероссийская зимняя школа-семинар магистрантов, аспирантов и молодых ученых «Актуальные проблемы науки и техники», Уфа, 2024

– Мавлютовские чтения, Уфа, 2025;

– XVIII Всероссийская зимняя школа-семинар магистрантов, аспирантов и молодых ученых «Актуальные проблемы науки и техники, Уфа, 2025];

– научных семинарах кафедры Автоматизированных систем управления Уфимского университета науки и технологий (2024-2026 гг.).

Разработанный метод квалификации неструктурированных данных и программное решение, реализующий предложенные алгоритмы, внедрены и

используются в составе промышленной информационно-аналитической системы для выявления деструктивного контента, для модерации контента и в учебном процессе, что подтверждено соответствующими актами о внедрении.

Публикации. Результаты исследований получили отражение в публикациях автора в научных периодические изданиях, которые включены в Российский индекс научного цитирования. Наиболее значимые результаты, выносимые на защиту, получили отражение в следующих научных публикациях автора, в журналах входящих в перечень, утверждённый Высшей аттестационной комиссией. Также автором разработаны несколько программ для ЭВМ и схема баз данных, на которые получены свидетельства о государственной.

Личный вклад автора.

Автором диссертационного исследования непосредственно выполнен полный цикл исследования: формулировка положения и теоретических выводов, проектирование архитектуры, разработка алгоритмов, создание модели, проведение экспериментов и координация опытной эксплуатации модуля квалификации неструктурированных данных с последующим анализом его эффективности.

Связь исследований с научными программами. Исследование выполнялось при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках основной части государственного задания высшим учебным заведениям FEUE-2023-0007 и действующей №FRRR-2026-0006.

Структура и объём работы.

Диссертация объемом 200 машинописных страниц, содержит введение, пять глав и заключение, список литературы (147 наименований), 46 формул, 19 таблиц, 24 рисунка, 6 приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении дается характеристика проблематики квалификации больших текстовых данных, для которой требуются новые методы и алгоритмы. Проведенный первичный анализ показал, то множество используемых моделей и подходов не могут перекрыть очевидные проблемы при отсутствии гибридного подхода.

В первой главе проведено исследование фундаментальных основ семантического анализа данных. Семантический анализ сформировался на стыке лингвистики и математики, а ключевыми операциями обработки информации являются идентификация, структурирование и квалификация. Эволюция методов обработки естественного языка характеризуется переходом от статистических и классификационных подходов к нейросетевым и гибридным моделям. Словарные правила и шаблонные реакции не обеспечивают гибкой квалификации текста, что определяет необходимость моделей, удерживающих контекст и выявляющих скрытые семантические связи, на современном этапе эту функцию выполняют нейросетевые архитектуры. Для задачи квалификации

потоковых данных из интернет-источников ни один из существующих подходов в отдельности не является достаточным: пользовательский контент социальных медиа характеризуется высокой плотностью лингвистического шума, полисемией коротких сообщений, быстрой эволюцией языка, культурной и субкультурной спецификой. На основе выявленных ограничений и перспективных направлений в диссертационной работе поставлены задачи по разработке формальной модели квалификации с иерархией «признак → категория → сет» и механизмом множественной принадлежности семантических единиц, созданию гибридного алгоритма на основе трансформерных архитектур, тонального анализа и стилометрических метрик, проектированию программной структуры для обработки данных в режиме, близком к реальному времени, а также внедрению решения в действующую систему автоматизированного сбора данных.

Во второй главе разработана формальная модель квалификации неструктурированных текстовых данных, задаваемая кортежем

$M=<W,C,R,\Omega,\Phi>$, где: $W=\{w_1,w_2,...,w_n\}$ - множество семантических единиц; $C=\{c_1,c_2,...,c_k\}$ - множество категорий, к которым могут быть отнесены семантические единицы; $R=\{r_1,r_2,...,r_m\}$ - множество отношений между категориями и единицами (иерархические, ассоциативные, логические); Ω - онтологическая спецификация, задающая ограничения на допустимые комбинации категорий и признаков; $\Phi: W \to \Delta(C)$ - функция квалификации, отображающая множество семантических единиц в пространство категорий, где $\Delta(C)$- множество распределений вероятностей на множестве категорий C, отражающее принцип многопринадлежности. Функция может быть реализована в двух режимах: одноместочный, используется, когда онтология требует единственной категории и многометочный - каждая категория получает вероятность независимо. Формирование сета выполняется согласно следующими правилу: $Set = f\ (Cat, \Omega)$, где Cat - множество категорий, Ω - онтологическая спецификация. Сет определяется как максимальный класс онтологии, покрывающий все предсказанные категории и не противоречащий аксиомам.

Модель реализует механизм многопринадлежности через вектор вероятностей, трёхуровневую структуру «признак → категория → сет» (рисунок 1).

