АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 811.161.1’322.2

А.А. Голубев, Н.В. Лукашевич

Исследование моделей нейронных сетей типа BERT для анализа тональности текстов на русском языке*

Описываются результаты применения стандартных архитектур нейронных сетей (CNN, LSTM, BiLSTM) и недавно появившихся их моделей BERT на ранее размеченных данных для анализа тональности текстов на русском языке. Сравниваются два варианта русскоязычной модели BERT (RuBERT) и различные способы ее применения.

Ключевые слова: *анализ тональности, русский язык, нейронные сети, архитектура BERT*

DOI: 10.36535/0548-0027-2021-01-3

ВВЕДЕНИЕ

В Интернете и в социальных сетях огромное количество людей излагает свои мнения на различные темы. Автоматическое извлечение и анализ этих мнений представляют значительный интерес для дальнейшего развития научных исследований в области русского языка.

Современные подходы к решению задач анализа тональности текстов основаны на применении методов машинного обучения, требующих наличия специальных обучающих и тестовых наборов данных (коллекций). Наибольшее количество различных размеченных коллекций анализа тональности создано для английского языка, например, Stanford Sentiment Treebank SST [1], коллекция отзывов пользователей о фильмах IMDB [2], коллекции твитов, размеченных по тональности в рамках тестирования SemEval [3, 4].

Для других языков существует значительно меньше размеченных данных. Примерами русскоязычных размеченных коллекций для анализа тональности являются наборы данных с тестирований РОМИП 2012-2013 и SentiRuEval 2015-2016 [5-7], включающие размеченные отзывы пользователей на фильмы, книги и цифровые камеры, а также цитаты из новостей и короткие сообщения из социальной сети Твиттер. Лучшие результаты анализа тональности текстов среди участников тестирования были получены с применением классических подходов машинного обучения: метода опорных векторов (SVM) [5], ранних нейронных архитектур [8], а также инженерных подходов, основанных на правилах [9].

Повышение качества автоматического анализа тональности связано с применением относительно новых архитектур нейронных сетей, например, BERT [10]. Поэтому можно предположить, что с использованием новых подходов результаты на русскоязычных коллекциях данных могут быть значительно улучшены.

В настоящей работе сравниваются несколько подходов к анализу тональности на основе модели BERT [10], хорошо зарекомендовавшей себя при решении широкого спектра задач по автоматической обработке текстов. Подходы на основе этой модели сопоставляются с методом машинного обучения SVM, а также со стандартными нейронными архитектурами (CNN, LSTM, BiLSTM). Лучшие результаты анализа продемонстрировал вариант модели BERT, обученный на разговорных текстах – диалоги, комментарии и др. [11]. Лучшей архитектурой оказалась модель BERT-NLI, рассматривающая задачу классификации по тональности как задачу текстового вывода (Natural Language Inference) [12]. На одной из задач эта модель практически достигает уровня человеческих ответов. В той же работе проводится анализ ошибок лучших подходов, а также приводятся примеры, которые сегодня не удается правильно классифицировать автоматически.

* Исследование выполнено при финансовой поддержке РФФИ (проект 20-07-01059)

Таблица 1

Размеченные коллекции, используемые в исследовании

Коллекция	Обучающая	Тестовая	Метрика	Лучшие результаты анализа	Метод
Новости РОМИП-2013[1]	4260	5500	$F_1 macro$	62,1	Правила
SentiRuEval-2015 Операторы[2]	5000	5322	$F_1^{+-} macro$	50,3	SVM
SentiRuEval-2015 Банки[2]	5000	5296	$F_1^{+-} macro$	36,0	SVM
SentiRuEval-2016 Операторы[3]	8643	2247	$F_1^{+-} macro$	55,9	GRU
SentiRuEval-2016 Банки[3]	9392	3313	$F_1^{+-} macro$	55,1	GRU

Таблица 2

Распределение классов по наборам данных (%)

Коллекция	Обучающая выборка, классы			Тестовая выборка, классы		
	Полож.	Отриц.	Нейтр.	Полож.	Отриц.	Нейтр.
Новости РОМИП-2013	16	36	48	11	33	56
SentiRuEval-2015 Операторы	19	32	49	10	23	67
SentiRuEval-2015 Банки	7	34	59	8	15	79
SentiRuEval-2016 Операторы	15	29	56	10	46	44
SentiRuEval-2016 Банки	8	18	74	10	22	68

ИСХОДНЫЕ ДАННЫЕ ИССЛЕДОВАНИЯ

Ранее были рассмотрены пять русскоязычных наборов данных, созданных в рамках предыдущих тестирований по анализу тональности: коллекция новостных цитат РОМИП-2013 [5] и четыре коллекции размеченных сообщений (твитов) из социальной сети Твиттер с тестирования SentiRuEval 2015-2016 [6, 7]. В табл. 1 представлены описание размеченных коллекций, включающее обучающие и тестовые выборки, меры качества результатов, а также лучшие методы и продемонстрированные ими результаты анализа; в табл. 2. – распределения классов тональностей для упомянутых наборов.

