

ОПРЕДЕЛЕНИЕ ЭМОЦИОНАЛЬНОГО ФОНА ТЕКСТОВ НА ОСНОВЕ МОДЕЛЕЙ И МЕТОДОВ NLP

Павлюченко К.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Парамонов А.И. – канд. техн. наук, доцент

Выполнен анализ методов автоматической обработки текстов на естественном языке с целью определения эмоционального фона. Рассмотрены существующие модели и методы машинного обучения, которые можно применять для классификации эмоций. Представлена их общая архитектура и особенности применения. Предложен метод на основе трансформеров для определения эмоционального фона текстов.

Сентимент-анализ, или анализ тональности текста, – это развивающееся направление компьютерной лингвистики, основной задачей которого является выявление в документе эмоционально окрашенной лексики и эмоциональной оценки объектов автором [1].

Выявление эмоционального контекста в текстах – сложная задача обработки естественного языка (NLP), направленная на определение эмоциональной окраски сообщений, таких как комментарии, рецензии или твиты. Это требует не только выявления базовых эмоций, но и учета контекста и языковых особенностей.

Существующие методы анализа естественно-языковых текстов имеют ограничения при обработке сложных текстов, где эмоции часто скрыты за сложными синтаксическими структурами или распределены между разными уровнями текста. Известные методы, такие как наивный байесовский классификатор и метод опорных векторов, имеют ограничения при работе со сложными контекстными зависимостями и не всегда способны учесть многообразие способов выражения эмоций в тексте.

В последние годы активно разрабатываются подходы для классификации эмоций, включая методы машинного обучения и нейронные сети [2, 3]. В работе рассматривается решение задачи анализа эмоциональной окраски текста с помощью современных моделей машинного обучения. В качестве признаков для анализа эмоциональной окраски могут использоваться и отдельные слова, и биграммы. Каждая модель имеет свои преимущества и ограничения, и их выбор зависит от конкретной задачи, целей проекта и доступных ресурсов. Важным моментом является то, что не все существующие модели могут удовлетворять требования по точности, скорости или специфике анализа, что делает выбор подходящего решения критически важным для успешной реализации проекта. Особый интерес при этом проявляется к моделям на основе трансформеров, которые способны эффективно обрабатывать контекстную информацию и анализировать сложные лингвистические конструкции.

Сегодня среди популярных и активно используемых моделей можно выделить BERT и GPT, также их модификации и методы: RoBERTa, Encoder-Decoder, Multi-Head Attention, Self-Attention, Attention Mechanism, DistilBERT.

Bidirectional Encoder Representations from Transformers (BERT) – одна из самых эффективных на сегодня моделей для обработки текста, которая активно используется в задачах анализа эмоций и классификации текста [4]. Одним из главных преимуществ данной модели является, то, что ее можно натренировать один раз на огромном количестве данных, а потом эффективно тренировать для других задач, в которых нет большого количества данных, что делает ее мощным инструментом для различных задач NLP, включая классификацию эмоций. Однако его применение требует значительных вычислительных ресурсов, а также времени на обучение, что может быть ограничением для небольших проектов или если требуется быстрое развертывание модели. RoBERTa [5] – популярная и уже зарекомендовавшая себя модификация BERT, которая характеризуется более тщательной и разумной оптимизацией обучающих параметров для модели BERT. Считается более производительной своей базовой модели BERT практически во всех задачах. Модель DistilBERT реализует оптимизацию обучения для модели BERT за счет уменьшения размерности и увеличения скорости обработки текста. В частности, DistilBERT имеет объем на 40% меньше, чем оригинальная BERT-модель, и она на 60% быстрее ее, при этом сохраняет 97% функциональности.

На основе проведенного исследования проблемы текстового выражения эмоциональных состояний в лингвистике [6], в работе предлагается метод автоматической обработки и анализа текстов, направленный на точное выявление эмоционального контекста. В основе метода лежит трансформерная модель, базирующаяся на современной архитектуре энкодера (например RoBERTa), специально адаптированная и дообученная для задачи распознавания эмоций в текстах. Выбор трансформерной архитектуры обусловлен ее доказанной эффективностью в задачах обработки естественного языка, где требуется глубокое понимание контекста и семантики. Трансформерные модели демонстрируют особую результативность при анализе текстов разной длины и структуры, что

критически важно для обработки сообщений в цифровой коммуникации, которые могут варьироваться от коротких реплик до развернутых высказываний.

