УДК 004

Драгунский Н.С., Хусамов Р.Р., Ганиев Ш.И., Шилин С.В.
Уфимский университет науки и технологий, Уфа

Научный руководитель:
Валеев С.С.
Уфимский университет науки и технологий, Уфа

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ DISTILBERT

Аннотация. В статье рассматривается задача анализа эмоциональной окраски текста с использованием модели DistilBERT, являющейся облегченным вариантом BERT. Приведены основные этапы подготовки данных, методология обучения модели, а также результаты экспериментов на различных корпусах. Обоснована применимость модели в прикладных задачах анализа пользовательского контента.

Ключевые слова: анализ тональности, DistilBERT, BERT, трансформеры.

Большие языковые модели – это системы машинного обучения. Они построены на глубоких нейронных сетях и содержат огромное количество параметров, позволяющих эффективно решать разнообразные задачи обработки естественного языка (NLP). Кратко рассмотрим основные характеристики и особенности больших языковых моделей.

Архитектура. Большинство современных крупных языковых моделей основаны на архитектуре трансформеров. Трансформеры используют механизм внимания, позволяющий моделировать зависимости между словами независимо от расстояния между ними в тексте. Это значительно улучшает качество понимания контекста и семантики текста. Основные компоненты архитектуры трансформера:

Encoder – преобразует входной текст в векторное представление, учитывая контекст каждого слова; Decoder – генерирует выходной текст на основе полученных представлений; Attention Mechanism – вычисляет связи между элементами последовательности, позволяя лучше учитывать контекст.

Современные большие языковые модели применяются во множестве различных областей:

– обработка естественного языка: создание диалоговых агентов, чат-ботов, виртуальных помощников;

– автоматическое написание текстов: генерация статей, писем, отчетов, сценариев и другого контента;

– перевод текстов: высококачественный машинный перевод на разные языки;

– анализ настроений и эмоций: выявление положительных/отрицательных отзывов пользователей;

– создание вопросов и ответов: автоматическая генерация FAQ разделов, поддержка клиентов;

– генерация изображений по текстовому описанию: использование совместно с моделями компьютерного зрения позволяет создать мультимедийные продукты.

Вот некоторые известные крупные языковые модели:

– GPT (Generative Pre-trained Transformer) – серия моделей от OpenAI, начиная с GPT-1 и заканчивая актуальной версией GPT-4;

– BERT (Bidirectional Encoder Representations from Transformers) – модель широко используемая для классификации текстов и распознавания сущностей;

– RoBERTa, DistilBERT, XLNet – различные вариации BERT с улучшенными характеристиками;

– YandexGigaChat – российская большая языковая модель от Яндекса, доступная пользователям;

– GigaChat – российская большая языковая модель от Sbera, доступная пользователям.

Эти модели отличаются объемом параметров, количеством слоев и уровнем подготовки, что делает их пригодными для решения конкретных задач.

Преимущества: высокая точность и эффективность в обработке сложных запросов, способность понимать и генерировать осмысленный контент, возможность интеграции в разнообразные приложения и сервисы, универсальность применения в разных областях NLP.

Недостатки: высокие требования к вычислительным ресурсам для тренировки и эксплуатации, возможности возникновения ошибок при недостаточной точности тренировочных данных, риск генерации недостоверной или предвзятой информации, необходимость постоянного обновления и адаптации к новым данным.

Актуальность задачи анализа эмоциональной окраски текстов возрастает с развитием цифровых коммуникаций. Анализ эмоционального состояния позволяет решать широкий спектр задач – от мониторинга социальных настроений до улучшения качества обслуживания клиентов.

Классические методы, такие как словарный анализ, уступают по точности и гибкости современным нейросетевым моделям, основанным на архитектуре трансформеров [1].

Одной из наиболее перспективных моделей является DistilBERT – упрощенная версия BERT, обладающая меньшими требованиями к ресурсам при сохранении высокой точности. DistilBERT обучается на корпусе BERT посредством метода знаний-дистилляции, позволяющего сохранить ключевые языковые закономерности [2–6].

Для проведения исследований была выбрана модель distilbert-base-uncased, предварительно дообученная на задаче классификации эмоциональной окраски текста. В качестве корпуса использовалась размеченная выборка твитов с платформы Kaggle, содержащая тексты, относящиеся к категориям «позитивный», «негативный» и «нейтральный».

Предварительная обработка данных включала: очистку текста от эмодзи, ссылок и HTML-тегов; лемматизацию с использованием библиотеки spaCy; токенизацию средствами HuggingFace Tokenizer.

Далее проводилось дообучение модели на размеченном датасете. Использовался кросс-энтропийный лосс и оптимизатор AdamW. Процесс обучения длился 3 эпохи с батч-сайзом 32. Контроль переобучения осуществлялся через сохранение наилучшей модели по метрике accuracy на валидационной выборке.

Отдельного внимания заслуживает визуализация внимания (attention scores), которая показала, что модель эффективно выделяет маркеры эмоционального окраса – такие как «прекрасный», «ужасно», «ненавижу», и др.

Применение данной модели может быть эффективно в: службах поддержки (автоматическая оценка эмоционального состояния клиента); социальных сетях (мониторинг настроений); образовании (оценка эмоционального восприятия учебных материалов); HR-аналитике (анализ отзывов сотрудников).

Таким образом, модель DistilBERT демонстрирует высокую эффективность при решении задачи анализа эмоциональной окраски текстов и может быть внедрена в прикладные информационные системы без значительных вычислительных затрат.

Таким образом, модель DistilBERT демонстрирует высокую эффективность при решении задачи анализа эмоциональной окраски текстов и может быть внедрена в прикладные информационные системы без значительных вычислительных затрат.

Список использованных источников:

1. Свидетельство о государственной регистрации программы для ЭВМ № 2023617147 Российская Федерация. Программное обеспечение для выявления противоправного контента: № 2023616049: заявл. 28.03.2023: опубл. 05.04.2023 / Р.Ф. Исмагилов, Н.Д. Лушников, А.С. Исмагилова;

заявитель федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский университет науки и технологий».

2. Vaswani A., et al. Attention is All You Need. Advances in Neural Information Processing Systems. 2017.

3. Sanh V., et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019.

4. Devlin J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. 2018.

5. Wolf T., et al. Transformers: State-of-the-Art Natural Language Processing. EMNLP. 2020.

6. Liu Y., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. 2019.

Dragunskiy N.S., Khusamov R.R., Ganiev S.I., Shilin S.V.
Ufa University of Science and Technology, Ufa

Scientific supervisor:
Valeev S.S.
Ufa University of Science and Technology, Ufa

TEXT SENTIMENT ANALYSIS USING THE DISTILBERT MODEL

Abstract. The article explores the task of text sentiment analysis using the DistilBERT model, a lightweight version of BERT. The paper outlines data preprocessing, model training methodology, and experiment results on various corpora. The applicability of the model to real-world user-generated content analysis is demonstrated.

Keywords: sentiment analysis, DistilBERT, emotionality, transformers.