Коллекция новостных цитат

В рамках тестирования РОМИП-2013 для создания коллекции новостных цитат из статей извлекались высказывания, записанные в форме прямой или косвенной речи [5]. Предполагалось, что подобные цитаты содержат высокую долю оценочных (позитивных или негативных) мнений. Задача тестирования состояла в отнесении высказываний к одному из трех классов: позитивный, негативный или нейтральный. Из табл. 2 видно, что данные по классам достаточно сбалансированы. Основной метрикой тестирования являлась макро F мера ($F_1 macro$).

Участники тестирования экспериментировали с классическими подходами машинного обучения: наивным байесовским классификатором и методом опорных векторов. Однако лучшие результаты анализа текстов были получены при использовании подхода, основанного на знаниях и правилах: 62,1 F_1 меры и 61,6 Accuracy. Это может объясняться широким разнообразием тем и используемой оценочной лексикой в цитатах, а также недостаточным объемом обучающей выборки [5, 9].

Коллекция твитов

В рамках двух тестирований SentiRuEval 2015-2016 [6, 13] рассматривалось задание по мониторингу репутации компаний на примере банков и телекоммуникационных операторов, заключавшееся в поиске мнений с положительной или отрицательной тональностью, а также положительных и отрицательных фактов о компаниях. Таким образом, задание можно рассматривать как задачу анализа тональности с ори-

[1] http://romip.ru/en/collections/sentiment-news-collection-2012.html

[2] https://drive.google.com/drive/folders/0BxlA8wH3PTU-ffll5LUM0SmVvZ1puc2NaalQtWmdEbEw1Yi0zYkl1cjdDN-2puel FIRDBHdVU

[3] https://drive.google.com/drive/folders/0BxlA8wH3PTU-fV1F1UTBwVTJPd3c

ентацией на заданный объект (*entity-oriented sentiment analysis*, *targeted sentiment analysis* [14]). За два года исследований создано четыре набора данных (см. табл. 1). В 2016 г. обучающие наборы были получены с помощью объединения прошлогодних обучающих и тестовых данных. Это позволило существенно увеличить размеры обучающих выборок [7].

Участникам тестирования было необходимо провести классификацию по трем классам. Из табл. 2 видно, что нейтральный класс превалирует во всех коллекциях. По этой причине главной метрикой качества была выбрана $F_1^{+-}macro$ мера, вычисляемая как среднее значение между F_1 мерами положительного и отрицательного классов. Аналогичная F_1 мера нейтрального класса игнорировалась, поскольку эта категория чаще всего не представляет интереса. Стоит уточнить, что выбор этой метрики не сводит задачу к бинарной классификации, поскольку ошибки на нейтральном классе негативно влияют на F_1^+ и F_1^- метрики. В дополнение к этому подсчитывалась мера $F_1^{+-}macro$, учитывающая неравенство объемов положительного и отрицательного классов [6, 7].

Как показано в табл. 1, результаты анализа тональности в тестировании 2016 г. значительно превышают результаты 2015 г. на аналогичных задачах. Это объясняется как увеличением обучающей выборки для второго соревнования, так и использованием участниками исследования более продвинутых методов, включая современные нейронные сети на основе векторных представлений слов (эмбеддингов) [7].

МЕТОДЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ

В настоящей работе исследовались классические архитектуры нейронных сетей для автоматической обработки текстов – сверточная и рекуррентная, – а также нейронная сеть архитектуры BERT.

В качестве базового подхода (*baseline*) для экспериментов был выбран метод опорных векторов, использующий предобученные эмбеддинги для описания признаков. Вариант эмбеддингов от FastText[4] выбран из-за лучших результатов в исследовании по сравнению с эмбеддингами ELMo[4], Word2Vec[5] и GloVe[6]. Для подачи данных в SVM использовался подход усреднения эмбеддингов по предложению. Для поиска оптимальных гиперпараметров применялся механизм подбора параметров библиотеки scikit-learn[7].

Предобработка текстов для всех наборов данных состояла из следующих шагов:

- приведение к нижнему регистру;
- замена ссылок токеном *ссылка*;
- замена упоминаний пользователей токеном *пользователь*;
- замена хэштегов токеном *хэштег*;
- замена адресов электронной почты токеном *почта*;
- замена телефонных номеров токеном *телефон*;
- замена всех эмотиконов соответствующими смысловыми токенами;
- удаление всех специальных символов за исключением пунктуации;
- сокращение числа повторений любой буквы подряд до двух раз;
- лемматизация и удаление стоп-слов.

