Lexicon-based methods and BERT model for sentiment analysis of Russian text corpora

Anastasiya Kotelnikova, Danil Paschenko and Elena Razova

Vyatka State University, 36, Moskovskaya st., Kirov, 610000, Russian Federation
kotelnikova.av@gmail.com, depashchenko@mail.ru,
razova.ev@gmail.com

Abstract. The article discusses two approaches to solving the problem of sentiment analysis: lexicon-based approach and deep learning. Within the first approach two well-known lexicon-based methods has been adapted for the Russian language – SO-CAL and SentiStrength. For these methods a unified sentiment lexicon has been prepared using a voting procedure based on the existing lexicons. The second approach has used the RuBERT deep learning model. The SentiRuEval-2015 corpora, which provides reviews and tweets, has been used as training and test data. Analysis of the results showed that deep learning model demonstrates higher accuracy compared to lexicon-based methods.

Keywords: Sentiment analysis, deep learning, BERT, RuBERT, sentiment lexicons, SO-CAL, SentiStrength.

1 Introduction

Sentiment analysis is a field of computational linguistics aimed at automated research of people's opinions and assessments in relation to various objects mentioned in the text, for example, products, services, organizations, persons, events [1]. Sentiment is represented as a value on a certain scale: binary (positive / negative attitude), ternary (adding neutral or contradictory), n-ary or real (for example, [–5, 5]).

There are many studies in the field of sentiment analysis, mainly for the English language [2–5], but in the last decade works for the Russian language have also appeared [6].

There are three main approaches to the sentiment analysis in texts – lexicon-based, machine learning, and hybrid, in which the two indicated approaches are combined [2–3].

Examples of existing lexicon-based sentiment analysis methods are SO-CAL [7] and SentiStrength [4]. Both methods use sentiment lexicons and assess the strength of positive and negative sentiments in texts. The methods take a text as input and produce a numerical value that averages the sentiment of the words in text found from

the sentiment lexicon. They were originally designed for only English texts. In general, a lexicon-based approach requires high-quality sentiment lexicon, the analysis process is quite fast, it doesn't need training data, but the accuracy is often not high enough.

The second approach to sentiment analysis is machine learning, within which there are two areas: traditional machine learning (for example, such methods as SVM, Naïve Bayes, and Gradient Boosting) and deep learning (for example, models based on the Transformer architecture such as BERT) [8–9]. The best results have recently been obtained based on deep learning models [10]. However, such models, having high accuracy, are poorly interpretable compared to lexicon-based methods [11], their application requires high-quality labeled training data, while a significant amount of time is spent on the training procedure. In addition, deep models do not take into account the knowledge about sentiment words contained in the corresponding lexicons.

The purpose of this work is to compare the performance of the lexicon-based methods SO-CAL and SentiStrength adapted for the Russian language and the BERT deep learning model for sentiment analysis, as well as to explore the possibility of adding information from sentiment lexicons to deep learning models.

2 Materials and methods

2.1 Sentiment analysis methods

SO-CAL[†] (Semantic Orientation CALculator) is a method developed by Maite Taboada that determines the sentiment of texts [7]. SO-CAL works for English and Spanish. We have adapted it to the Russian language.

The first change affected preprocessing – for Russian texts it is required to determine not only the part of speech of tokens, as in the original version, but also their initial form. The *rnnmorph*[‡] module was used for this.

Secondly, we used a lexicon created by combining existing sentiment lexicons (described below). With the help of *rnnmorph* the combined lexicon was split into four separate lexicons for different parts of speech: nouns, adjectives, verbs, adverbs. If an element was attributed to several of the specified parts of speech, it was used in several lexicons.

Thirdly, Russian-language lists of special words and expressions, used in the algorithm, were formed, in particular, a lexicon of modifiers (words that affect the sentiment of the words to which they belong, for example, *less*, *absolutely*) and lists of negations (for example, *nothing*, *without*).