На первом уровне (признаковый анализ) исходный текст преобразуется в структурированное числовое представление: формируются эмбеддинги токенов, частотные профили, тональная разметка и синтаксические деревья; синтаксический анализ строит деревья зависимостей, частотный профиль формируется с использованием BPE-токенизации, тональный анализ сочетает словарный подход и нейросетевой классификатор. На втором уровне (семантический анализ) вводится контекст и смысловые связи: строятся синтаксические зависимости и семантическая сеть, вычисляется семантическая близость через косинусное сходство, применяется модифицированная дельта

Берроуза, адаптированная для коротких текстов за счёт BPE-токенизации и включения в интегральный вектор признаков.

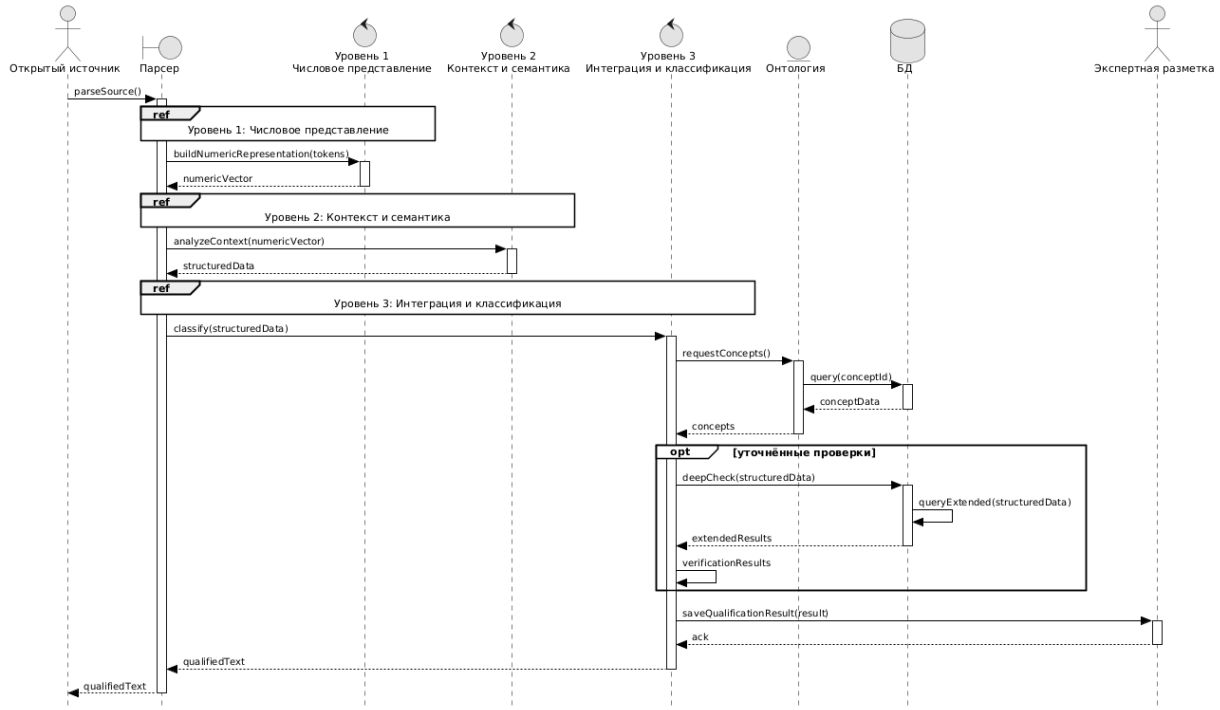


Рисунок 1 - Предлагаемая структура гибридного метода

На третьем уровне (интеграция и классификация) используется свёрточная нейронная сеть (CNN), на вход которой подаются признаки, извлечённые на предыдущих уровнях; выходной слой формирует распределение вероятностей по категориям (softmax в одноместочном режиме, независимые сигмоиды - в многометочном). Онтологическая спецификация обеспечивает интерпретируемость решений: для каждого предсказания формируется объяснение - набор онтологических отношений, графовых связей и ключевых маркеров; противоречие аксиомам запускает углублённую проверку или экспертную разметку. Накопление несоответствий между предсказаниями и онтологическими ограничениями служит индикатором концептуального дрейфа и инициирует процедуру обновления модели и онтологии.

В третьей главе разработан алгоритм и математический аппарат квалификации неструктурированных текстовых данных, обеспечивающий полный цикл обработки - от токенизации до формирования сета. Математическое представление алгоритма квалификации отличается комплексным объединением контекстуализированных векторных представлений, тонального анализа и модифицированной дельты Берроуза на уровне признакового пространства, а также использованием порядковой регрессии с обучаемыми порогами.