Заключительный пункт проводился только для метода опорных векторов и классических нейронных сетей. Для моделей, основанных на архитектуре BERT, это не привело к изменениям и дало лишь незначительный прирост в районе 0,01%.

Классические нейронные сети

Архитектура сверточной сети, используемая в нашем исследовании, основана на подходах, описанных в [15, 16]. Входные данные представляются в виде матрицы размера $s \times d$, где s – количество токенов в предложении, а d – размерность пространства эмбеддингов. Оптимальная высота матрицы $s = 50$ была выбрана экспериментальным путем. При необходимости предложение обрезалось или дополнялось нулевыми векторами до требуемой длины.

Далее операции свертки с фильтрами различных размеров применяются к матрице параллельно. Одному фильтру соответствует матрица $w \in R^{h \times d}$ с размером фильтра h, равным числу слов, которое этот фильтр покрывает. Затем к результатам использования каждого фильтра применяется операция *max-pooling*. Это помогает извлечь наиболее важную информацию независимо от положения признака в тексте. После проведения всех сверточных операций полученные векторы конкатенируются и отправляются на полносвязный слой, результат которого передается на слой *softmax* для получения конечного распределения вероятностей по классам. Число сверток было выбрано равным четырем с окнами (2, 3, 4, 5) соответственно. Для снижения эффекта переобучения было добавлено два слоя *dropout* с параметром $p = 0{,}5$: первый – после слоя *max-pooling*, второй – после полносвязного слоя.

Другой известной архитектурой нейронной сети, часто используемой для автоматической обработки текстов, является рекуррентная сеть LSTM. Основная ее идея – введение специальной внутренней ячейки состояния размерности $c_t \in R^d$, равной размерности скрытого слоя сети. В рассматриваемом нами случае размерность d была выбрана равной размерности пространства эмбеддингов.

Одним из недостатков архитектуры сети LSTM считается невозможность учитывать информацию идущих впереди слов, поскольку предложение читается сетью только в прямом направлении [17]. Для решения этой проблемы обычно используется двунаправленная BiLSTM сеть. Две эти LSTM сети проходят по предложению с двух сторон, а их ячейки состояния конкатенируются и позволяют получить вектор размерности $c_t \in R^{2d}$. Как и в случае обыч-

[4] http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html
[5] https://rusvectores.org/ru/models/
[6] http://www.cs.cmu.edu/~afm/projects/multilingual_embeddings.html
[7] https://scikit-learn.org/stable/

ной LSTM, этот вектор подается на полносвязный слой размерности 40, выход которого передается далее на слой *softmax* для получения итоговых вероятностей принадлежности классам.

В обеих рекуррентных архитектурах для борьбы с переобучением применялись *dropout* слои с параметром вероятности $p = 0{,}5$, расположенные до и после полносвязных слоев. Для всех описанных архитектур были использованы предобученные русскоязычные эмбеддинги FastText размерности 300.

Настройка моделей BERT

Подходы, используемые для анализа тональности на основе BERT [10], могут быть разделены на две группы: 1) классификация по одному предложению; 2) классификация с использованием вспомогательных предложений [18], в процессе которой задача анализа тональности преобразуется в задачу классификации пары предложений. Модель BERT способна принимать на вход как одно, так и два предложения, рассматривая их как задачу оценки смысловой связности.

Для задачи объектно-ориентированного анализа тональности для каждого объекта оценки в тексте имеется метка тональности. Это дает возможность маскировать реальное название сущности специальным токеном. Например, исходный текст "*Сбербанк – надежное место для хранения ваших сбережений"* преобразуется к виду "*MASK надежное место для хранения ваших сбережений"*. В случае наличия в предложении двух сущностей задача решается дважды с поочередным маскированием каждой из них.

Модель с одним предложением (BERT-single) использует только исходный текст и представляет стандартную архитектуру BERT с дополнительным линейным слоем с матрицей $W \in R^{K \times M}$. Здесь K определяет количество классов, а M – размерность выходного представления исходного предложения. Входное представление текста конструируется в виде суммы эмбеддинга токена, а также позиционных и разделяющих эмбеддингов. Итоговое распределение вероятностей вычисляется с помощью слоя *softmax*. Модель этой архитектуры проводит классификацию на основе контекста маскированной сущности.

Модель с двумя предложениями (BERT-pair) имеет некоторые отличия. Входное представление преобразует два предложения в одно путем добавления токена *[SEP]* между ними. Для задачи классификации в начало входной последовательности добавляется специальный токен *[CLS]*. Поверх выходного представления данного токена размерности H добавляется линейный слой классификации с матрицей $W \in R^{K \times H}$. В этом случае модель основывается не только на контексте, но и на дополнительном вопросе.

Модели, использующие вспомогательные предложения, основаны на задачах ответа на вопрос (QA) и вывода из текста (NLI):

- pair-NLI: *"Тональность MASK равна"*
- pair-QA: *"Что вы думаете о тональности MASK?"*

Следствием к предложению должен являться один из элементов множества *Положительно, Отрицательно, Нейтрально*.