As in the original version of the method, the Russian version ignores the sentiment words in the sentence if there is a condition word from the special list in the sentence. Such list includes conditional markers (for example, *if*), some verbs (like *expect*, *doubt*), questions and words enclosed in quotation marks (which may be factual, but

[†] https://github.com/sfu-discourse-lab/SO-CAL.
[‡] https://github.com/IlyaGusev/rnnmorph.

do not necessarily reflect the opinion of the author). If repeating the same sentiment word, there is a decrease in weight for each subsequent repetition.

The version of SO-CAL adapted for the Russian language, just like the original one, for each text produces a numerical value corresponding to the degree of sentiment of the text: a value greater than zero indicates a positive sentiment, a value less than zero indicates a negative sentiment.

On the training part of the corpora a threshold is determined that separates the texts into positive and negative ones (for a three-class classification two thresholds are selected to divide into positive, neutral and negative ones). Next, the sentiment of the texts of the test part is determined, taking into account the found thresholds.

SentiStrength is a lexicon-based method developed by Mike Thelwall et al. [4]. For texts in English it gives two numerical values: the first score is from –1 for texts that are not negative, to –5 for extremely negative texts, the second one is from 1 for texts that are not positive, and up to 5 for extremely positive texts.

The method was originally developed for the English language, we adapted it for the Russian language by changing the linguistic resources required for the algorithm. These are a list of sentiment words, a list of modifiers – words that raise or lower the sentiment score of the following words (for example, *bad*, *a little*, *very*, *extremely*), a list of negations (for example, *not*, *never*), a list of words that indicate the presence of a question in a sentence (for example, *how*, *when*, *why*), etc. The replacement of such language-independent resources as the list of emoticons was not carried out.

Experiments with SentiStrength were performed with two versions of the datasets. The first version is raw, unprocessed data. The second version is preprocessed (lemmatized) data.

RuBERT. In addition to the SO-CAL and SentiStrength methods, we used a model based on the Russian-language version of BERT – RuBERT [9]. The multilingual version of the BERT-base is used as an initialization for RuBERT. The model was trained on the Russian part of Wikipedia and news articles. In our experiments the models were fine-tuned based on training data and tested on test data.

In experiments with the RuBERT model two variants of text corpora were used: a corpus without preprocessing and a corpus in which positive words present in the combined lexicon were replaced by *good*, and negative words by *bad*. This preprocessing procedure made it possible to test the hypothesis of a potential improvement in the performance of sentiment analysis based on RuBERT when adding information from the sentiment lexicon.

2.2 Linguistic resources

The main linguistic resources for solving the problem of sentiment analyses are sentiment lexicons and text corpora labelled by sentiment.

Sentiment lexicons. The combined sentiment lexicon was formed using nine publicly available lexicons for the Russian language [6; 12].

1. *EmoLex* [13]. Created with crowdsourcing. Words in the lexicon are associated with positive and negative sentiments and with emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. This lexicon has been translated into more than 100 languages (including Russian) using Google Translate.
2. *Chen-Skiena's lexicon* [14]. Automatically built for 136 languages, including Russian, using graph propagation.
3. *LinisCrowd* [15]. Created with crowdsourcing. An initially selected list of 7,546 words based on a list of high-frequency adjectives, the lexicon ProductSentiRus [16], an explanatory dictionary and a translation of the English-language sentiment lexicon was labelled by annotators. We considered positive and negative words that received the majority of labels of the corresponding sentiment; contradictory and neutral words were ignored.
4. *RuSentiLex* [17]. For each word the sentiment (positive, negative, neutral) and the source (opinion, fact, feeling) are indicated. To create this lexicon, first, lists of sentiment words were generated based on the RuThes thesaurus, existing sentiment lexicons, news articles and Twitter, then linguists analyzed the resulting lists to form a final lexicon. We use the 2017 version and only words and combinations with positive or negative sentiment.
5. *SentiRusColl* [18]. Russian sentiment lexicon of collocations. To create a lexicon a corpus of reviews for ten domains was used, combinations of candidate words were automatically selected from it, which were then labelled by three annotators. The lexicon contains the collocations that received the majority of votes.
6. *Word Map* [19]. Online thesaurus of words and expressions of the Russian language. The sentiment lexicon developed within this project contains the Russian words, supplied with label and strength of sentiment (positive, negative or neutral). Crowdsourcing was applied to create the lexicon. We used positive and negative words from the 2019 version of this lexicon.
7. *Blinov's lexicon* [20]. A manually compiled list of 969 most positive and 1,138 most negative words from the lexicon ProductSentiRus was automatically expanded with synonyms and antonyms from the Russian Wiktionary.
8. *Kotelnikov's lexicon* [21]. An automatically selected list of words from five domains was labelled by four annotators. We took words, on the sentiment of which at least three out of four annotators agreed.
9. *Tutubalina's lexicon* [22]. A manually created lexicon based on strictly positive and negative user reviews about cars has been expanded with synonyms.