Первый этап - преобразование текста в последовательность чисел, понятную модели. Применяется BPE-токенизация: текст разбивается на подсловные единицы - токены, каждая из которых получает идентификатор из словаря. Это позволяет работать с редкими словами, обеспечивает кросс-лингвистичность и устойчивость к опечаткам, поэтому BPE стал стандартом в

GPT и BERT. Каждому токену сопоставляется плотный вектор - эмбеддинг; последовательность идентификаторов преобразуется в матрицу эмбеддингов, и текст переходит в непрерывное векторное пространство, где близкие по смыслу токены располагаются рядом. Далее извлекаются базовые признаки. Синтаксические - одномерной свёрточной сетью, обрабатывающей эмбеддинги фильтрами разных размеров, такая архитектура обеспечивает высокую скорость ответа и низкую задержку. Семантические - методом латентного размещения Дирихле, представляющим текст как распределение вероятностей над скрытыми темами. Средняя лексическая тональность вычисляется усреднением оценок отдельных токенов, а итоговая - softmax-нормализацией, дающей распределение по классам «позитив», «нейтрально», «негатив».

Параллельно производится количественная оценка степени аномальности анализируемого текста относительно эталонных текстов, принадлежащих к определенному сету. В качестве меры аномальности используется дельта Берроуза - метрика, изначально разработанная для задач стилометрии и проверки авторства, но адаптированная в настоящем исследовании для выявления потенциальных вбросов в рамках категории:

$$\Delta_B = \frac{1}{|V_{pr}|}\sum_{i\in V_{pr}}^{N}\left|\frac{f_i-\mu_i}{\sigma_i}\right|,$$

где: Δ_B - дельта Берроуза; V_{pr} - множество токенов BPE-словаря, присутствующих в анализируемом тексте; $|V_{pr}|$ - число присутствующих токенов; μ_i - средняя частота токена в стандартных текстах; f_i - частота токена в анализируемом массиве; σ_i - стандартное отклонение токена.

Значение дельты Берроуза приводится к диапазону $[0,1]$ сигмоидным преобразованием с обучаемым параметром масштаба γ, после чего нормированная аномальность δ используется как компонент интегрального вектора признаков и как регуляризующий множитель, усиливающий признаки аномальных текстов.

Далее вычисляется косинусное расстояние между усреднёнными эмбеддингами первой и последней третей сообщения - чем выше значение, тем сильнее различаются начало и конец текста. Такая адаптация позволяет улавливать динамику тональности внутри текста, характерную для саркастических высказываний, которые часто строятся на контрасте: позитивное начало и негативная концовка (или наоборот).

$$D_{parts} = 1-\cos(Emb_1,...Emb_n)\in[0,2],$$

где: $Emb_1 = \frac{1}{k}\sum_{i=1}^{k} X_i$ - усреднённый эмбеддинг первых k токенов (обычно k=⌊n/3⌋);

$Emb_n = \frac{1}{k}\sum_{i=n-k+1}^{n} X_i$ - усреднённый эмбеддинг последних k токенов; n - длина последовательности; X_i - эмбеддинг токена; $\cos(u,v) = \frac{u*v}{\|u\|*\|v\|}$ - косинусное сходство.

Для оценки семантического противоречия между темой текста и тем, что о ней говорится, применяется разметка, позволяющая извлечь именные группы -

они и представляют тему сообщения. Для каждой именной группы вычисляется средняя тональность входящих слов, а параллельно - тональность слов, не входящих в именные группы, то есть того, что о теме говорится. Сопоставление этих двух показателей позволяет выявить семантический диссонанс, характерный для сарказма, - например, когда о нейтральной теме говорится в резко негативном или подчёркнуто позитивном ключе.

Помимо базовых признаков, для детекции сарказма вычисляются четыре дополнительные лингвистические метрики, адаптированные для коротких текстов социальных медиа.

Сопоставление тональности темы и тональности контекста позволяет выявить семантический диссонанс, характерный для сарказма, - например, когда о нейтральной теме говорится в резко негативном или подчёркнуто позитивном ключе.

$$J_{sem} = | sent_{topic} - sent_{context} | \in [0,2],$$

где: $sent_{topic}$ - тональность темы (о чём текст); $sent_{context}$ - тональность контекста (что о теме говорят).

Далее необходимо вычислить коэффициент эмоциональной интенсивности, так как сарказм часто сопровождается резкими перепадами эмоционального тона [92].

$$\varepsilon_{\text{int}} = \frac{1}{n-1}\sum_{i=1}^{n-1} | emo(t_i) - emo(t_{i+1}) | * \mathbb{I}[emo(t_i) - emo(t_{i+1}) | > \tau_e],$$

где: $emo(t_i) \in [0,1]$ - интенсивность эмоции токена; $\tau_e \in [0,1]$ - порог чувствительности (обучаемый параметр); $\mathbb{I}[\cdot]$ - функция принадлежности элемента к подмножеству.

Кроме того, необходима оценка преувеличений, характерных для сарказма:

$$H_{hyp} = \frac{1}{n}\sum_{i=1}^{n} hyper(t_i) * \mathbb{I}[hyper(t_i) > 0] \in [0,1],$$

где: $hyper(t_i)$ - вес гиперболичности токена; n - число токенов.