В случае задачи общего анализа тональности метки принадлежат не объектам, а целым предложениям. Поскольку объекты для маскирования отсутствуют, было решено присваивать токен целому предложению. Таким образом, исходное предложение *"56% Rambler Group было продано Сбербанку"* преобразуется к виду *"MASK = 56% Rambler Group было продано Сбербанку"*. Для этого варианта задачи конструируются аналогичные вспомогательные предложения.

В настоящей статье сравниваются две предобученные модели BERT из библиотеки DeepPavlov [19]:

1) RuBERT, Russian, cased, 12-layer, 768-hidden, 12-heads, 180M параметров, обучение на русской части Википедии и новостях[8];

2) Conversational (диалоговый) RuBERT, Russian, cased, 12-layer, 768-hidden, 12-heads, 180M параметров, обучение на различных форумах, Пикабу и подкорпусе постов социальных сетей корпуса Тайга[8].

Обучение моделей производилось параметрами: вероятность на слое dropout 0,1; число эпох 5; начальный шаг обучения размер батча, равный 12.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Для сравнения рассматриваемых моделей вычислялись стандартные метрики: точность (Accuracy) и $F_1 macro$. Помимо них были вычислены метрики, необходимые для сравнения с участниками соревнований: $F_1^{+-} macro$ и $F_1^{+-} micro$, учитывающие только положительный и отрицательный классы. Все представленные результаты работы нейронных сетей являются усреднением по пяти попыткам. Для разделения двух предобученных BERT моделей используется метка (C) для диалогового (*conversational*) варианта.

Результаты моделей на коллекции новостных цитат

Результаты использования моделей на коллекции новостных цитат РОМИП-2013 продемонстрированы в табл. 3. Как отмечалось ранее, участники соревнования применяли традиционные методы машинного обучения (метод опорных векторов, наивный байесовский классификатор и т.д.), а также инженерные методы, основанные на знаниях и правилах и продемонстрировавшие лучшие результаты.

Рассматриваемое задание оказалось сложным даже для моделей, применяющих эмбеддинги (SVM, CNN, LSTM, BiLSTM). Среди классических нейронных сетей двунаправленная BiLSTM архитектура показала лучшие результаты, при этом лишь незначительно повысив результаты инженерного подхода по F_1 метрике. Использование модели BERT радикальным образом повышает результаты. Лучшей конфигурацией является модель BERT-pair-NLI на базе диалоговой модели RuBERT. Модель BERT-pair-NLI присваивает маскирующий токен всему предложению и трактует задачу анализа тональности как формирование вывода по тексту (*textual entailment*). Результаты этой модели выше результатов BiLSTM на 9 процентных единиц.

[8] http://docs.deeppavlov.ai/en/master/features/models/bert.html

Таблица 3

Результаты моделей на коллекции новостных цитат

Модель	Accuracy	$F_1 macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
РОМИП-2013 [5]	61,60	62,10	–	–
SVM	69,12	61,63	74,82	75,07
CNN	68,57	60,43	73,51	74,55
LSTM	73,61	62,31	77,02	78,20
BiLSTM	74,14	62,78	77,61	78,94
BERT-single	78,90	68,07	84,33	84,45
BERT-pair-QA	79,06	68,54	84,33	84,45
BERT-pair-NLI	79,68	69,45	84,96	85,08
BERT-single (C)	79,81	**71,12**	85,05	85,10
BERT-pair-QA (C)	78,95	70,16	84,71	84,83
BERT-pair-NLI (C)	**80,28**	70,62	**85,52**	**85,68**

Таблица 4

Результаты моделей на коллекции твитов SentiRuEval-2015 (операторы связи)

Модель	Accuracy	$F_1 macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
SentiRuEval-2015 [6]	–	–	48,80	53,60
SVM	62,86	58,29	50,27	54,70
CNN	60,80	57,52	49,92	53,23
LSTM	64,46	58,94	52,10	56,03
BiLSTM	65,54	59,35	53,01	56,83
BERT-single	72,48	67,04	58,43	62,53
BERT-pair-QA	74,00	67,83	58,15	62,92
BERT-pair-NLI	74,66	68,24	59,17	64,13
BERT-single (C)	76,55	**69,12**	61,34	66,23
BERT-pair-QA (C)	**76,63**	68,54	**63,47**	**67,51**
BERT-pair-NLI (C)	76,40	68,83	63,14	67,45
Manual	–	–	70,30	70,90

Результаты моделей на коллекции твитов

Результаты применения моделей на двух коллекциях твитов с соревнования SentiRuEval-2015 отражены в табл. 4 и 5. Особенностью этого тестирования стало то, что в период между сбором обучающей и тестовой коллекций прошло около 6 месяцев 2013-2014 гг. В этот период произошли события на Украине, которые отразились в тематике твитов о проблемах операторов связи и банков, что привело к большим различиям между обучающими и тестовыми коллекциями.