Each lexicon has been separately processed as follows:

- neutral words have been removed;
- all words have been converted to lower case;
- words that are both positive and negative in the lexicon have been removed (including the analyzed words with the spelling "e" and "ë");
- words containing non-Cyrillic letters in the spelling have been removed;
- one occurrence of each element has been left.

The size of lexicon after preprocessing is given in Table 1.

Table 1. Size of preprocessed lexicons.

Lexicon	Size	Positive elements	Negative elements
Blinov's lexicon	3,665	1,692	1,973
EmoLex	4,685	2,020	2,665
Chen-Skiena's lexicon	2,876	1,246	1,630
LinisCrowd	3,992	1,128	2,864
RuSentiLex	11,941	3,093	8,848
SentiRusColl	6,577	4,008	2,569
Word Map	11,324	4,550	6,774
Kotelnikov's lexicon	3,238	1,036	2,202
Tutubalina's lexicon	2,482	1,051	1,431

At the next stage a combined sentiment lexicon was obtained from the preprocessed lexicons. This lexicon includes words that are simultaneously found in at least four source lexicons: 1,444 negative words (67%) and 712 (33%) positive words, a total of 2,156 words. At the same time only those words were left in which the sentiment was clearly defined, i.e. there were no controversial cases. A controversial case was considered when the number of lexicons in which a word was classified as negative coincided with the number of lexicons in which it was positive. For the rest of the cases the words were assigned to the prevailing class, i.e. the voting was used. Other values were investigated for the minimum number of lexicons, but for four the best performance scores were obtained on the training data.

We created two versions of the combined lexicon – in the first one positive elements were assigned a sentiment score of 3, negative ones – a score of –3. This version of the lexicon is hereinafter referred to as *CLex*. In the second version of the lexicon the sentiment score was determined by the number of source lexicons that contained a given word. The second version of the lexicon is weighted and is further called *WCLex*.

Text Corpora. For experimental research the corpora of the SentiRuEval-2015[§] competition were taken. The data consists of labelled reviews of restaurants and cars, as well as tweets about banks and telecommunications companies. Corpora sizes are very different: there are fewer reviews than tweets, but the average review length is almost 10 times longer than the length of a tweet (830 characters versus 85 for training data and 830 versus 82 for test data). This is due to the maximum possible tweet length. The organizers of the competition provided training and test data. The training sample consists of 403 reviews and 9,722 tweets. The size of the test sample is 403 reviews and 8,308 tweets (Table 2).

[§] http://www.dialog-21.ru/evaluation/2015/sentiment/.

Table 2. Corpora size.

		Reviews		Tweets	
Category	Sentiment	Cars	Restaurants	Banks	Telecom
Training	Negative	30	28	1,059	1,585
	Neutral	58	36	3,470	2,346
	Positive	115	136	354	908
	Total	**203**	**200**	**4,883**	**4,839**
Test	Negative	26	26	654	847
	Neutral	76	31	3,534	2,585
	Positive	98	146	346	342
	Total	**200**	**203**	**4,534**	**3,774**

3 Results

Three models were tested – the lexicon-based methods SentiStrength (SS) and SO-CAL adapted for the Russian language, as well as the RuBERT deep learning model.

For lexicon-based methods, two variants of the combined lexicon were used: CLex and WCLex. For SO-CAL each corpus preprocessed with *rnnmorph* was evaluated on both lexicons. For SentiStrength estimates were obtained on the original raw data and on the preprocessed lemmatized data. For the RuBERT model two variants of text corpora were used: corpora without preprocessing and corpora with replaces to *good* and *bad* of words from sentiment lexicon.