Трансформер, или трансформерная модель, реализует ключевую концепцию «механизм внимания», который изучает контекстуальные отношения между словами (или лексемами) в тексте. Архитектура трансформера оказала огромное влияние на область обработки естественного языка (NLP), заменив более старые архитектуры, такие как RNN и LSTM, благодаря своим преимуществам в реализации параллелизма и способности эффективно обрабатывать длинные зависимости в данных. В своей общей архитектуре трансформер включает два основных механизма – кодировщик, который считывает и предобрабатывает введенный текст, и декодер, который выдает прогноз для задачи. В отличие от направленных (векторных) моделей, которые считывают вводимый текст последовательно (в одном направлении), кодировщик трансформера считывает сразу всю последовательность слов. Эта особенность позволяет модели изучать контекст слова на основе всего его окружения (и слева и справа от слова). Общая структура кодировщика трансформера представлена на рисунке 1. Вход представляет собой последовательность токенов, которые сначала встраиваются в векторы, а затем передаются и обрабатываются в нейронной сети. Выход представляет собой последовательность векторов, каждый из которых соответствует входному токену с тем же индексом. В работе предложена идея использовать механизм «внутреннего внимания» (*self-attention*), который предполагает, что каждый элемент входной последовательности может учитывать все остальные элементы последовательности при обработке.

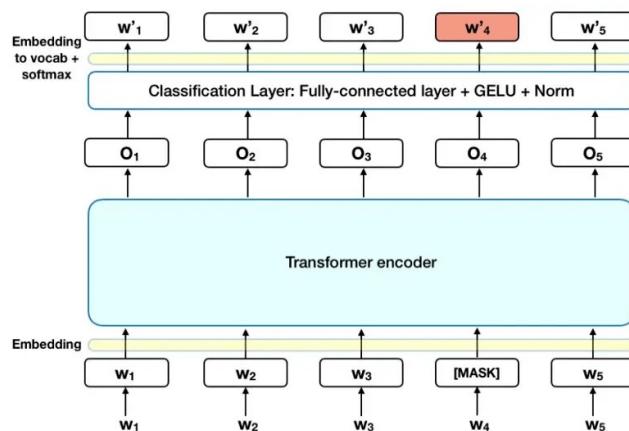


Рисунок 1 – Классическая структура кодировщика трансформера

Предлагаемый метод использует многоуровневый подход, который учитывает как глобальный, так и локальный контексты за счет последовательной обработки текста на нескольких уровнях: анализ общей эмоциональной окраски текста и выделение эмоций для отдельных предложений.

Предложенное решение показало улучшенную точность при работе с короткими сообщениями, характерными для современных платформ социальных медиа и мессенджеров. Это достигается благодаря способности системы извлекать максимум информации из ограниченного контекста путем комплексного анализа всех доступных эмоциональных маркеров. Проводится серия компьютерных экспериментов для определения показателей достигаемой точности выявления эмоционального фона для текстов разной размерности. На данный момент предварительная оценка общей точности классификации оценивается на уровне 80%, что считается хорошим показателем.

Список использованных источников:

1. Gelbukh A. (Ed.) *Computational Linguistics and Intelligent Text Processing // 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I.* – Springer-Verlag Berlin Heidelberg, 2013. – 598 p.
2. Картер, Д. Нейросети. Обработка естественного языка / Д. Картер – «Автру», 2003. – 156 с.
3. Романов, А. С. Анализ тональности текста с использованием методов машинного обучения / А. С. Романов, М. И. Васильева, А. В. Куртукова, Р. В. Мещеряков // R. Piotrowski's Readings in Language Engineering and Applied Linguistics : Proceedings, Saint-Petersburg, 27 ноября 2017 года / St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Herzen State Pedagogical University of Russia. – Saint-Petersburg: Creative Commons CCO, 2018. – С. 86-95.
4. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. // arXiv preprint arXiv:1810.04805. (2018). URL: <https://arxiv.org/pdf/1810.04805.pdf> (дата обращения: 10.04.2025).
5. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Liu, Yinhan & Ott, Myle & Goyal, Naman & Du, Jingfei & Joshi, Mandar & Chen, Danqi & Levy, Omer & Lewis, Mike & Zettlemoyer, Luke & Stoyanov, Veselin. // arXiv preprint arXiv:10.48550/arXiv.1907.11692. (2019). URL: <https://arxiv.org/pdf/1907.11692.pdf> (дата обращения: 10.04.2025).
6. Павлюченко, К. А. Анализ методов выявления эмоционального контекста в текстах / К. А. Павлюченко // Информационные системы и технологии : сборник статей 60-ой юбилейной научной конференции аспирантов, магистрантов и студентов, Минск, 22–26 апреля 2024 г. / Институт информационных технологий Белорусского государственного университета информатики и радиоэлектроники ; редкол.: А. И. Парамонов [и др.]. – Минск, 2024. – С. 48–45.