

АНАЛИЗ ТОНАЛЬНОСТИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Сметанин Сергей Игоревич

Студент

Факультет бизнеса и менеджмента НИУ ВШЭ, Москва, Россия

E-mail: sismetanin@edu.hse.ru

Автоматическая классификация текстов по тональности является важной прикладной задачей, позволяющей анализировать мнения потребителей о товаре либо услуге. В данной работе рассматривается использование сверточной нейронной сети для классификации русскоязычных отзывов из интернет-магазинов на пять классов.

Свёрточные нейронный сети (CNN), изначально предназначенные для обработки изображений, демонстрируют высокие показатели качества при решении задач классификации тональности. В рамках соревнования по анализу тональности англоязычных текстов SemEval-2017 [1] и русскоязычных текстов SentiRuEval-2016 [2] подходы с использованием CNN продемонстрировали одни из самых высоких результатов.

Обучающая выборка общим объемом 25000 текстов, которая была сформирована на основе отзывов о товарах из интернет-магазина Aliexpress, содержит по 50000 текстов для каждого из классов тональности. Меткой класса для каждого отзыва является рейтинг отзыва, проставленный автором. На основе сформированного набора текстов была обучена модель Word2Vec [3], которая далее была использована для построения embedding слоя нейронной сети. Предварительно тексты были приведены к нижнему регистру, все знаки препинания были удалены. Архитектура сверточной нейронной сети была основана на работе [4], где сверточные слои параллельно обрабатывают биграммы, триграммы и 4-граммы.

Для построения CNN была использована нейросетевая библиотека Keras с фреймворком Theano. При обучении в качестве функции потерь была использована категориальная кросс-энтропия, а в качестве алгоритма оптимизации – adam. Наивысший показатель $F_1(30\%)757.6\%$.

Литература

1. Rosenthal S., Farra N., Nakov P. SemEval-2017 task 4: Sentiment analysis in Twitter // Proceedings of the 11th International

- Workshop on Semantic Evaluation (SemEval-2017), 2017, P. 502–518.
2. Loukachevitch N., Rubtsova Y. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue, 2016, P. 375–384.
 3. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // In Advances in Neural Information Processing Systems, 2013, P. 3111–3119.
 4. Zhang Y., Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification // arXiv preprint arXiv:1704.06125, 2015.