Д. Э. Пащенко, Е. В. Котельников
Вятский государственный университет, г. Киров
УДК 004.89

СОВМЕСТНЫЙ АНАЛИЗ НЕЙРОСЕТЕВОЙ ЯЗЫКОВОЙ МОДЕЛИ И СЛОВАРНОГО МЕТОДА В ЗАДАЧЕ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВ

Аннотация. В статье представлены результаты исследований по поиску зависимостей между тональностью, сопоставляемой тексту нейросетевой моделью ruRoberta, и вниманием, которое она уделяет оценочным словам. Предсказания языковой модели сравнивались со словарным методом SO-CAL, с помощью которого подбирался словарь оценочной лексики. Результаты показали, что модель ruRoberta уделяет больше внимания негативной оценочной лексике.

Ключевые слова: обработка естественного языка, анализ тональности, нейронные сети, внимание, словарный метод, BERT, SO-CAL.

Введение. Анализ тональности — область компьютерной лингвистики, направленная на поиск мнений и оценок людей по отношению к упомянутым в тексте объектам — продуктам, услугам, событиям и т. д. [1]. Наиболее успешными в этой области являются глубокие нейросетевые языковые модели, в частности, семейство моделей BERT [2], основанных на архитектуре Transformer [3].

Альтернативой нейронным сетям являются словарные методы. Они отличаются высокой скоростью вычислений, ввиду того что не требуют обучения, а также хорошей интерпретируемостью [4]. Тем не менее, словари оценочный лексики, на которых основывается работа этих методов, достаточно трудоемки в разработке и зачастую узконаправлены. Кроме того, словарные методы все еще не обеспечивают высокое качество анализа тональности, чтобы составлять конкуренцию нейросетевым моделям [5].

Проблема исследования. Задача повышения интерпретируемости нейронных сетей является весьма актуальной в современной науке, в частности, для понимания причин наблюдаемых результатов нейросетевой языковой модели при предсказании тональности текстов. Предполагается, что это, с одной стороны, позволит улучшить работу существующих хорошо интерпретируемых методов,

основанных на словарях, и, с другой стороны, повысить качество работы нейросетевых моделей за счет учета словарей оценочной лексики. В настоящем исследовании выдвигается гипотеза, что существует корреляция между весами внимания, уделяемых языковой моделью, и оценкой, которую она присваивает тексту.

Материалы и методы. В работе использовалась модель ruRoberta-large [6], демонстрирующая высокие результаты среди моделей типа BERT в рейтинге Russian SuperGLUE [7]. Эта модель была дообучена на большом количестве русскоязычных текстов, которые включали 6 текстовых корпусов, размеченных на классы «позитивный», «негативный» и «нейтральный» [8]. В корпусах присутствовали как короткие тексты — сообщения в социальных сетях, так и длинные — отзывы на фильмы и отели. В процентном соотношении данные разделились следующим образом: позитивные — 38,9%, негативные — 25,9% и нейтральные — 35,2% (табл. 1).

Таблица 1

Обучающие (80%) и валидационные (20%) данные

Источник	*Позитивные*	*Негативные*	*Нейтральные*	*Сумма*
Linis Crowd	3 582	19 005	20 526	43 113
Ru Reviews	29 999	30 000	30 000	89 999
Ru Sentiment	10 113	3 912	12 720	26 745
Russian Hotel Reviews	47 339	3 492	6 373	57 204
Senti Ru Eval 2015	2 445	4 255	12 136	18 836
Senti Ru Eval 2016	484	1 770	3 246	5 500
Итого	93 962	62 434	85 001	241 397

В качестве тестовых данных в данной работе рассматривался корпус RuNews, который включает 1 823 новостные статьи, размеченные на три класса тональности: 147 позитивные, 550 негативные и 1 126 нейтральные [8].

Подбор количества эпох обучения нейросетевой модели производился при отделении 20% валидационных данных от обучающих. Лучшие результаты модель показала на двух эпохах.