Стилистический диссонанс - отклонение стиля текста от эталонного для данной темы. Саркастические тексты часто имеют аномальный стиль по сравнению с «нормальными» текстами на ту же тему:

$$S_{dis} = \frac{\| H_{syn} - H_{syn}^{base}(k) \|}{\| H_{syn}^{base}(k) \| + \varepsilon} \in [0, \infty),$$

где: $H_{\text{syn}}^{\text{base}}(k)$ - усредненный синтаксический профиль текстов темы k в обучающей выборке; $\|\cdot\|$ - евклидова норма; $\varepsilon = 10^{-8}$ - малая константа, предотвращающая деление на ноль.

Для определения важности токенов при выявлении сарказма применяется механизм масштабированного скалярного произведения: формируется запрос, агрегирующий все признаки сарказма, для каждого токена вычисляются ключ и значение, а сопоставление запроса с ключами даёт веса внимания, нормируемые softmax. На основе весов формируется вектор контекстной противоречивости -

взвешенная сумма значений токенов, аккумулирующая вклад фрагментов текста в ощущение противоречивости. Последний шаг заключается в преобразовании этого вектора в скалярную оценку сарказма с помощью линейного слоя с сигмоидой, полученный скаляр ρ используется как самостоятельный признак и как компонент интегрального вектора при категоризации и оценке тяжести.

Вычисляется коэффициент контекстного диссонанса, адаптированный под контент социальных медиа:

$$C_{dis} = 1 - \cos(SP_{current}, SP_{context}) \in [0,2],$$

где: $SP_{current}$ - семантическое пространство текущего сообщения; $SP_{context} = \frac{1}{m}\sum_{j=1}^{m} SP_j$ - усредненное семантическое пространство предыдущих m сообщений в диалоге; m - размер окна контекста (гиперпараметр, обычно $m = 3 - 5$).

Семантическое пространство текста - его интегральная смысловая характеристика, получаемая как взвешенная сумма эмбеддингов токенов, где вес каждого токена отражает его значимость для представления всего текста. Полученное пространство сопоставляется с эталонным с помощью коэффициента конвергенции - косинусного сходства, показывающего, насколько текст типичен для данной предметной области. Для сопоставимости с другими компонентами модели семантическое пространство подвергается Z-стандартизации, что приводит компоненты к единому масштабу и обеспечивает численную устойчивость вычислений.

Интегральный вектор признаков формируется путём конкатенации всех извлечённых компонент. В отличие от известных подходов, где признаки объединяются на уровне решений, здесь интеграция выполняется на уровне признакового пространства, что позволяет учитывать корреляции между стилем, тональностью и семантикой уже на этапе обучения.

$$\Pr iz_{sarc}^{new} = [H_{syn} \oplus H_{sem} \oplus CCV \oplus \rho \oplus J_{sem} \oplus E_{\text{int}} \oplus D_{parts} \oplus H_{hyp} \oplus S_{dis} \oplus C_{dis}] * (1 + \lambda * \delta) \in R^{D_{sarc}},$$

где: D_{sarc} - полная размерность; $\lambda \in \mathbb{R}$ - обучаемый параметр регуляризации аномальностью. При $\lambda = 0$ аномальность игнорируется; при $\lambda > 0$ аномальные тексты усиливают свои признаки; δ - нормированная аномальность; $\bigoplus$ - операция конкатенации; J_{sem} - индекс семантической инконгруэнтности; D_{parts} - косинусное расстояние между частями текста.

Все компоненты интегрального вектора признаков центрируются и масштабируются по статистикам обучающей выборки, сохранённым в артефактах модели, что исключает зависимость результата от состава текущего пакета и обеспечивает корректную работу для одиночных объектов. Нормализованный вектор подаётся на линейный классификатор с softmax-нормализацией, формирующий распределение вероятностей по категориям; в одноточном режиме итоговая категория определяется как класс с максимальной вероятностью. Параллельно вычисляется скалярный коэффициент уверенности на основе конвергенции, который используется для масштабирования лингвистической оценки категорий и отражает степень типичности текста для предметной области.

Нормализованное семантическое пространство масштабируется коэффициентом уверенности к, после чего проецируется в пространство категорий с помощью обучаемой матрицы. Поэлементное умножение на к позволяет динамически регулировать вклад семантической информации: для типичных текстов вклад семантики выше, для аномальных - ниже. На выходе получается лингвистическая векторная оценка категорий, которая в дальнейшем комбинируется с нейросетевой оценкой при многозадачном обучении. Новизна решения состоит в динамическом регулировании вклада семантической информации: для типичных текстов вклад семантики выше, для аномальных ниже:

$$ling_{vect} = W_{proj} * (SP_{norm} \odot \kappa) \in R^C,$$

где: SP_{norm} - нормализованное семантическое пространство; $ling_{vect}$ - лингвистическая оценка категории; $\odot$ - поэлементное умножение; κ - скаляр, транслируемый на все размерности SP_{norm}, $W_{proj} \in \mathbb{R}^{C \times d}$- матрица проекции размерности.

