Объектно-ориентированный анализ тональности твитов о банках и операторах сотовой связи

И.Ф. Жолобов, e-mail: ilyazh47@google.com

Вятский государственный университет

Аннотация. *В работе рассматривается задача объектно-ориентированного анализа тональности на корпусе SentiRuEval-2016, включающего твиты о банках и операторах сотовой связи. Исследуются методы на основе словарей оценочной лексики, машинного обучения и глубоких нейронных сетей. Для методов машинного обучения используются векторные представления на основе tf-idf и представления, полученные после обучения нейросетевой модели RuBERT. Полученные результаты для методов, использующих словари, хуже оценок предыдущих работ на 14–27 процентных пунктов. Методы машинного обучения с использованием tf–idf дали оценки, отличающиеся от результатов предыдущих работ не более чем на 11–13 процентных пунктов в меньшую сторону. Применение представлений, полученных на основе нейросетевой языковой модели RuBERT позволило превзойти результаты предыдущих работ на 6–11 процентных пунктов.*

Ключевые слова: *Словарь оценочной лексики, токенизация, представление текста, методы машинного обучения, нейронные сети, RuBERT.*

Введение

Люди привыкли давать оценку различным аспектам собственной жизни: своему окружению, событиям в жизни, компаниям и их услугам и товарам, которыми они пользуются, и делиться своим мнением в сети.

Компании заинтересованы в получении оценок относительно своих и конкурентных предложений на рынке. Особо ценными для них являются положительные и негативные отзывы о своем продукте – можно использовать как преимущество перед конкурентами или улучшить свой продукт, чтобы не уступать конкурентам.

Задачу объектно-ориентированного анализа тональности можно сформулировать следующим образом. Пусть задано некоторое конечное множество объектов (например, банков или операторов сотовой связи), шкала оценок относительно данных объектов (например, позитивно, негативно и нейтрально), а также высказывание, в котором упоминается

хотя бы один объект из данного множества. Объектно-ориентированный анализ тональности – поиск оценки относительно каждого объекта, упомянутого в высказывании [8].

SentiRuEval-2016 – корпус твитов, предложенный для одноименного соревнования в рамках конференции «Диалог-2016» [11]. Корпус разделен на две категории объектов: высказывания пользователей о банках и об операторах связи. Каждое сообщение содержит мнение пользователя об одном или нескольких объектах, принадлежащих к данной категории.

В настоящей работе рассматриваются методы с использованием словарей оценочной лексики, методы машинного обучения с различными способами получения векторного представления текстов, а также нейросетевая модель RuBERT [4]. Данные модели не были исследованы в предыдущих работах [1, 2, 2] для корпуса SentiRuEval-2016.

1. Текстовый корпус

Текстовый корпус SentiRuEval-2016 включает твиты и поделен на две категории объектов: банки и операторы связи.

Данные наборы имеют следующие общие данные:

– twitid – id твита;
– date – дата публикации твита;
– text – текст твита.

Остальные поля в наборе данных представляют собой объекты, которые могут упоминаться в твите и для них необходимо найти оценку.

Множества объектов для категории банков: Сбербанк, ВТБ, Газпром, Альфа банк, Банк Москвы, Райффайзен, Уралсиб, Россельхозбанк; а также для операторов связи: Билайн, МТС, Мегафон, Теле2, Ростелеком, Комстар, Ситилинк.

Соотношение объектов и классов на обучающем и тестовом наборе данных для банков и операторов связи приведено в таблице 1.

Таблица 1

Соотношение классов и объектов

Набор данных	Категория	Классы			
		Нейтрально	Негативно	Позитивно	Оценка отсутствует
Обучающий	Банки	7158	2807	760	64411
	Операторы связи	5209	2611	1382	51292

Тестовый	Банки	2316	784	318	23086
	Операторы связи	1062	2611	231	13269

Таким образом, можно заметить, что данный корпус является сильно не сбалансированным.

Наборы данных характеризуются следующими значениями, представленными в таблице 2:

Таблица 2

Статистика по корпусу

Характеристика	Банки		Операторы связи	
	Обучающий	Тестовый	Обучающий	Тестовый
Количество твитов	9 392	3 313	8643	2247
Количество объектов	8	8	7	7
Количество твитов, имеющих более чем одну оценку	857	101	435	193
Количество твитов, имеющих разноклассовую оценку	23	11	131	49

Таким образом, точность оценок будет смещена в пользу примеров с наибольшем количеством оценок по данной тональности. В предыдущих работах [1, 2, 2] при подсчете метрик не учитывается нейтральный класс тональности, и в нашей работе тоже не будем его учитывать при подсчете метрик, но для обучения моделей примеры с данной тональностью будут использоваться.