Two series of experiments were carried out. In the first series only texts with a positive and negative sentiment were used, thus a binary classification was carried out. The second series of experiments was carried out for a three-class classification – texts with a neutral sentiment were also used.

Table 3 shows the values of the macro F1-score metric on test data for binary classification.

Table 3. Binary classification results: macro F1-score.

	Reviews		Tweets	
Model	Cars	Restaurants	Banks	Telecom
SO-CAL, CLex	0.8247	0.8453	0.6099	0.6868
SO-CAL, WCLex	0.8666	0.8398	0.5747	0.6987
SS, no preprocessing, CLex	0.8817	0.8092	0.2521	0.3218
SS, preprocessed, CLex	0.9072	0.8470	0.2521	0.3232
SS, no preprocessing, WCLex	0.8853	0.9236	0.5305	0.5281
SS, preprocessed, WCLex	0.8817	0.8467	0.5256	0.5252
RuBERT, no preprocessing	0.9495	**0.9701**	**0.8579**	**0.8198**
RuBERT, with replaces	**0.9700**	0.9664	0.8441	0.8058

Among the lexicon-based methods, on average, for all experiments, SO-CAL showed better results by 10 percentage points (pp), while in the classification of reviews, the results were better for SentiStrength, tweets – for SO-CAL. The use of a weighted lexicon for both methods sometimes even worsens, but on average not significantly improves the performance (by about 9 pp on average, and by 28 pp maximum for tweets about banks when using SentiStrength). In general, among the lexicon-based methods the best result was shown by SO-CAL when using WCLex.

The RuBERT model is superior to lexicon-based methods. On average for all experiments its result exceeds the results of lexicon-based methods by 22 pp (in comparison with the best lexicon-based method – by 15 pp). The experiment carried out to replace words with subsequent fine-tuning of the RuBERT model showed a good result only in one case – for reviews on cars, the quality increased by 2 pp, in other cases such preprocessing deteriorated the performance.

The performance of the classification of reviews is on average 32 pp higher than the performance of tweet classification, the largest difference (by 59 pp) for SentiStrength on lemmatized data with CLex.

Table 4 shows the values of the macro F1-score metric on test data for a three-class classification.

Table 4. Three-class classification results: macro F1-score.

	Reviews		Tweets	
Модель	Cars	Restaurants	Banks	Telecom
SO-CAL, CLex	0.6118	0.6412	0.5091	0.5005
SO-CAL, WCLex	0.6354	0.6247	0.5516	0.4884
SS, no preprocessing, CLex	0.4992	0.4682	0.4024	0.4026
SS, preprocessed, CLex	0.4400	0.4781	0.4022	0.4040
SS, no preprocessing, WCLex	0.5503	0.4601	0.5286	0.4811
SS, preprocessed, WCLex	0.5238	0.4199	0.5308	0.4786
RuBERT, no preprocessing	**0.6806**	**0.6837**	**0.7043**	**0.6434**
RuBERT, with replaces	0.6134	0.5991	0.7010	0.6331

With a three-class classification among the lexicon-based methods on average for all experiments again by 10 pp SO-CAL performed better, with better results for both reviews and tweets. The use of a weighted lexicon for both methods has very little effect on the performance. In general, among the lexicon-based methods the best result was shown by SO-CAL when using WCLex.

The best results for the three-class, as well as for the binary classification, are shown by the RuBERT model. On average for all experiments its results exceeds the results of lexicon-based methods by 16 pp (in comparison with the best lexicon-based method – by 10 pp). Replacing words in the original data did not improve the result. The performance of the classification of reviews and tweets on average differs slightly (by 3.5 pp).

4 Discussion

The experiments showed that the RuBERT deep learning model in all cases gets better results than lexicon-based methods (compared to the best SO-CAL lexicon-based model with a weighted lexicon – by 15 pp for a binary and 10 pp for a three-class classification). However, the results of the neural network model are difficult to interpret, and the lexicon-based methods provide detailed information about the sentiment words and expressions found in the text. This allows, in particular, analyzing the errors that these methods made.

