УДК 004.838.2

АНАЛИЗ ЭМОЦИОНАЛЬНОЙ ТОНАЛЬНОСТИ: ОБЗОР И СРАВНИТЕЛЬНЫЙ АНАЛИЗ СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Мухин И.С. (ИТМО)

Научный руководитель – кандидат педагогических наук, доцент Авксентьева Е.Ю. (ИТМО)

Введение. Рынок обработки естественного языка (natural language processing – далее NLP) активно развивается на протяжении последних нескольких лет. Он демонстрирует непрерывный рост, что свидетельствует о высоком спросе на соответствующие технологии. И несмотря на это, предложение на рынке русскоязычного сегмента все еще довольно ограничено, что может быть связано со сложностью обработки языка. Это касается не только NLP решений в целом, но и тех, что связаны с анализом эмоциональной тональности (sentiment analysis – далее SA). При этом изучение подходов SA может способствовать дополнительным аспектом для анализа фондовых рынков, удовлетворенности различными услугами со стороны потребителей, и даже влияния человеческих эмоций на мировые события и процессы.

Основная часть. В основе множества решений SA лежат следующие методы [1]:

- С применением машинных моделей;
- Лингвистические алгоритмы:
    - Основанные на правилах;
    - Основанные на словарях тональности.

Учитывая, что эмоциональная оценка может отличаться от подхода к подходу, как правило, выделяют несколько классов для проведения прогнозов. Это может быть бинарная оценка (позитивная и негативная), тринарная, как это было представлено в работах SentiRuEval 2015 [2], и даже с применением шкалы, к примеру, от крайне негативного тона до крайне позитивного, примером чего служит словарь тональности linis-crowd 2015. Более подробно обзор датасетов и словарей тональности со ссылками на источники и количеством классов тональности представлен в таблице 2 и 3.

Таблица 1 – Список открытых датасетов размеченных русскоязычных текстов

Наименование датасета	Ссылка на источник	Количество классов тональности
SentiRuEval-2015	https://www.dialog-21.ru/evaluation/2015/sentiment/	4
SentiRuEval-2016	https://www.dialog-21.ru/evaluation/2016/sentiment/	3
ROMIP 2011	https://www.dialog-21.ru/evaluation/2012/sentiment/	2, 3 и 5
Russian Hotel Reviews Dataset	https://drive.google.com/drive/folders/17sa3h4XHcG0MJGrbfOsbL-kDW29CuJul	5
Kaggle Russian News dataset	https://www.kaggle.com/c/sentiment-analysis-in-russian	3
Kaggle sentiment analysis dataset	https://www.kaggle.com/c/methodcompetition1/data	3

Таблица 2 – Список открытых датасетов русскоязычных словарей тональности

Ссылка на источник	Количество классов тональности
https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent	3
https://dmafanasyev.github.io/rulexicon/reference/hash_rusentilex_2017.html	3
https://github.com/text-machine-lab/sentimental/blob/master/sentimental/word_list/russian.csv	шкала от -5 до 5
https://linis-crowd.org/static/collection_(docs_words)_2016_all_labels.zip	шкала от -2 до 2

Как уже упоминалось, как правило, существует два подхода к построению классификатора – на основе алгоритма машинного обучения или лингвистических

алгоритмов. Можно выделить следующие методы, используемые в современных решениях, представленных рынке разметки как русскоязычных, так и англоязычных текстов (таблица 3).

Таблица 3 – Список SA решений

Наименование решения	Ссылка на источник	Алгоритм
Stanford CoreNLP	https://nlp.stanford.edu/sentiment/	Рекурсивные нейронные сети
TextBlob	https://textblob.readthedocs.io/en/dev/advanced_usage.html#sentiment-analyzers	Лингвистические правила и байесовский классификатор
SentiFinder	https://eurekaengine.ru/ru/description/#sentiFinder	Случайный лес и градиентный бустинг
SentiScan	https://youscan.io/kz/blog/automatic-sentiment/	Гибрид машинных моделей и лингвистических правил
SentiStrengh	http://sentistrength.wlv.ac.uk/	Словарь тональности и лингвистические правила

В дополнение к алгоритмам, которые в то или иное время легли в основу коммерческих продуктов, в данной работе представлен перечень моделей в хронологическом порядке на основе информации с портала международной конференции Dialogue [3]:

- ROMIP 2011 — в первое время становления конференции по вопросам SA наибольшей популярностью пользовался метод опорных векторов (support vector machine – далее SVM), при этом в некоторых случаях можно было наблюдать применение байесовского классификатора и лингвистических алгоритмов, основанных на правилах;
- ROMIP 2012 — помимо SVM и подхода, основанном на лингвистических правилах распространение получило применение словарей тональности;
- Sentiment analysis 2016 — наблюдается применение нейронных сетей, так, наиболее популярными являются решения, основанные на сверточных и рекуррентных моделях;
- RuSentNE 2023 — к этому времени авторы работы занимаются fine-тюнингом на базе таких моделей, как Bert/RuBert, ruT5 и HAlf a MAsked Model (HAMAM).

Выводы. Проведен обзор существующих решений по анализу эмоциональной тональности, преимущественно русскоязычных текстов. В дальнейшем планируется разработка собственного подхода, расширяющего возможности рассмотренных решений. Также предстоит решить одну из наиболее часто встречающихся проблем SA – низкая точность распознавания корректного уровня тональности в случаях присутствия в тексте сарказма, поговорок, фразеологизмов, метафор и т.д.

Список использованных источников:

1. Devika M.D., Sunitha C., Ganesh A. Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science. – 2016. – Ч. 87. – С 44–49.
2. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. 2015. SentiRuEval: Testing Object-oriented Sentiment Analysis Systems in Russian // Компьютерная лингвистика и интеллектуальные технологии. – 2015. – № 14(2). – С. 3–15.
3. Dialogue. International the conference in computational linguistics and intelligent technologies [Электронный ресурс] URL: https://www.dialog-21.ru/ (Дата обращения: 16.01.2024)

Мухин И.С. (автор) _____

Авксентьева Е.Ю. (научный руководитель) _____