Для оценки моделей, как и в предыдущих статьях [1, 2, 2] для данного корпуса будут использоваться метрик F-macro и F-micro для получения оценивания результатов работы модели на всем наборе данных, а так же отдельно $F1_{pos}$, $F1_{neg}$ – метрика F1-score для каждого класса в корпусе, исключая нейтральный класс.

2. Предобработка текста

Перед применением методов машинного обучения из текстов твитов удаляются ссылки и знаки пунктуации, и производится токенизация

На основе токенов будет возможно построить представление текстов, которое необходимо для обучения моделей.

В нейросетевой модели RuBERT, являющейся стеком энкодеров [4], есть встроенный механизм токенизации. Токен – слово из словаря нейросетевой модели; если данного слова нет в словаре, оно разбивается на части, которые в нем есть.

Для методов, основанных на словарях оценочной лексики, и методов машинного обучения, для получения токенов была использована библиотека NLTK [6].

В работе использовались три различных подхода:

- методы на основе словарей оценочной лексики;
- методы машинного обучения;
- нейронная сеть RuBERT [4].

3. Подходы к решению

Методы на основе словарей оценочной лексики являются наиболее простыми и менее ресурсозатратными, чем методы машинного обучения и RuBERT. Было рассмотрено два метода классификации.

Производился поиск слов из твита в словаре оценочной лексики. На основе найденных слов составлялась оценка и использовалась в качестве предсказанной оценки для всех объектов в данном примере. Также данный метод был модифицирован – для каждого из объектов были построены словари синонимов, производился поиск синонима, и поиск слов из словаря оценочной лексики производился в области 10 слов вокруг синонима.

Следующим подходом к решению задачи стали методы машинного обучения для решения задач классификации из библиотеки scikit-learn [**Ошибка! Источник ссылки не найден.**]. Для методов машинного обучения необходимо получить векторное представление документов корпуса. Также для выбранных методов машинного обучения, будут подбираться гиперпараметры моделей, с проверкой результатов обучения при помощи кросс-валидации.

Кросс-валидация – метод оценивания результатов обучения различных моделей машинного обучения. Данный процесс позволяет оценить работу модели на всем наборе данных и при этом усреднить результаты в зависимости от входных данных в процессе обучения модели.

Еще одним подходом к решению задачи является использование нейронной сети RuBERT [4]. Основным назначением данной нейронной сети является получение наилучших представлений входных данных. Так как данная нейронная сеть формирует новые представления в

зависимости от входных данных, то полученные представления можно будет использовать для обычных методов машинного обучения.

4. Представление текста

Для метода, основанного на словарях оценочной лексики, достаточно произвести токенизацию, но для моделей машинного обучения еще необходимо получить представление текста – перевести документы в векторное пространство, с сохранением особенностей как всего корпуса, так и отдельно взятого документа.

В данной работе для моделей машинного обучения было использовано tf-idf. tf-idf – это статистическая мера, используемая для оценки важности в контексте документа, который является частью коллекции документов или корпуса [10].

tf (term frequency) – отношение вхождения некоторого слова к общему числу слов документа:

$$tf(t,d) = \frac{n_t}{\sum_k n_k}, \quad (1)$$

где idf (inverse term frequency) – инверсия частоты, с которой некоторое слово встречается в документах коллекции:

$$idf(t,D) = log \frac{|D|}{|\{d_i \hat{I} D \mid t \hat{I} d_i\}|}, \quad (2)$$

где итоговый вес термина t для документа d рассчитывается по формуле:

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D). \quad (3)$$

Вес tf-idf позволяет оценить важность слова в контексте документа, являющегося частью корпуса.

Набором слов в данном случае будут нормальные формы слов, полученные на основе токенов при помощи библиотеки pymorphy [5].

Для методов машинного обучения было построено векторное представление tf-idf на основе словаря из нормальных форм, который подвергся удалению, таких частей речи как: числительные, союзы, предлоги, частицы, междометия.