Описанные сложности негативно повлияли на результаты банковской коллекции 2015 г. [6]. Задача оказалась трудной для методов SVM+FastText, CNN, LSTM и BiLSTM. Только модели на основе BERT смогли значительно улучшить результаты анализа текстов. Как и в предыдущем случае, лучшим вариантом стала диалоговая модель BERT в конфигурации задачи формирования вывода по тексту.

Один из участников SentiRuEval-2015 загрузил ручную разметку тестовой коллекции данных операторов связи и получил результаты, приведенные в табл. 4 (Manual) [10]. Надо отметить, что лучшие показатели моделей BERT вплотную приблизились к показателям ручной разметки.

Результаты моделей на двух коллекциях твитов с соревнования SentiRuEval-2016 представлены в табл. 6 и 7. В отличие от предыдущего конкурса, базовый порог результатов 2016 г. (лучшие показатели участников) превысил результаты SVM+FastText и CNN моделей. Это может быть объяснено тем фактом, что участники применяли нейронные сети в сочетании с эмбеддингами, а также комбинировали метод опорных векторов с существующими русскоязычными словарями оценочной лексики [7, 8]. Лучшие результаты также были получены моделями BERT на основе двух предложений, которые на 9–14 процентных пунктов превышают результаты участников тестирований 2016 г.

Таблица 5

Результаты моделей на коллекции твитов SentiRuEval-2015 (банки)

Модель	Accuracy	$F_1 macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
SentiRuEval-2015 [6]	–	–	36,00	36,60
SVM	49,23	43,39	33,08	36,62
CNN	47,91	42,87	31,62	34,18
LSTM	51,89	44,12	35,85	39,55
BiLSTM	53,21	46,43	36,93	40,18
BERT-single	83,78	74,57	57,82	60,64
BERT-pair-QA	84,24	75,34	56,65	57,41
BERT-pair-NLI	85,14	77,59	60,46	63,15
BERT-single (C)	85,80	78,71	64,90	66,95
BERT-pair-QA (C)	86,28	78,62	62,37	67,27
BERT-pair-NLI (C)	**86,88**	**79,51**	**67,44**	**70,09**

Таблица 6

Результаты моделей на коллекции твитов SentiRuEval-2016 (операторы связи)

Модель	Accuracy	$F_1 macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
SentiRuEval-2016 [7]	–	–	55,94	65,69
SVM	65,89	55,34	53,13	65,87
CNN	65,28	54,87	52,62	64,40
LSTM	66,71	56,74	56,93	67,18
BiLSTM	67,30	57,11	57,23	67,93
BERT-single	72,85	65,12	60,29	71,70
BERT-pair-QA	74,24	66,34	63,86	73,26
BERT-pair-NLI	74,51	67,48	62,81	73,39
BERT-single (C)	75,20	67,89	64,96	73,91
BERT-pair-QA (C)	75,27	68,11	65,91	**74,22**
BERT-pair-NLI (C)	**75,71**	**68,42**	**66,07**	74,11

Таблица 7

Результаты моделей на коллекции твитов SentiRuEval-2016 (банки)

Модель	Accuracy	$F_1 macro$	$F_1^{+-} macro$	$F_1^{+-} micro$
SentiRuEval-2016 [7]	–	–	55,17	58,81
SVM	66,46	57,85	51,12	53,74
CNN	67,15	58,43	52,06	54,96
LSTM	70,80	61,17	57,22	59,71
BiLSTM	71,44	61,86	58,40	61,06
BERT-single	81,20	73,21	68,19	69,56
BERT-pair-QA	80,35	72,61	66,61	68,18
BERT-pair-NLI	80,91	72,68	65,62	67,65
BERT-single (C)	80,47	72,59	66,95	69,46
BERT-pair-QA (C)	**82,28**	**74,06**	**69,53**	**71,76**
BERT-pair-NLI (C)	81,28	73,34	65,82	68,03

АНАЛИЗ ОШИБОК МЕТОДОВ

Качество анализа тональности на рассмотренных наборах данных значительно улучшили методы на основе нейронных сетей. Вместе с тем сохранилась значимая доля ошибок. Для анализа были извлечены примеры, в которых ошиблись все методы, т.е. предсказали не ту тональность, которую указали аннотаторы. Статистические данные по коллекциям примеров, в которых ошиблись все методы, продемонстрированы в табл. 8. Интересно отметить, что доля подобных твитов среди коллекций телекоммуникационных операторов существенно выше, чем среди банковских.

