УДК 004.8

ИССЛЕДОВАНИЕ СИСТЕМ РАСПОЗНАВАНИЯ ОТЗЫВОВ СГЕНЕРИРОВАННЫХ БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ МЕТОДАМИ КЛАССИЧЕСКОГО МАШИННОГО ОБУЧЕНИЯ

Казанцев О. П.[1], Афанасьев М. А.[1], Колосов Н. А.[1]
Научный руководитель – канд. техн. наук, преподаватель Евстафьев О. А.[1]
[1]Университет ИТМО
kazantsev@niuitmo.ru

Введение

Генеративные модели продолжают развиваться, демонстрируя рост качества и одновременное снижение стоимости их применения [1]. Тексты, созданные с применением больших языковых моделей (англ. Large Language Models, LLM; далее – БЯМ), все сложнее отличить от текстов, написанных человеком, что повышает риск их недобросовестного использования, например, в создании недостоверных пользовательских отзывов. В связи с этим актуально исследовать возможность автоматического определения автора отзыва на русском языке при помощи маломощных классических методов машинного обучения, а также провести интерпретацию полученных результатов. В работе рассматривается задача бинарной классификации отзывов по категориям: созданные человеком или сгенерированные БЯМ.

Основная часть

В качестве источника отзывов, написанных человеком, использовался набор данных «RuReviews», концентрирующийся на отзывах о товарах в сегменте женской одежды и аксессуаров [2]. Для получения сгенерированных отзывов, была проведена аугментация данных посредством генерации текста с использованием БЯМ. Использовалась модель «mistral-large-2512» [3] и составная промпт-инструкция с различными комбинациями параметров-метаданных. Совокупный размер датасета составляет 19 894 строк с 11 параметрами (текст, автор отзыва; метаданные: тональность, товар, портрет покупателя, кому покупался товар, длина отзыва, грамотность письма). Доли отзывов, написанных человеком и сгенерированных БЯМ, равны. В обучении модели использовались текст отзыва и целевой класс – автор отзыва, остальные параметры применялись для анализа результатов после обучения модели.

В основе решения лежит обучение ансамбля моделей, каждая из которых оперирует различными типами признаков, извлекаемых из текста:

– стилометрическое представление – вектор статистических характеристик текста (например, средние длины структур, встречаемость знаков препинания и другие), отражающий устойчивые стилевые и структурные паттерны;

– TF-IDF на основе слов – признаковое пространство, фиксирующее устойчивую лексику [4];

– TF-IDF на основе символов (n-грамм) – признаковое пространство, фиксирующее орфографию, пунктуацию и наличие опечаток [4];

– эмбеддинг документа – семантическое представление отзыва, полученное с помощью модели-трансформера «paraphrase-MiniLM-L3-v2», предназначенной для работы с длинными фрагментами текста (англ. sentence transformer) [5].

Каждое представление текста подается на вход собственной модели логистической регрессии [6]. Этот тип модели был выбран из-за простоты и удобства интерпретирования. Итоговый прогноз формируется методом накопления (англ. stacking): вероятности базовых моделей используются в качестве входных данных для метамодели логистической регрессии.

Анализ и интерпретация результатов проведены в два этапа:

– изучение зависимости между метаданными и качеством прогнозирования;

– интерпретация весовых коэффициентов обученной метамодели и оценка значимости прогнозов базовых моделей.

Выводы

В ходе исследования была разработана система распознавания и классификации текстовых отзывов, характеризующаяся высокой точностью прогнозирования и обеспечивающая возможность интерпретации факторов, влияющих на качество принимаемых решений.

По результатам исследования установлено влияние метаданных на точность прогнозирования, а также выполнена сравнительная оценка эффективности различных способов векторного представления текста применительно к решаемой задаче:

– краткость и высокая языковая нормативность отзыва являются факторами, снижающими точность прогнозирования модели;

– интерпретация моделей, построенных на TF-IDF признаках, показала, что тексты, принадлежащие человеку, характеризуются устойчивой тематической лексикой (лексемы «товар», «продавец», «заказ») и избыточным использованием восклицательных знаков («!!», «!!!»). Тогда как для отзывов, сгенерированных БЯМ, типичны дискурсивные маркеры разговорного регистра (частица «ну») и пунктуационные конструкции с тире («—»);

– анализ весовых коэффициентов метамодели показал, что наибольший вклад в итоговое решение вносят прогнозирования модели, основанные на символьных и лексических TF-IDF характеристиках. Вклад эмбеддинговой модели характеризуется как умеренный, тогда как модель на стилометрических признаках демонстрирует наименьшую прогностическую значимость.

В дальнейшем, для повышения точности прогнозирований и универсальности модели целесообразно расширить набор данных другими сегментами товаров и провести аугментацию данными, сгенерированными при помощи других БЯМ и типов промптов. Полученные результаты работы формируют перспективную основу для разработки вычислительно нетребовательной системы классификации отзывов, пригодной для промышленной эксплуатации.

Литература

1. Miao X. et al. Towards efficient generative large language model serving: A survey from algorithms to systems // ACM Computing Surveys. 2025. Vol. 58, no. 1. P. 1–37.
2. Smetanin S. RuReviews: Russian Reviews Dataset [Электронный ресурс]. – Режим доступа: https://github.com/sismetanin/rureviews (дата обращения: 10.02.2026).
3. Mistral AI. Mistral Large 3 (v25.12) [Электронный ресурс]. – Режим доступа: https://docs.mistral.ai/models/mistral-large-3-25-12 (дата обращения: 10.02.2026).
4. Patil R. et al. A survey of text representation and embedding techniques in NLP // IEEE Access. 2023. Vol. 11. P. 36120–36146.
5. Sentence-Transformers. paraphrase-MiniLM-L3-v2 [Электронный ресурс] // Hugging Face. – Режим доступа: https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L3-v2 (дата обращения: 10.02.2026).
6. Gasparetto A. et al. A survey on text classification algorithms: From text to predictions // Information. 2022. Vol. 13, no. 2. P. 83.