Итоговая комбинация формируется как взвешенная сумма трёх компонент - нейросетевой оценки категорий, лингвистической векторной оценки и скаляра саркастичности, умноженного на обучаемый вектор весов влияния по каждой категории, что является отличительной чертой. Такая тройная композиция позволяет учитывать, как статистические закономерности, выявленные нейросетью, так и семантическую близость текста к эталонным представлениям, а также вклад саркастичности - причём её влияние зависит от категории, а не является одинаковым для всех классов:

$$combined_{out} = \alpha' * neural_{out} + \beta' * ling_{vect} + \gamma' * (\rho \odot w_\rho)_c,$$

где: $wp_p \in R^C$ - обучаемый вектор весов влияния саркастичности на каждую категорию; ρ - скаляр саркастичности, транслируемый на все размерности.

Чтобы объединить нейросетевую и лингвистическую оценки, вводятся адаптивные веса. Три обучаемых параметра α', β' и γ' пропускаются через softmax-нормализацию, в результате чего получаются нормированные веса α', β' и γ', сумма которых равна единице. Такое нормирование позволяет модели динамически перераспределять вклад каждой компоненты в зависимости от их полезности для текущей задачи, не нарушая баланса итогового решения.

Полученная комбинация преобразуется в распределение вероятностей по категориям с помощью softmax - в одноместочном режиме все вероятности положительны и в сумме дают единицу, что обеспечивает корректную интерпретацию результата. Итоговая категория определяется как класс с максимальной вероятностью, при этом для многометочного режима вместо softmax используется сигмоида, позволяющая каждой категории получать независимую вероятность и одновременно относить объект к нескольким классам.

После получения распределения вероятностей по категориям выполняется формирование сета - верхнего уровня иерархии «признак → категория → сет».

Сет S допустим для множества категорий Cat, если он покрывает все предсказанные категории и не противоречит аксиомам и правилам онтологии:

$$Adm(S, Cat, \Omega) = \mathbf{1}[Cat \subseteq S.cat] * \mathbf{1}[\forall \alpha \in A: \alpha(S, Cat) = true] * \mathbf{1}[\forall \theta \in \Theta: \theta(S, Cat) = true],$$

где: $\mathbf{1}[\cdot]$ - индикаторная функция (1 - истина, 0 - ложь), S - сет (набор категорий/понятий онтологии), Cat - множество предсказанных категорий, $\Omega = (A, \Theta)$ - онтология, где A - множество аксиом, Θ - множество правил, $S.cat$- множество категорий, покрываемых сетом, $\alpha(S, Cat)$ - проверка аксиомы для сета и категорий, $\theta(S, Cat)$ - проверка правила для сета и категорий.

Среди всех допустимых сетов выбирается максимальный, который покрывает наибольшее число предсказанных категорий:

$$Set = arg \max_{S \in Sets} \sum_{c \in Cat} 1[c \in S.cat] * Adm(S, Cat, \Omega),$$

где: $Sets$ - множество всех рассматриваемых сетов (наборов категорий онтологии), Cat - множество предсказанных категорий, $S.cat$- множество категорий, покрываемых сетом, $\mathbf{1}[\cdot]$ - индикаторная функция (1 - истина, 0 - ложь), $\Omega = (A, \Theta)$ - онтология, где A - множество аксиом, Θ - множество правил, $Adm(S, Cat, \Omega)$ - предикат допустимости сета.

Если несколько сетов имеют одинаковую мощность, выбирается тот, у которого выше суммарная вероятность покрытых категорий:

$$Set = arg \max_{S \in Sets} \sum_{c \in Cat \cap S.cat} cat_prob[c] * Adm(S, Cat, \Omega),$$

где: $Sets$ - множество всех рассматриваемых сетов (наборов категорий онтологии), Cat - множество предсказанных категорий, $S.cat$- множество категорий, покрываемых сетом, $c \in Cat \cap S.cat$ - категории, которые и предсказаны**,** и покрыты сетом, $cat_prob[c]$ - вероятность категории, $\Omega = (A, \Theta)$ - онтология, где A - множество аксиом, Θ - множество правил, $Adm(S, Cat, \Omega)$ - предикат допустимости сета.

Уверенность сета определяется как минимум вероятностей категорий, вошедших в сет:

$$conf(Set) = \min_{c \in Cat \cap S.cat} cat_prob[c],$$

где: Set - выбранный сет, Cat - множество предсказанных категорий, $S.cat$- множество категорий, покрываемых сетом, $c \in Cat \cap S.cat$ - категории, которые и предсказаны, и покрыты сетом, $cat_prob[c]$ - вероятность категории, $conf(Set)$ - уверенность сета (минимальная вероятность среди покрытых предсказанных категорий).

Если ни один сет не является допустимым формируется сет «Прочее»:

$$\forall S \in Sets: Adm(S, Cat, \Omega)=0 ,$$

где: Set - выбранный сет, Cat - множество предсказанных категорий, $\Omega = (A, \Theta)$ - онтология, где A - множество аксиом, Θ - множество правил, $Adm(S, Cat, \Omega)$ - предикат допустимости сета.