Для сравнения были извлечены примеры, в которых только одна модель предсказала тональность правильно. Среди этих моделей не оказалось метода опорных векторов и классических нейронных сетей. Лучшей конфигурацией по усредненному результату стала BERT-single на основе Conversational RuBERT модели. В табл. 9 представлено распределение подобных твитов по методам и коллекциям (для экономии места BERT архитектуры обозначены соответствующими сокращениями из табл. 3–7).

Примеры, в которых ошиблись все модели, могут быть разделены на несколько групп.

Первая группа включает очень короткие примеры, тональность в которых выражается одним, возможно, неоднозначным словом. Например, в следующих примерах аннотатор ставит отрицательную тональность, а модели предсказывают нейтральную.

- *Альфа-клик у всех лежит, да?*
- *Сбербанк навязывает кредитную карту.*

Во второй группе ошибки возникают из-за использования автором в твите количественных величин.

- *Расчетно-кассовое обслуживание Сбербанке стоит 2,5 т.р./месяц, в Татфондбанке 150 целковых.*
- *Пока ждем сотрудника Сбербанка, могла уже 3 раза сходить пообедать.*
- *Нормально @sberbank зарабатывает – размен 5% от суммы.*
- *Сбербанк делает мою карту уже 1 месяц и 7 дней*

Для определения тональности высказывания в таких случаях требуются специальные знания, которые невозможно получить из обучающей выборки.

В третьей группе примеров в твитах упоминается несколько сущностей с разной тональностью. В отчетах организаторов прошедших тестирований [9, 10] 2015-2016 гг. было указано, что модели не справились с подобными примерами. Большинство методов проставляло одинаковую тональность всем сущностям в твитах, но были и подходы, пытавшиеся применить более детальный анализ. При обеих стратегиях были получены не очень высокие результаты анализа таких твитов.

Современные модели на основе архитектуры BERT способны справляться со значимой долей твитов с несколькими сущностями. В табл. 10 представлены результаты моделей на подобных примерах. Однако в некоторых твитах ни одна модель не смогла отличить разные тональности по отношению к разным упоминаемым организациям. Например, в следующих твитах отношение автора к Билайну положительное, но модели ставят отрицательную тональность и Билайну, и МТС.

- *Я сдаюсь! И перехожу после НГ на Билайн, сохранив старый номер. Ибо МТС интернет жутко лютый, причем в любой точке города.*
- *МТС не работает! Вечно вне зоны доступа. Связь постоянно прерывается. Всю семью переводим на Билайн.*

Таблица 8

Статистика по примерам, на которых ошиблись все модели

Коллекция	Объём	Ошибки	Доля (%)
SentiRuEval-2015 Операторы	4173	621	14,88
SentiRuEval-2015 Банки	4613	213	4,62
SentiRuEval-2016 Операторы	2460	345	14,02
SentiRuEval-2016 Банки	3418	306	8,95
Всего	14664	1485	10,13

Таблица 9

Статистика по примерам, для которых только одна модель дала правильный ответ

Коллекция	Кол-во	Доля (%)	BS	BPQ	BPN	BS-C	BPQ-C	BPN-C
Операторы 2015	160	3,83	16,25	18,75	20,00	15,00	14,37	15,63
Банки 2015	222	4,81	11,71	4,05	17,57	50,45	4,05	12,17
Операторы 2016	101	4,11	19,8	12,87	13,86	22,78	17,82	12,88
Банки 2016	113	3,31	6,19	24,78	28,32	10,62	12,39	7,7
Всего	596	4,06	1,49	15,11	19,94	24,71	12,16	14,6

Таблица 10

Результаты моделей на примерах, содержащих две различные тональности

Коллекция	Кол-во	Доля (%)	BS	BPQ	BPN	BS-C	BPQ-C	BPN-C
Операторы 2015	59	1,41	6,78	3.39	6,78	3,39	5,08	6,78
Банки 2015	14	0,3	7,14	0	10	0	0	0
Операторы 2016	43	1,75	6,98	6.98	9,3	9,3	13,96	9,3
Банки 2016	11	0,32	0	0	0	18,18	18,18	0
Всего	127	0,95	5.23	2,59	4,02	7,72	9,31	4,02

В этом случае лучшие результаты по усреднению демонстрирует конфигурация BERT- pair-QA на основе модели Conversational RuBERT.

Еще одна – четвертая группа сообщений содержит явные оценки или эмоции автора сообщения, но не по отношению к упоминаемому банку или оператору – здесь аннотаторы твитов проставляли нейтральную тональность, а модели определяли превалирующую тональность.

