РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ И ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

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

Аннотация: *В работе рассматриваются методы разработки интеллектуальных алгоритмов для автоматической классификации текстовых сообщений по эмоциональной окраске и тематическим категориям с применением технологий обработки естественного языка. Исследуются современные подходы машинного и глубокого обучения, включая векторные представления текста и трансформерные архитектуры. Предлагается алгоритмическая модель, обеспечивающая повышение точности анализа тональности и семантической структуры сообщений. Практическая значимость исследования заключается в возможности применения разработанных решений в системах мониторинга социальных сетей, анализа отзывов и автоматизированной обработки клиентских обращений.*

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

В современном цифровом обществе объем текстовой информации, создаваемой пользователями интернета, организациями и автоматизированными системами, стремительно возрастает. Социальные сети, онлайн-платформы, службы технической поддержки, новостные ресурсы и корпоративные базы данных ежедневно генерируют огромные массивы текстовых сообщений. В условиях информационной перегрузки возникает необходимость в разработке интеллектуальных инструментов автоматической обработки и анализа текстов.

Одним из наиболее востребованных направлений является классификация текстовых сообщений по тональности и семантическим категориям. Анализ тональности позволяет определить эмоциональную окраску текста (позитивную, негативную или нейтральную), что особенно важно при мониторинге общественного мнения, анализе клиентских отзывов и управлении репутацией компаний. Семантическая классификация обеспечивает тематическое распределение сообщений, позволяя структурировать информацию и выявлять ключевые проблемные области или тенденции.

Современные методы обработки естественного языка (Natural Language Processing, NLP) и алгоритмы искусственного интеллекта открывают широкие

возможности для решения данных задач. Классические методы машинного обучения, основанные на векторизации текста (TF-IDF, Bag-of-Words), постепенно уступают место моделям глубокого обучения, таким как нейронные сети и трансформеры. Архитектуры типа BERT и других языковых моделей демонстрируют высокую точность за счёт учета контекстуальных зависимостей и скрытых семантических связей между словами.

Актуальность исследования обусловлена необходимостью повышения эффективности автоматизированного анализа текстовых данных и создания универсальных алгоритмов, способных адаптироваться к различным предметным областям. Практическая значимость работы заключается в возможности применения разработанных моделей в системах медиамониторинга, службах поддержки клиентов, маркетинговых исследованиях и интеллектуальных информационных системах.

Таким образом, разработка алгоритмов искусственного интеллекта для классификации текстовых сообщений по тональности и семантическим категориям является перспективным направлением, сочетающим научную новизну и высокую прикладную ценность.

Современные исследования в области обработки естественного языка показывают, что эффективность классификации текстов напрямую зависит от выбранной архитектуры модели, качества обучающей выборки и методов предварительной обработки данных. В рамках анализа алгоритмов искусственного интеллекта для определения тональности и семантических категорий целесообразно рассмотреть сравнительные характеристики различных подходов.

Прежде всего, традиционные методы машинного обучения основаны на формальном представлении текста в виде набора признаков. Такие модели, как наивный байесовский классификатор, логистическая регрессия и метод опорных векторов, используют статистические показатели частоты слов или их комбинаций. Достоинством данных методов является относительная простота реализации и низкие вычислительные затраты. Однако они имеют ограничение: игнорируют сложные синтаксические и контекстуальные зависимости. Например, предложения «Это было неожиданно хорошо» и «Это было неожиданно плохо» отличаются всего одним словом, но классические модели могут испытывать трудности при анализе усилительных или отрицательных конструкций[1]

С развитием глубокого обучения векторные представления слов (word embeddings) позволили учитывать семантическую близость понятий. Модели типа Word2Vec и GloVe формируют плотные числовые представления слов, отражающие их смысловые связи[2]. Это значительно повысило качество классификации, особенно в задачах тематического анализа. Тем не менее данные

подходы фиксируют значения слов вне контекста, что ограничивает их точность при обработке многозначных выражений.

Революционным этапом стало внедрение трансформерных архитектур. Механизм self-attention обеспечивает анализ взаимосвязей между всеми словами предложения одновременно, что позволяет учитывать сложные грамматические конструкции, инверсии и скрытые эмоциональные оттенки. Например, в предложении «Хотя сервис не идеален, в целом я остался доволен» трансформерная модель способна выделить доминирующую позитивную оценку, несмотря на наличие отрицательной характеристики[3].

В рамках семантической классификации также важна проблема многозадачного обучения (multi-task learning). Современные алгоритмы позволяют одновременно определять тональность и тематическую категорию текста. Это особенно актуально в прикладных системах, где необходимо не только выявить эмоциональную окраску сообщения, но и определить его тип: жалоба, рекомендация, информационный запрос или техническое обращение. Интеграция нескольких задач в единую модель повышает согласованность результатов и снижает вычислительные затраты.

Отдельного внимания заслуживает вопрос интерпретируемости алгоритмов. В прикладных сферах, таких как банковский сектор или государственное управление, важно понимать, какие именно признаки повлияли на решение модели. Использование методов объяснимого искусственного интеллекта (Explainable AI) позволяет анализировать вклад отдельных слов и фраз в итоговую классификацию, что повышает доверие к системе[4]

Кроме того, эффективность алгоритмов зависит от качества обучающих данных. Наличие шумов, сарказма, неоднозначных формулировок и языковых вариаций может существенно снижать точность классификации. В связи с этим актуальной задачей является создание сбалансированных корпусов текстов и применение методов аугментации данных.

Таким образом, аналитический обзор показывает, что современные трансформерные модели демонстрируют наилучшие результаты по сравнению с традиционными методами, однако их применение требует значительных вычислительных ресурсов и тщательно подготовленных данных. Оптимальная стратегия разработки алгоритма заключается в комбинировании преимуществ различных подходов с учётом специфики предметной области и практических требований к системе.

Проведённое исследование показало, что разработка алгоритмов искусственного интеллекта для классификации текстовых сообщений по тональности и семантическим категориям является актуальным и перспективным направлением в области обработки естественного языка. Стремительный рост

объёма текстовой информации требует внедрения автоматизированных систем анализа, способных быстро и точно обрабатывать большие массивы данных.

Анализ существующих подходов показал, что традиционные методы машинного обучения отличаются простотой реализации и сравнительно низкими вычислительными затратами, однако ограничены в учёте контекстуальных и семантических связей. Современные модели глубокого обучения, основанные на архитектуре трансформеров, обеспечивают более высокую точность благодаря механизму внимания и способности анализировать текст с учётом двустороннего контекста.

Эффективность алгоритмов во многом зависит от качества обучающих данных, корректной разметки и выбора адекватных метрик оценки. Практическая значимость разработанных решений заключается в возможности их применения в системах медиамониторинга, анализе клиентских отзывов, службах поддержки и интеллектуальных информационных системах.

Таким образом, интеграция методов машинного обучения и современных нейросетевых архитектур открывает широкие перспективы для создания высокоэффективных систем автоматической классификации текстовой информации.

Список литературы:

1. Журафский Д., Мартин Дж. Х. Обработка речи и языка. — Пирсон, 2023.
2. Мэннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. — Кембридж: Cambridge University Press, 2009.
3. Васвани А. и др. Механизм внимания как основа трансформеров (Attention Is All You Need) // Материалы конференции NeurIPS. — 2017.
4. Девлин Дж., Чанг М.-В., Ли К., Тутанова К. BERT: Предобучение глубоких двунаправленных трансформеров для понимания языка // Материалы конференции NAACL-HLT. — 2018.