Отличительной чертой является то что переход от категорий к сету можно вычислить, а не просто объявить: аксиомы и правила проверяются отдельно, результат получается бинарным, а если подходящего сета нет — формируется класс «Прочее» с пометкой «Требует подтверждения эксперта».

Параллельно с категоризацией выполняется бинарная классификация сарказма. Полученные параметры обучаются совместно с категориальной

головой в режиме многозадачного обучения, что позволяет модели использовать общие представления для обеих задач и повышает качество каждой из них. Полученный бинарный флаг сарказма в дальнейшем используется при оценке степени тяжести и при интерпретации решений.

Контент получает оценку степени тяжести от 0 до 4 независимо от категории, что является отличительной чертой. Вектор для этой оценки строится из сочетания признаков:

$$sev = \text{Priz}_{clean} \oplus H_{ton} \oplus \delta \oplus \sigma(K) \oplus SP_{norm} \oplus \log(1+\Delta_B) \oplus J_{sem} \oplus \varepsilon_{\text{int}} \oplus \rho \in R^{D_{sev}},$$

где: $D_{\text{sev}} = (h_{\text{syn}} + h_{\text{sem}})$ - полная размерность; $\log(1 + \Delta_B)$ - сглаживание для подавления выбросов при больших частотах; $\mathcal{I}_{\text{sem}}$ - индекс семантической сходимости; ρ - скаляр саркастичности.

После формирования вектора признаков непрерывная скалярная оценка s вычисляется линейным слоем, отражая степень серьёзности контента. Для перевода оценки в пять упорядоченных уровней (0-4) используется порядковая регрессия с четырьмя обучаемыми порогами $\theta_1 < \theta_2 < \theta_3 < \theta_4$, а логиты вычисляются как отрицательный квадрат разности между оценкой и порогом, что обеспечивает плавное распределение вероятностей. Полный вектор логитов нормализуется softmax, а итоговый уровень определяется как индекс максимального элемента, такая схема сохраняет порядковую природу шкалы, при которой уровень 3 всегда «тяжелее» уровня 2.

Одним из этапов обработки является адаптация признаков для коротких сообщений. Для коротких сообщений (до n_0 токенов) увеличивается вес локальных признаков и уменьшается вес глобальных:

$$\text{Priz}_{short} = [\alpha(n) * Pri\,z_L || \beta(n) * Pri\,z_G],$$

где: $Priz_L$ - локальные признаки; $Priz_G$ - глобальные признаки; $\alpha(n) = \sigma\left(\frac{n-n_0}{\tau}\right)$ - функция адаптации локальных признаков; $\beta(n) = 1 - \alpha(n)$ - функция адаптации глобальных признаков; n - длина сообщения (число токенов); $n_0 = 20$ - порог короткого сообщения; $\tau = 10$ - гиперпараметр; σ - сигмоидная функция.

В четвертой главе представлена практическая реализация разработанного метода семантического анализа в виде библиотеки и ее интеграции в промышленную парсинговую систему. Существующий механизм квалификации основан на словарном подходе и характеризуется отсутствием контекстного и тонального анализа, высокой зависимостью от ручного контроля, низкой точностью при полисемии и ограниченной масштабируемостью. Разработана модульная программная структура в виде подключаемой библиотеки Python, рассмотрены преимущества библиотечного подхода и его ограничения, компенсируемые многоуровневым тестированием и мониторингом; определены этапы внедрения (анализ, рефакторинг, интеграция, верификация). Разработан механизм HQP, реализующий пять уровней обработки запроса: валидацию, семантическое обогащение, каноническую трансформацию, квалификацию и форматирование.

Диаграмма (Рисунок 2) отражает жизненный цикл одного запроса квалификации. Сначала запрос проходит три проверки - доступна ли библиотека,

есть ли готовый ответ в кэше и разрешён ли вызов через «CircuitBreaker». Если проверки пройдены, запускается обработка из пяти уровней: проверка данных, обогащение текста, приведение к единому формату, сама квалификация и формирование ответа. Запрос может завершиться одним из трёх способов: успешный ответ (SUCCESS), резервный ответ при недоступности библиотеки (FALLBACK) или ошибка на любом из пяти уровней (ERROR).