- *Сейчас в Сбербанке бабушка с Альцгеймером пыталась снять деньги со счёта. Угнетающее зрелище, дерьмовая болезнь.*
- *Этот момент, когда ты не успел до закрытия Сбербанка на несколько минут, сейчас застрял у бабушки, надо дойти до киви, а телефон сейчас вырубится*
- *Когда придут счета от МТС, нужно просто объявить что симку похитили. Это вроде спасает, когда уходишь в минус.*

Как и ожидалось, есть ещё и пятая группа ироничных высказываний, которые выглядят по лексике позитивно или нейтрально, а на деле имеют негативную окраску:

- *Отлично, моя карта со стипухой в другом Сбербанке... Проехать полгорода и узнать об этом – всегда мечтала прям.*
- *В следующий раз возьму с собой в Сбербанк вязание.*

ДРУГИЕ РАЗМЕЧЕННЫЕ КОЛЛЕКЦИИ И ПОДХОДЫ ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Наиболее объемным набором данных для анализа тональности текстов на русском языке обладает коллекция RuSentiment [20], содержащая более 30 тыс. постов из социальной сети «ВКонтакте». Каждый пост отнесен к одному из трех классов (позитивный, негативный, нейтральный) по тональности. Для классификации сообщений по тональности авторы этой коллекции применяли классические методы машинного обучения (логистическая регрессия, метод опорных векторов с линейным ядром, градиентный бустинг) и нейронные сети. Лучший результат был получен с помощью четырехслойной полносвязной нейронной сети в сочетании с эмбеддингами FastText, предобученными на тех же данных и составил 71,7 F_1 меры. В работе [11] авторы получили 87,73 F_1 меры с использованием мультиязычных моделей BERT и RuBERT, предобученных на русскоязычных текстах.

Другой известный набор данных – это коллекция твитов, автоматически размеченных по эмотиконам (RuTweetCorp) [21]. Эта коллекция содержит более 200 тыс. постов 2013-2014 гг., размеченных по тональности на два класса.

В работе [22] к данным RuTweetCorp применялся метод опорных векторов с эмбеддингами Word2Vec. Авторы [23] тестировали нейросетевые модели LSTM+CNN и BiGRU на коллекциях RuSentiment и RuTweetCorp. А. Звонарев и А. Билый в [24] представили более высокие результаты с использованием сверточной нейронной сети по сравнению с логистической регрессией и XGBoost классификатором на данных RuTweetCorp.

В [25] описан корпус RuSentRel, включающий аналитические статьи, посвященные международным отношениям. В корпусе размечено авторское отношение к упоминаемым именованным сущностям, а также отношения упоминаемых именованных сущностей между собой. В работе [26] исследовались подходы к распознаванию оценочных отношений между упоминаемыми сущностями с использованием сверточных нейронных сетей и подхода опосредованного обучения (*distant supervision*) на данных корпуса RuSentRel.

ЗАКЛЮЧЕНИЕ

В настоящей работе были протестированы стандартные нейронные архитектуры (CNN, LSTM, BiLSTM) и современные BERT-модели на русскоязычных коллекциях данных, подготовленных в рамках ранее проведенных тестирований по анализу тональности текстов. Помимо стандартной BERT архитектуры исследованы модели, сводящие задачу анализа тональности к задаче ответа на вопрос (QA), а также к задаче формирования вывода по тексту (NLI). Сравнивались два варианта предобученных русскоязычных моделей BERT. Было показано, что на большинстве задач диалоговый вариант модели BERT, обученный на текстах социальных сетей, демонстрирует лучшие результаты анализа тональности. Наиболее удачной конфигурацией стала BERT-NLI архитектура, рассматривающая задачу классификации по тональности как задачу логического вывода по тексту. На одном наборе данных модель практически достигла результатов человеческого уровня. Исходный код программы (https://github.com/ antongolubev5/Targeted-SA-for-Russian-Datasets) и используемые в работе данные (https://github.com/LAIR-RCC/ Russian-Sentiment-Analysis-Evaluation-Datasets) находятся в публичном доступе.