Для RuBERT [4] получать представления отдельно не нужно – в процессе обучения данная нейронная сеть обучается получать представление текстов и использует их для решения поставленной задачи. После обучения модели можно отдельно от результатов решения использовать новые представления данных, полученные в результате

работы нейросетевой модели RuBERT, для методов машинного обучения.

5. Результаты экспериментов

В работе было использовано несколько словарей оценочной лексики, которые были построены, основываясь на других словарях, аналогично работе [9]. Происходило голосование словарей – если *n* и более словарей голосовали за то, что данное слово носит негативную оценку, то оно учитывалось как негативное, аналогично для позитивно-окрашенных слов. Были построены словари, где *n* = 1, 2, 3. В таблицах 3 и 4 представлены результаты, полученные с использованием словарей оценочной лексики. Можно заметить, что для словарей, где *n* = 1 и 2 результаты идентичны.

Таблица 3

Результаты на словарях оценочной лексики для банков

Метод	n словарей	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
На всем тексте примера	1	0.3917	0.2835	0.3376	0.3394
	2	0.3917	0.2835	0.3376	0.3394
	3	**0.3919**	**0.3368**	**0.3644**	**0.3695**
На основе синонимов	1	0.3139	0.2826	0.2983	0.2995
	2	0.3139	0.2826	0.2983	0.2995
	3	0.2611	0.2921	0.2766	0.2739

Таблица 4

Результаты на словарях оценочной лексики для операторов связи

Метод	n словарей	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
На всем тексте примера	1	0.5305	0.2253	0.3779	0.4001
	2	**0.5305**	0.2253	0.3779	**0.4001**
	3	0.4927	0.2236	0.3581	0.3883
На основе синонимов	1	0.4764	0.3302	0.4033	0.3954
	2	0.4764	**0.3302**	**0.4033**	0.3954
	3	0.4214	0.3289	0.3751	0.3746

Были протестированы следующие модели машинного обучения [**Ошибка! Источник ссылки не найден.**]:

- логистическая регрессия;
- метод опорных векторов;
- k-ближайших соседей;
- дерево решений;
- случайный лес;
- адаптивный бустинг (AdaBoost);
- градиентный бустинг;
- наивный байесовский классификатор;
- полиноминальный байесовский классификатор.

Лучшими среди протестированных моделей оказались:

- логистическая регрессия;
- метод опорных векторов;
- случайный лес.

Для данных моделей был произведен выбор оптимальных значений гиперпараметров с использованием 5-кратной кросс-валидации на обучающих данных.

Таблица 5

Результаты лучших моделей машинного обучения для банков

Модель	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
SVC (C= 10, kernel='rbf')	0.4700	**0.3333**	0.4016	0.4348
LogisticRegression (solver= 'saga', penalty='none')	**0.5065**	0.3228	**0.4146**	**0.4512**
RandomForestClassifier (criterion= 'gini', n_estimators= 100, max_features= 'log2')	0.2820	0.2171	0.2496	0.2625

Таблица 6

Результаты лучших моделей машинного обучения для операторов

Модель	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
SVC(C= 10, kernel='rbf')	**0.6084**	**0.2544**	**0.4314**	**0.5568**
LogisticRegression(solver= 'newton-cg', penalty='l2')	0.5884	0.2284	0.4084	0.5368

RandomForestClassifier(criterion= 'gini', n_estimators= 150, max_features= 'log2')	0.4910	0.1500	0.3205	0.4431

Для нейронной сети RuBERT производился выбор оптимального количества эпох обучения. Лучший результат был достигнут на обучающем наборе данных как для банков, так и для операторов связи при количестве эпох обучения, равном 2.

Так как основной задачей данной модели является получение эффективных векторных представлений для данных, то полученные векторных представления были использованы в качестве входных данных для лучших методов машинного обучения с подбором гиперпараметров.

