РАЗРАБОТКА АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ КЛАССИФИКАЦИИ ТЕКСТОВЫХ СООБЩЕНИЙ ПО ТОНАЛЬНОСТИ И СЕМАНТИЧЕСКИМ КАТЕГОРИЯМ НА ОСНОВЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Абдукаримов Абдужалолиддин
магистрант направления «Data Science», Tashkent International University of Education» (TIUE), г. Ташкент, Республика Узбекистан

Аннотация: *В статье рассматривается разработка алгоритмов искусственного интеллекта для автоматической классификации текстовых сообщений по тональности и семантическим категориям на основе методов обработки естественного языка (NLP). Описывается полный цикл построения модели: сбор и разметка данных, предобработка текста, извлечение признаков и обучение классификаторов, а также оценка качества по метрикам точности, полноты, F1-меры и ROC-AUC. Отдельное внимание уделяется сравнению традиционных подходов (TF-IDF, SVM, логистическая регрессия) и современных нейросетевых моделей на базе трансформеров, способных учитывать контекст и смысл высказывания.*

Ключевые слова: *искусственный интеллект, обработка естественного языка (NLP), анализ тональности, семантическая классификация.*

В условиях стремительного развития цифровых технологий и глобализации информационного пространства объем текстовых данных, генерируемых пользователями, организациями и автоматизированными системами, увеличивается с экспоненциальной скоростью. Социальные сети, онлайн-платформы, службы клиентской поддержки, электронные СМИ и мессенджеры ежедневно создают миллионы текстовых сообщений, отражающих общественные настроения, мнения потребителей и различные социальные процессы. В этой связи особую актуальность приобретает задача автоматизированного анализа текстовой информации с использованием алгоритмов искусственного интеллекта.

Одним из ключевых направлений интеллектуальной обработки текстов является классификация сообщений по тональности (позитивная, негативная, нейтральная) и семантическим категориям (тематика, тип обращения, содержание проблемы и др.). Анализ тональности позволяет выявлять эмоциональную окраску высказываний, оценивать репутацию компаний и брендов, а также отслеживать общественные настроения. Семантическая классификация, в свою очередь, обеспечивает структурирование больших массивов текстовых данных, что существенно упрощает их последующий анализ и принятие управленческих решений.

Современные методы обработки естественного языка (Natural Language Processing, NLP) предоставляют широкий спектр инструментов для решения

указанных задач. К традиционным подходам относятся статистические модели и алгоритмы машинного обучения, основанные на извлечении признаков (например, TF-IDF) и использовании классификаторов, таких как логистическая регрессия, метод опорных векторов (SVM) и наивный байесовский классификатор. Однако в последние годы наибольшую эффективность демонстрируют нейросетевые модели глубокого обучения, в частности архитектуры трансформеров (BERT, RoBERTa и др.), способные учитывать контекст и скрытые семантические связи в тексте.

Разработка алгоритмов искусственного интеллекта для классификации текстовых сообщений предполагает комплексный подход, включающий этапы сбора и разметки данных, предобработки текстов, выбора модели, обучения и оценки качества. Важным аспектом является также обеспечение интерпретируемости результатов и адаптивности алгоритмов к различным предметным областям и языковым особенностям.

Актуальность исследования обусловлена необходимостью повышения эффективности анализа больших объемов текстовой информации, автоматизации процессов мониторинга и поддержки принятия решений. Практическая значимость работы заключается в возможности применения разработанных алгоритмов в системах анализа отзывов, управлении качеством обслуживания, медиамониторинге, маркетинговых исследованиях и других сферах, где требуется оперативная и точная обработка текстовых данных.

Проблема классификации текстовых сообщений по тональности и семантическим категориям широко исследуется в рамках обработки естественного языка (NLP) и машинного обучения. В научной литературе можно выделить несколько этапов эволюции подходов: от лексико-статистических методов к глубоким нейросетевым архитектурам.

На раннем этапе исследования анализ тональности базировался преимущественно на словарных и статистических методах. В работе Pang и Lee (2008) показано, что задачи sentiment analysis могут эффективно решаться с использованием классических алгоритмов машинного обучения, таких как наивный байесовский классификатор и метод опорных векторов (SVM). Авторы подчеркивают значимость представления текста в виде векторных моделей (Bag-of-Words, n-граммы), что стало основой для дальнейших исследований.

Manning, Raghavan и Schütze (2009) в фундаментальном труде «Introduction to Information Retrieval»[1] детально описывают методы представления текстов через TF-IDF и косинусное сходство, а также механизмы классификации документов. Их исследования показали, что корректный выбор признакового пространства оказывает критическое влияние на точность классификации.

С развитием глубокого обучения произошёл качественный скачок в области анализа текста. Mikolov и соавторы (2013)[2] предложили модели Word2Vec, которые позволили представлять слова в виде плотных векторных представлений (embeddings), отражающих их семантическую близость. Это стало важным этапом

перехода от поверхностных статистических признаков к семантически насыщенным моделям.

В 2018 году Devlin и соавторы представили модель BERT (Bidirectional Encoder Representations from Transformers), основанную на архитектуре трансформеров, предложенной Vaswani и соавторами (2017)[3]. Данные исследования продемонстрировали значительное повышение качества задач классификации текста за счёт учета двустороннего контекста. В отличие от традиционных моделей, трансформеры способны анализировать зависимость между словами на больших расстояниях, что особенно важно при определении тональности и сложных семантических категорий.