СПИСОК ЛИТЕРАТУРЫ

1. Socher R., Perelygin A., Wu J., Chuang J., Manning C.D., Ng A.Y., Potts C. Recursive deep models for semantic compositionality over a sentiment treebank // Proceedings of the 2013 conference on empirical methods in natural language processing. – Seattle: Association for Computational Linguistics, 2013. – P. 1631–1642.
2. Maas A., Daly R., Pham P., Huang D., Ng A.Y., Potts C. Learning word vectors for sentiment analysis // Proceedings of the 49th annual meeting of the association for computational linguistics. Vol. 1. – Portland: Association for Computational Linguistics, 2011. − P. 142–150.
3. Nakov P., Ritter A., Rosenthal S., Sebastiani F., Stoyanov V. Semeval-2016 task 4: Sentiment analysis in twitter // Proceedings of the 10th International Workshop on Semantic Evaluations SemEval-2016. – San Diego: Association for Computational Linguistics, 2016. – P. 502–518.
4. Rosenthal S., Farra N., Nakov P. Semeval-2017 task 4: Sentiment analysis in twitter // Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017). – Vancouver: Association for Computational Linguistics, 2017. – P. 502–518.
5. Chetviorkin I., Loukachevitch N. Evaluating sentiment analysis systems in Russian // Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing. – Sofia: Association for Computational Linguistics. – 2013. – P. 12–17.
6. Loukachevitch N., Rubtsova Y. Entity-oriented sentiment analysis of tweets: results and problems // International Conference on Text, Speech, and Dialogue. – Cham: Springer, 2015. – P. 551−559.
7. Loukachevitch N., Rubtsova Y. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // Proceedings of International Conference Dialog-2016. – Moscow: Russian State University for Humanities, 2016. – P. 416-426.
8. Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of neural network architectures for sentiment analysis of Russian tweets. // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue. – Moscow: Russian State University for Humanities, 2016. – P. 50−59.
9. Kuznetsova E., Loukachevitch N., Chetviorkin I. Testing rules for a sentiment analysis system // Proceedings of International Conference Dialog. – Moscow: Russian State University for Humanities, 2013. – P. 71−80.
10. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. – Minneapolis: Association for Computational Linguistics, 2019. – P. 4171—4186.
11. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for russian // Proceedings of International Conference Dialog. – Moscow: Russian State University for Humanities, 2019. – P. 333-339.
12. Bowman S. R., Angeli G., Potts C., Manning C. D. A large annotated corpus for learning natural language inference // Proceedings of EMNLP-2015. – Lisbon: Association for Computational Linguistics, 2015. – P. 632-642.
13. Amigo E., De Albornoz J.C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., De Rijke M., Spina D. Overview of replab 2013: Evaluating online reputation monitoring systems // International conference of the cross-language evaluation forum for European languages – Berlin, Heidelberg: Springer, 2013. – P. 333−352.
14. Zhang M., Zhang Y., Vo D.T. Gated neural networks for targeted sentiment analysis // Thirtieth AAAI Conference on Artificial Intelligence. – Phoenix: AAAI Press, 2016. – P. 3087–3093.
15. Cliche M. BB twtr at SemEval-2017 task 4: Twitter Sentiment Analysis with CNNs and LSTMs // Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). – – Vancouver: Association for Computational Linguistics, 2017. – P. 573−580.
16. Zhang Y., Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification // Proceedings of the 8th International Joint Conference on Natural Language Processing. – Taipei, Taiwan: AFNLP, 2015. – P. 253 −263.
17. Chiu J. P. C., Nichols E. Named entity recognition with bidirectional LSTM-CNNs // Transactions of the Association for Computational Linguistics. – 2016. – Vol. 4. – P. 357-370.
18. Sun C., Huang L., Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentences // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Minneapolis: Association for Computational Linguistics, 2019. – P. 380−385.
19. Burtsev M. et al. DeepPavlov: Open-Source Library for Dialogue Systems // Proceedings of ACL 2018, System Demonstrations. – Melbourne: Association for Computational Linguistics, 2018. – P. 122−127.
20. Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. Rusentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th International Conference on Computational Linguistics. – Santa Fe: Association for Computational Linguistics, 2018. – P. 755−763.
21. Rubtsova Y. Constructing a corpus for sentiment classification training. Software and Systems (109) // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016". – Moscow: Russian State University for Humanities, 2016. – P. 72−78.

22. Rubtsova Y. Reducing the deterioration of sentiment analysis results due to the time impact // Information. – 2018. – Vol. 9, № 8. – P. 184-194.
23. Svetlov K., Platonov K. Sentiment analysis of posts and comments in the accounts of russian politicians on the social network // 25th Conference of Open Innovations Association (FRUCT) – Helsinki: IEEE, 2019. – P. 299–305.
24. Zvonarev A., Bilyi A. A comparison of machine learning methods of sentiment analysis based on Russian language twitter data // The 11th Majorov International Conference on Software Engineering and Computer Systems – Saint Petersburg: ITMO, 2019. – URL: http://ceur-ws.org/Vol-2590/short35.pdf
25. Loukachevitch N., Rusnachenko N. Extracting sentiment attitudes from analytical texts // Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2020. – Moscow: Russian State University for Humanities, 2020. – P. 459–468.
26. Rusnachenko N., Loukachevitch N., Tutubalina E. Distant supervision for sentiment attitude extraction // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). – Varna: Shaumen, 2019. – P. 1022–1030.

Материал поступил в редакцию 21.10.20.

Сведения об авторах

ГОЛУБЕВ Антон Александрович – студент магистратуры Московского государственного технического университета имени Н.Э. Баумана
e-mail: antongolubev5@yandex.ru

ЛУКАШЕВИЧ Наталья Валентиновна – доктор технических наук, ведущий научный сотрудник НИВЦ Московского государственного университета имени М.В. Ломоносова
e-mail: louk_nat@mail.ru
тел. +79261446163