The analysis showed that for reviews the most common causes of errors are, for example, the following: incorrect search for negation, search for not all sentiment words, prevalence of vocabulary of the opposite sentiment, a problem of the third class, when sentiment 0 is used not only for neutral texts, but for contradictory ones also. There are also two other common types of errors for tweets: lack of knowledge of the context and incomplete phrases, which are related to the specifics and limitation of the number of characters in one text message.

5 Conclusion

The use of the RuBERT deep learning model has a higher performance (by 22 pp for two-class classification and by 16 pp for three-class classification) compared to the adapted versions of the SO-CAL and SentiStrength lexicon-based methods, but its results are difficult to interpret. The ability to analyze errors allows you to identify ways of improving lexicon-based methods.

Adding information from the lexicon during preprocessing data for RuBERT led to an improvement in the result only in one case out of eight.

References

1. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press (2015).
2. Poria, S., Hazarika, D., Majumder, N., Mihalcea, R.: Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. In: Computing Research Repository, arXiv: 2005.00357 (2020).
3. Taboada, M.: Sentiment Analysis: An Overview from Linguistics. In: Annual Review of Linguistics, 2, 325–347 (2016).
4. Thelwall, M. et al: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558 (2010).
5. Mohammad, S.M.: Emotion Measurement (Second Edition), Chapter 11 – Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text, edited by H.L. Meiselman, Woodhead Publishing, 323–379 (2021).
6. Kotelnikov, E.V. et al: Modern sentiment lexicons for opinion mining in English and Russian (analytical survey). Nauchno-tekhnicheskaya informaciya, 12, 16–33 (2020).
7. Taboada, M. et al: Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267–307 (2011).

8. Devlin, J. et al: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language (2018).
9. Kuratov, Y., Arkhipov, M.: Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. MIPT, Dialogue 2019 (2019).
10. Howard, J., Ruder, S.: Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia. Association for Computational Linguistics, 328–339 (2018).
11. Yu, R., Alì, G.: What's Inside the Black Box? AI Challenges for Lawyers and Researchers. Legal Information Management, 19(1), 2–13 (2019).
12. Kotelnikov, E.V. et al: A comparative study of publicly available Russian sentiment lexicons. – Communications in Computer and Information Science: 7th conference on Artificial Intelligence and Natural Language (AINL-2018), 139–151. Springer (2018).
13. Mohammad, S.M., Turney, D.P.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436–465 (2013).
14. Chen, Y., Skiena, S.: Building Sentiment Lexicons for All Major Languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 383–389, Baltimore (2014).
15. Koltsova, O.Yu., Alexeeva, S.V., Kolcov, S.N.: An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2016", 15 (22), 277–287 (2016).
16. Chetviorkin, I., Loukachevitch, N.: Extraction of Russian Sentiment Lexicon for Product Meta-Domain. In: Proceedings of COLING 2012, 593–610, Mumbai (2012).
17. Loukachevitch, N., Levchik, A.: Creating a General Russian Sentiment Lexicon. In: Proceedings of Language Resources and Evaluation Conference LREC-2016, 1171–1176 (2016).
18. Kotelnikova, A., Kotelnikov, E.: SentiRusColl: Russian Collocation Lexicon for Sentiment Analysis. In: Artificial Intelligence and Natural Language. AINL 2019. Communications in Computer and Information Science, 1119, 18–32, Springer, Cham (2019).
19. Kulagin, D.: Russian Word Sentiment Polarity Dictionary: a Publicly Available Dataset. Poster in: Artificial Intelligence and Natural Language. AINL 2019 (2019).
20. Blinov, P.D. et al: Research of lexical approach and machine learning methods for sentiment analysis. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2013", 12(19). 51–61 (2013).
21. Kotelnikov, E. et al: Manually Created Sentiment Lexicons: Research and Development. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2016", 15 (22), 300–314 (2016).
22. Tutubalina, E.V.: Extraction and summarization methods for critical user reviews of a product. Kazan Federal University, Kazan, Russia (2016).