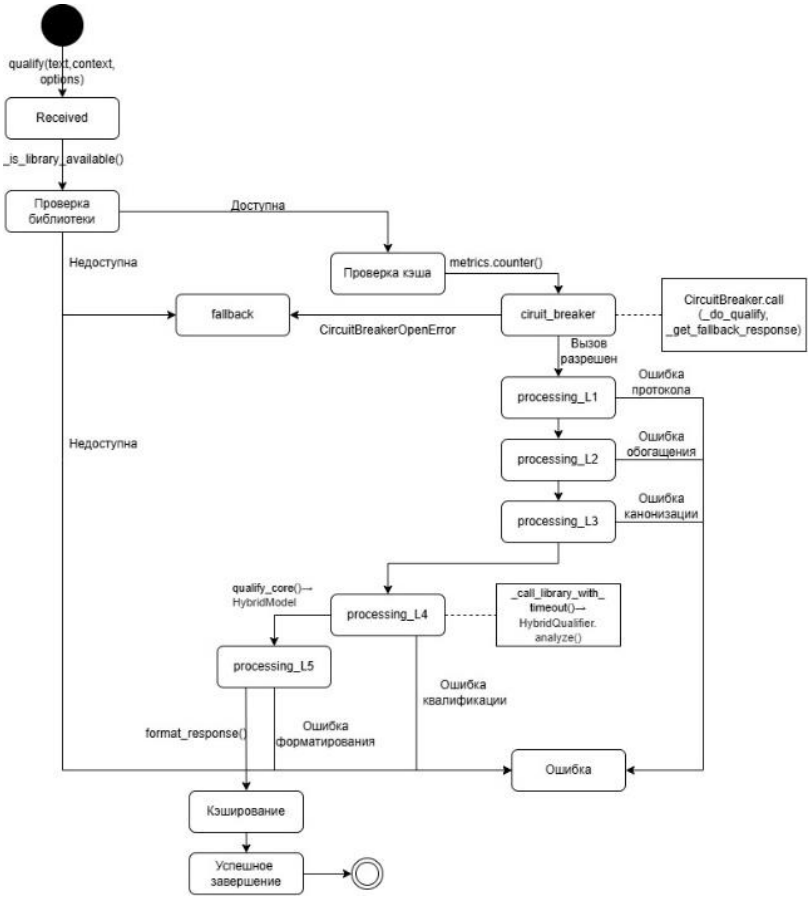


Рисунок 2 - Диаграмма состояний обработки запроса

В пятой главе проведена экспериментальная оценка эффективности метода квалификации. Экспериментальная оценка выполнена на трёх корпусах - специализированном, синтетическом и Kaggle Customer Support; сравнение с дообученной DistilBERT показало прирост точности на 5,40% на специализированном корпусе, +0,57%. на синтетическую и сопоставимую точность на Kaggle (22,58%). Установлено, что синергетический эффект достигается за счёт взаимодействия трёх групп методов: статистические извлекают частотные характеристики, нейросетевые предоставляют контекстуализированные представления, лингвистические обеспечивают интерпретируемость. Отличительной особенностью программной реализации является модульная структура подключаемой библиотеки с пятиуровневой обработкой запроса и специализированным протоколом взаимодействия. Полученные результаты образуют основу для практического применения метода в системах мониторинга информационного пространства.

Таким образом, исходя из теоретических результатов, реализована модель, представленная в главе 2, на ее основе реализованы математическое

представление и алгоритм, представленные в главе 3, в главе 4 проведена апробация, в главе 5 - оценка применимости предлагаемого решения.

ОСНОВНЫЕ ВЫВОДЫ И РЕЗУЛЬТАТЫ РАБОТЫ

В диссертационной работе решена научно-практическая задача повышения точности квалификации неструктурированных текстовых данных за счёт разработки гибридного метода, интегрирующего контекстуальные векторные представления, тональный и стилометрический анализ.

1. Анализ предметной области показал, что существующие подходы к семантическому анализу - символьные, статистические, нейросетевые - обладают взаимодополняющими ограничениями, а пользовательский контент социальных медиа характеризуется лингвистическим шумом, полисемией, быстрой эволюцией языка и субкультурной спецификой; это определило необходимость перехода к гибридным решениям.

2. Разработана формальная модель квалификации неструктурированных текстовых данных, задаваемая кортежем $M=<W,C,R,\Omega,\Phi>$, реализующая механизм многопринадлежности через вектор вероятностей, два режима категоризации (однометочный и многометочный) и правило формирования сета $Set = f(Cat,\Omega)$.

3. Разработана трёхуровневая структура гибридного метода квалификации: первый уровень (признаковый анализ) извлекает количественные характеристики; второй уровень (семантический анализ) вводит контекст, строит семантическую сеть и вычисляет дельту Берроуза; третий уровень (интеграция и классификация) принимает итоговое решение на основе свёрточной сети и сопоставляет результат с онтологическими классами.

4. Разработан математический аппарат, формализующий все этапы обработки - от BPE-токенизации до формирования сета. Предложенное математическое представление алгоритма отличается комплексным объединением контекстуализированных векторных представлений, тонального анализа и модифицированной дельты Берроуза на уровне признакового пространства, а также использованием порядковой регрессии с обучаемыми порогами для раздельной оценки степени тяжести.

5. Разработана модульная программная структура в виде подключаемой библиотеки Python и механизм HQP, обеспечивающий взаимодействие между платформой сбора данных и модулем квалификации и реализующий пять уровней обработки запроса (валидация, семантическое обогащение, каноническая трансформация, квалификация, форматирование).

6. Экспериментальная оценка выполнена на трёх корпусах: специализированном, синтетическом и Kaggle Customer Support. Сравнение с дообученной DistilBERT показало прирост точности на 5,40%. на специализированном корпусе, +0,5%. на синтетическую и сопоставимую точность на Kaggle (22,58%). Установлено, что синергетический эффект достигается за счёт взаимодействия статистических, нейросетевых и лингвистических методов.