Таблица 7

Результаты для баков на векторных представлениях, полученных RuBERT

Модель	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
SVC(C= 0.1, kernel='poly')	0.6833	**0.6252**	**0.6542**	**0.6670**
LogisticRegression(solver= 'saga', penalty='l1')	0.6775	0.6134	0.6455	0.6590
RandomForestClassifier(criterion= 'gini', n_estimators= 50, max_features= 'log2')	**0.6838**	0.6160	0.6499	0.6642
RuBERT	0.6756	0.6245	0.6500	0.6606

Таблица 8

Результаты для операторов связи на векторных представлениях, полученных RuBERT

Модель	$F1_{neg}$	$F1_{pos}$	Macro F1	Micro F1
SVC (C= 0.1, kernel='rbf')	0.7697	0.5324	0.6510	0.7327
LogisticRegression(solver= 'newton-cg', penalty='none')	**0.7707**	0.5246	0.6476	0.7314

RandomForestClassifier(criterion= 'gini', n_estimators= 150, max_features= 'log2')	0.7675	0.5182	0.6428	0.7289
RuBERT	0.7696	**0.5368**	**0.6532**	**0.7330**

На векторных преставлениях из RuBERT оценки для методов машинного обучения улучшились. Это свидетельствует о том, что RuBERT позволяет получить более информативные векторные представления чем tf-idf. Полученный результат может говорить о том, что при помощи RuBERT несбалансированность классов в корпусе может нивелироваться, чего не могут предложить модели машинного обучения с использованием tf–idf, которые имеют сильную зависимость от распределения классов в наборе данных (см. таблицы 5–8).

В таблице 9 представлены лучшие результаты предыдущих работ в сравнении с наибольшими оценками, полученными в процессе исследования.

Таблица 9

Сравнительная характеристика лучших результатов

	Операторы		Банки	
Лучшие результаты статей	Macro F1	Micro F1	Macro F1	Micro F1
[1]	0.5594	0.6813	0.5517	0.5881
[2]	0.5594	0.6525	0.5517	0.5881
[3]	0.5270	0.6840	0.5380	0.5770
SVC + RuBERT	**0.6510**	**0.7327**	**0.6542**	**0.6670**

Таким образом, использование нейросетевой модели RuBERT для получения векторных представлений тестов и SVC в качестве классификатора позволило улучшить результаты предыдущих статей для корпуса SentiRuEval-2016.

Заключение

В ходе работы были рассмотрены различные модели и методы для объектно-ориентированного анализа тональности. Для корпуса SentiRuEval-2016 были получены значения метрик качества, превышающие результаты предыдущих работ [1, 2, 2].

Было выявлено, что при помощи нейросетевой модели RuBERT можно получить такие векторные представления текстов, которые

позволяют сгладить несбалансированность классов изначального корпуса, что дает возможность улучшить результаты, получаемые при помощи стандартных методов машинного обучения.

Список литературы

1. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis / Loukachevitch N. [and all] // Proceedings of International Conference Dialog. – Moscow, June 1-4, 2016. – Режим доступа: https://www.dialog21.ru/media/3410/loukachevitchnvrubtsovayv.pdf
2. Comparison of neural network architectures for sentiment analysis of russian tweets / Arkhipenko K. [and all] // Proceedings of International Conference Dialog. – Moscow, June 1-4, 2016. – Режим доступа: https://www.dialog-21.ru/media/3380/arkhipenkoetal.pdf
3. Entity Based Sentiment Analysis Using Syntax Patterns and Convolutional Neural Network / Karpov I. [and all] // Proceedings of International Conference Dialog. – Moscow, June 1-4, 2016. – Режим доступа: https://www.dialog-21.ru/media/3396/karpoviaetal.pdf
4. Adaptation of deep bidirectional multilingual transformers for russian language / Kuratov Y. [and all] // Proceedings of International Conference Dialog. – Moscow, May 29 – June 1, 2016. – Режим доступа: https://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf
5. Morphological Analyzer and Generator for Russian and Ukrainian Languages / Korobov M. // Analysis of Images, Social Networks and Texts. AIST 2015, 2015.
6. Natural Language Toolkit [Электронный ресурс]: библиотека для python. – Режим доступа: https://www.nltk.org/index.html
7. scikit-learn: machine learning in Python [Электронный ресурс]: библиотека для python. – Режим доступа: https://www.nltk.org/index.html
8. A Survey of Multimodal Sentiment Analysis / M. Soleymani [and all] // Image and Vision Computing, 2017.
9. Lexicon-based Methods vs. BERT for Text Sentiment Analysis / Kotelnikova A. [and all] // 10th International Conference on Analysis of Images, Social Networks and Texts (AIST-2021), 2021.
10. An information-theoretic perspective of tf–idf measures / Aizawa A. // Information Processing & Management, 2003. – P. 45-65.
11. Диалог 2016 [Электронный ресурс]: сайт конференции Диалог 2016. – Режим доступа: https://www.dialog-21.ru/dialogue2016/results/