Работы Liu et al. (2019) по модели RoBERTa подтвердили, что дальнейшая оптимизация архитектуры трансформеров и обучение на больших корпусах данных позволяет достигать более высокой точности в задачах sentiment analysis и тематической классификации[4]

В современных исследованиях также рассматриваются вопросы интерпретируемости моделей (Ribeiro et al., 2016 — LIME), что особенно важно при использовании алгоритмов искусственного интеллекта в прикладных областях, таких как анализ отзывов клиентов, медиамониторинг и системы поддержки принятия решений. Интерпретируемость позволяет объяснять, какие именно фрагменты текста повлияли на итоговую классификацию.

Таким образом, анализ научной литературы показывает, что эволюция методов классификации текстов прошла путь от простых лексических моделей к сложным нейросетевым архитектурам глубокого обучения. При этом современные исследования акцентируют внимание не только на повышении точности, но и на адаптивности моделей, их устойчивости к шуму данных и возможности интерпретации результатов.

Следовательно, разработка алгоритмов искусственного интеллекта для классификации текстовых сообщений должна учитывать лучшие достижения как классических статистических методов, так и современных трансформерных моделей, обеспечивая баланс между точностью, вычислительной эффективностью и практической применимостью.

Разработка алгоритмов искусственного интеллекта для классификации текстовых сообщений по тональности и семантическим категориям представляет собой комплексный процесс, включающий несколько взаимосвязанных этапов. Прежде всего осуществляется предобработка текстовых данных. Данный этап предполагает очистку текста от лишних символов, удаление стоп-слов, приведение слов к начальной форме (лемматизация) и нормализацию регистра. Например, исходное сообщение «Обслуживание было ужасным, ждать пришлось слишком долго!» после предобработки трансформируется в набор ключевых слов: «обслуживание, ужасный, ждать, долго». Такая обработка позволяет снизить уровень шума и выделить наиболее значимые признаки для дальнейшего анализа.

Следующим этапом является классификация по тональности. В рамках традиционных методов используется представление текста в виде векторной модели (TF-IDF) с последующим применением алгоритмов машинного обучения, таких как логистическая регрессия или метод опорных векторов. Например, сообщение «Очень доволен качеством сервиса» будет отнесено к позитивной тональности, тогда как высказывание «Товар оказался некачественным и быстро сломался» — к негативной. Однако классические модели не всегда корректно учитывают контекст. Фраза «Не могу сказать, что сервис плохой» содержит слово «плохой», которое может привести к ошибочной негативной классификации. Современные нейросетевые модели, такие как BERT, анализируют двусторонний контекст и способны правильно интерпретировать подобные конструкции, определяя их как нейтральные или умеренно позитивные[5]

Семантическая классификация предполагает распределение текстов по тематическим категориям. Например, сообщение «Прошу заменить товар по гарантии» относится к категории «жалоба», а текст «Хотел бы узнать условия доставки» — к категории «информационный запрос». Для решения данной задачи применяются многоклассовые классификаторы, способные учитывать смысловую структуру текста. Современные трансформерные архитектуры позволяют одновременно выполнять классификацию по тональности и тематике, что повышает эффективность анализа.

Оценка качества алгоритмов осуществляется с использованием метрик accuracy, precision, recall и F1-score. Например, если модель корректно классифицирует 900 из 1000 сообщений, точность (accuracy) составляет 90%. Однако при несбалансированных выборках более информативной является F1-мера, позволяющая учитывать баланс между точностью и полнотой.

Практическая значимость разработанных алгоритмов заключается в их применении в системах анализа клиентских отзывов, мониторинга социальных сетей, маркетинговых исследованиях и службах поддержки. Автоматическая классификация текстов позволяет оперативно выявлять негативные сигналы, анализировать общественные настроения и структурировать большие массивы данных для последующего принятия решений.

Проведённый анализ подтверждает, что алгоритмы искусственного интеллекта на основе методов обработки естественного языка являются эффективным инструментом для классификации текстовых сообщений по тональности и семантическим категориям. Эволюция подходов от статистических моделей к нейросетевым архитектурам глубокого обучения существенно повысила точность и адаптивность систем анализа текста.

Современные трансформерные модели обеспечивают учет контекста и скрытых семантических связей, что особенно важно при обработке сложных языковых конструкций. Интеграция методов анализа тональности и тематической

классификации открывает широкие перспективы для создания интеллектуальных систем мониторинга и поддержки принятия решений.

Таким образом, разработка алгоритмов NLP-классификации является актуальным и перспективным направлением, сочетающим теоретическую новизну и практическую значимость.

СПИСОК ЛИТЕРАТУРЫ

1. Журафский Д., Мартин Дж. Х. Обработка речи и языка. — 3-е изд. — Пирсон, 2023.
2. Мэннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — Кембридж: Кембридж Юниверсити Пресс, 2009.
3. Девлин Дж., Чанг М.-В., Ли К., Тутанова К. BERT: Предобучение глубоких двунаправленных трансформеров для понимания языка // Материалы конференции НААКЛ-ХЛТ. — 2018.
4. Васвани А. и др. Механизм внимания как основа трансформеров (Attention Is All You Need) // Материалы конференции НейрИПС. — 2017.
5. Панг Б., Ли Л. Анализ тональности и извлечение мнений // Foundations and Trends in Information Retrieval. — 2008.