Перспективы дальнейшей разработки темы. В рамках дальнейших исследований планируется добавить мультимодальный анализ, объединяющий визуальные и аудиальные признаки, для расширения набора признаков анализируемого контента. Кроме того, предполагается реализовать визуализацию внутренних представлений нейросетей, расширить работу на другие языки и внедрить автоматизированный подбор гиперпараметров с использованием байесовской оптимизации а также постепенный переход к обучению с подкреплением.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных изданиях из Перечня рецензируемых научных изданий, рекомендованных ВАК

1. **Алексеева Д. С**., Кононов Н. А., Антонов В. В., Еникеев Р. Р. Алгоритм анализа сетевого трафика в многокомпонентных структурно сложных информационных системах // Вестник Уфимского государственного авиационного технического университета, - 2024. - Т. 28, № 1(103). - С. 92-98. - DOI 10.54708/19926502_2024_28110392.

2. **Алексеева Д. С.**, Пальчевский Е. В., Антонов В. В., Суворова В. А. Оптимизация математической основы IRT с помощью LLM-моделей // Системная инженерия и информационные технологии, - Уфа, - 2026. - Т. 8, № 1(25). - С. 3-18. - DOI: https://doi.org/10.54708/SIIT-2026-no1-p3.

3. **Алексеева Д. С.** Мультипараметрический метод установления авторства текста: интеграция стилометрического, тематического и прагматического анализа // Системная инженерия и информационные технологии, - Уфа, - 2026. - Т. 8, № 2(26). - С. 75-83. - DOI: https://doi.org/10.54708/SIIT-2026-no2-p75.

4. **Алексеева Д. С**., Юлина Е. В., Кононов Н. А., Антонов В. В., Дьячков А. С. Разработка модуля автоматизированной предобработки текстовых данных с адаптивным управлением для нейросетевых моделей // Вестник Уфимского государственного авиационного технического университета, - 2026. - Т. 30, № 2(112). - С. 41-48. - DOI 10.54708/19926502_2026_30211241.

Иные публикации автора

5. **Алексеева Д. С**., Кононов Н. А. Криптография средствами офисного пакета MS Excel: возможности и оценка реализации // World science: problems and innovations: сборник статей LXVII Международной научно-практической конференции. - Пенза, 2022. - С. 36-39.

6. **Алексеева Д. С**. Обзор юридического сопровождения средств криптографической защиты // Мавлютовские чтения: материалы XVI Всероссийской молодежной научной конференции. В 6-ти томах. - Уфа, 2022. - С. 451-455.

7. **Алексеева Д. С.**, Кононов Н. А. Разработка и оценка реализации алгоритма шифрования RSA с применением языка Visual Basic for Applications // Молодежный вестник УГАТУ. - 2023. - № 1 (27). - С. 10-14.

8. **Алексеева Д. С**., Антонов В. В., Харисова З. И., Ишмеева А. С. Возможности применения средств видеофиксации с обработкой данных на основе искусственного интеллекта при производстве следственных действий // IV Международная научная конференция по междисциплинарным исследованиям: сборник статей. - Екатеринбург, 2023. - С. 374-380.

9. **Алексеева Д. С**., Кононов Н. А. Подход к автоматизации процесса выбора программного средства защиты информации // Молодежный вестник УГАТУ. - 2024. - № 1 (30). - С. 5-9.

10. **Алексеева Д. С**., Суворова В. А. Метод обеспечения защиты системы автоматизированного сбора данных из открытых источников от встречного воздействия исследуемого объекта // Информационные технологии интеллектуальной поддержки принятия решений (памяти проф. Н. И. Юсуповой) ITIDS'2024: труды X Международной научной конференции. В 2-х томах. - Уфа, 2024. - С. 284-289.

11. **Алексеева Д. С**., Нуриахметов А. И., Юлина Е. В. Интеллектуальные технологии автоматизированного сбора и анализа данных в технических приложениях: архитектура, методы, перспективы // Лучшие научные исследования студентов и учащихся: сборник статей XI Международной научно-практической конференции. - Пенза, 2025. - С. 78-83.

Патенты и свидетельства о регистрации программ для ЭВМ

12. **Алексеева Д. С.** Свидетельство о государственной регистрации программы для ЭВМ № 2023661529 «Программа автоматизации деятельности торгового агента» // Зарег. 24.05.2023. - М.: Роспатент, 2023.

13. **Алексеева Д. С.** Свидетельство о государственной регистрации программы для ЭВМ № 2023661948 «Имитационная модель средства ирригации» // Зарег. 24.05.2023. - М.: Роспатент, 2023.

14. Нуриахметов А. И., **Алексеева Д. С.,** Еникеев Р. Р. Свидетельство о государственной регистрации базы данных № 2025623973 «База данных автоматизированного тестирования программных модулей» // Зарег. 07.08.2025. - М.: Роспатент, 2025.

Диссертант  Д.С. Алексеева