Далее подбирался наиболее подходящий для данной задачи словарь. Использовались словари из статьи [5], в которой разрабатывался универсальный словарь оценочной лексики. В названии словарей числа обозначают количество исходных словарей, в которых данное слово встретилось с одной и той же оценкой: например, dict 3+ означает, что этот словарь был сформирован за счет слов, которые встречаются с одинаковой оценкой не менее чем в трех существующих словарях оценочной лексики. Подбор оптимального словаря для целей настоящего исследования проводился по метрике f1-score, для этого использовалась адаптация словарного метода SO-CAL для русского языка [9]. Наиболее подходящим оказался словарь dict 5+ (f1-score = 0,53), содержащий 1 181 оценочное слово (449 позитивных и 732 негативных). Для модели ruRoberta также было получено значение метрики f1-score = 0,63.

Для дальнейшего анализа были сопоставлены эталонная оценка текста и результаты двух методов — нейросетевой модели ruRoberta и словарного метода SO-CAL со словарем dict 5+. После этого производилось разделение текстов на следующие группы. Название группы составлялось как конкатенация значений «Pos», «Neg», «Neut» или «X» для эталонной оценки, предсказания ruRoberta и предсказания словарного метода. Значение «X» означает, что оценка может быть любой. Например, группа PosNegX будет состоять из текстов, имеющих эталонную оценку «позитивный», при этом языковая модель присвоила им оценку «негативный», а оценка SO-CAL в данном случае может быть любой.

В эксперименте рассматривались три больших группы — PosXX, NegXX и NeutXX, они составлялись путем разделения по эталонной оценке. Каждая из этих групп дополнительно делилась по предсказанию, полученному методом SO-CAL, — так сформировано еще по 3 вложенных группы. В итоге рассматривались 12 групп и весь набор данных.

Вычислялось внимание, усредненное по всем головам нейросетевой модели, уделяемое каждому оценочному слову на первом и последнем слое, в соответствии с методикой, предложенной в работе [10]. Оценочные слова были взяты из словаря dict5+, который использовался методом SO-CAL. Величины внимания считались отдельно для позитивной и негативной лексики, чтобы далее можно было вычислить сумму внимания, уделяемого моделью всем позитивным и негативным словам в тексте. Предполагалось, что существует корреляция между полученными величинам и тональностью, которую языковая модель ruRoberta присвоила тексту. Полученные коэффициенты корреляции представлены в табл. 2.

Наибольшие положительные коэффициенты корреляции весов позитивных слов с оценкой ruRoberta для первого и последнего слоя наблюдаются в группе PosXPos (0,47 и 0,42). Практически ни в одной из групп веса позитивных слов не имеют отрицательную корреляцию с оценкой языковой модели.

Для весов негативных слов наибольшие отрицательные коэффициенты корреляции наблюдаются при рассмотрении всех данных (0,37 и 0,33). Примечательно, что для группы NegXPos корреляции весов негативных слов имеют положительное значение, причем весьма значительное — 0,19 и 0,28 соответственно для первого и последнего слоя модели.

В группе PosXNeg мало текстов (всего 13), всем им ruRoberta присвоила одно и то же значение («0»), в связи с чем значение коэффициента корреляции не вычисляется.

Введем дополнительно пороги в 15% по абсолютной величине, чтобы отбросить случаи незначительных корреляций. В этом случае позитивные веса только в 4 случаях из 12 имеют достаточно значимую корреляцию с оценкой нейронной языковой модели. Причем из таблицы видно, что эти случаи идентичны как для первого, так и для последнего слоя модели в одной группе (эти случаи выделены жирным шрифтом в табл. 2).

Негативные веса в 7 случаях из 12 имеют корреляцию, в абсолютном выражении превышающую 15%. Как и для случаев с позитивными весами, в 6 из 7 случаев значительные корреляции наблюдаются сразу на первом и на последнем слое модели в одной группе.

Таблица 2

Корреляция оценки ruRoberta

Тип данных	*Объем данных*	*Позитивные слова*		*Негативные слова*	
		Первый слой	*Последний слой*	*Первый слой*	*Последний слой*
Все данные	1 823	**0,24239**	**0,21230**	**-0,36996**	**-0,33000**
PosXX	147	**0,37100**	**0,21389**	-0,12666	-0,11023
PosXPos	86	**0,46904**	**0,42329**	**-0,20088**	**-0,15636**
PosXNeg	13	NaN	NaN	NaN	NaN
PosXNeut	48	0,02183	-0,08018	0,05567	**0,16910**
NegXX	550	-0,00464	0,03009	**-0,20284**	**-0,19688**
NegXPos	28	**0,19041**	**0,19340**	**0,19104**	**0,27512**
NegXNeg	441	-0,05484	-0,00954	**-0,19055**	**-0,18544**
NegXNeut	81	-0,03752	-0,03371	**-0,19474**	**-0,20957**
NeutXX	1 126	0,11482	0,09158	-0,14321	-0,12299
NeutXPos	251	0,09801	0,07807	-0,06717	-0,11972
NeutXNeg	349	0,10467	0,08568	**-0,15814**	-0,11419
NeutXNeut	526	0,07737	0,05011	-0,00819	-0,00495

Выводы. Усредненные величины внимания, уделяемые нейросетевой языковой моделью ruRoberta каждому оценочному слову на первом и последнем слое, имеют корреляцию с присваиваемой тональностью. Для позитивной лексики это наблюдалось лишь в 33,3% случаев, в то время как для негативной лексики — в 58,3%. Это говорит о том, что модель ruRoberta уделяет больше внимания негативно окрашенной оценочной лексике.

СПИСОК ЛИТЕРАТУРЫ

1. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015. P. 80-83. — Text: electronic.
2. Devlin, J. BERT: Pre training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova // Proceedings of 7th Annual Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). — 2019. — Pp. 4171–4186. — Text: electronic.
3. Vaswani, A. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones et al. // Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). — 2017. — Vol. 30. — Pp. 6000–6010. — Text: electronic.
4. Kotelnikova, A.V. Lexicon-based Methods and BERT Model for Sentiment Analysis of Russian Text Corpora / A.V. Kotelnikova, D.E. Paschenko, E.V. Razova // Information Technologies and Intelligent Decision Making Systems 2021 (ITIDMS-II-2021). CEUR Workshop Proceedings. — 2021. — Vol. 2922. — Pp. 73–81. — Text: electronic.
5. Kotelnikova, A., Paschenko, D., Bochenina, K., Kotelnikov, E. (2022). Lexicon-Based Methods vs. BERT for Text Sentiment Analysis. In: , et al. Analysis of Images, Social Networks and Texts. AIST 2021. Lecture Notes in Computer Science, vol 13217. Springer, Cham. https://doi.org/10.1007/978-3-031-16500-9_7. — Text: electronic.
6. Sberbank-ai/ruRoberta-large [сайт]. — URL: https://huggingface.co/sberbank-ai/ruRoberta-large (дата обращения: 01.05.2023). — Режим доступа: свободный. — Текст: электронный.
7. Russian SuperGLUE: Leaderboard [сайт]. — URL: https://russiansuperglue.com/leaderboard/2 (дата обращения: 02.05.2023). — Режим доступа: свободный. — Текст: электронный.
8. Kotelnikova A.V., Vychegzhanin S.V., Kotelnikov E.V. Cross-domain sentiment analysis based on small in-domain fine-tuning // IEEE Access. 2023. Vol. 11. P. 41061–41074. — Text: electronic.
9. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37(2), 267–307 (2011). — Text: electronic.
10. Razova E., Vychegzhanin S., Kotelnikov E. Does BERT look at sentiment lexicon? // 10th International Conference on Analysis of Images, Social Networks and Texts (AIST-2021). Communications in Computer and Information Science (CCIS). Vol. 1573. P. 55–67. — Text: electronic.