

Combining Lexicons and Neural Networks for Targeted Sentiment Analysis

Anton Golubev
Lomonosov Moscow State University
Moscow, Russia
antongolubev5@yandex.ru

Natalia Loukachevitch
Lomonosov Moscow State University
Moscow, Russia
louk_nat@mail.ru

Abstract—In this paper we considered the targeted sentiment analysis problem for the Russian language. We compared several machine learning approaches based on transformer architecture: BERT, RuBERT, RuRoberta. The main focus of our studies was directed to the use of additional sources of knowledge stored in sentiment lexicons. Two Russian sentiment lexicons: RuSentiLex and RuSentiFrames were involved in our study. We use the lexicons in two different ways. The first technique uses the RuSentiLex for automatic labeling of additional text data to enhance training dataset. Two variants of the lexicon-based annotated texts were obtained: general and domain-specific collections. Another technique creates embeddings for word labels in sentiment lexicons and, in such a way, integrate the lexicon knowledge into the whole process of neural network training. The combination of two Russian vocabularies and the GRU-TSC architecture achieved state-of-the-art results on five evaluation datasets.

Keywords—*target-oriented sentiment analysis, neural networks, transfer learning, sentiment lexicons, distant supervision*

I. INTRODUCTION

Sentiment analysis or opinion mining is an important natural language processing task used to determine sentiment attitudes in the text. Nowadays many people express their opinions on various topics in the Internet and social networks. Automatic extraction and analysis of this data is important for government organizations, companies, as well as ordinary users. One of its main applications is researching product reviews and monitoring customer feedback.

Currently, improving the quality of automatic sentiment analysis is associated with the use of modern transformer architectures, for example, BERT [1]. In recent years, transfer learning techniques have gained wide popularity. This approach includes a preliminary stage, which consists of identifying general ideas for the initial task, and an adaptation stage, which consists of applying the information obtained in the previous step to the target task [2].

Another way to improve sentiment classification quality is to use external knowledge sources such as sentiment vocabularies. In this work, we use an approach that

combines transformer architecture with the usage of Russian sentiment lexicons, which include words marked by sentiment. The results were improved by more than 4% compared to current performance levels for all evaluation datasets.

II. RELATED WORK

There exist various approaches of combining lexicons and deep learning approaches to increase the quality of sentiment analysis systems. In [3], the authors create specialized lexicon-based features and add them to other features in sentiment classification of tweets. With this aim, they gather three English manual and two automatically generated sentiment vocabularies. The authors used the sentiment score $score(w, p)$ for each token w and emotion or polarity p to determine: total count of tokens in the tweet with $score(w, p) > 0$, the sum of such scores, the greatest score and the score of last token in the text. SVM-classifier trained on these expanded lexicon-based features improved the performance on more than 8 percent points over the same approach without the lexicon-based features. This approach gained the best results on sentiment analysis task on Twitter data collection during SemEval-2013 [4]. The approach made it possible to mitigate the problem of previously unknown sentiment words that appeared in the test collection but absent in the training collection.

In current BERT-based approaches for sentiment analysis, embeddings for sentiment categories described in lexicons are trained [5, 6]. The authors of [7] tune pre-trained language models (PLMs) using task-specific prompts. The main idea of this approach called prompt-tuning is to add text sentences to the input and convert a classification problem into a masked language modeling problem, where a crucial step is to predict a masked word and transform the predicted words to the task label. Hu et al., 2022 [7] use external knowledge bases (vocabularies), which allows combining language models and lexicons.

Current state-of-the-art results on most available Russian datasets for sentiment analysis are achieved using transfer learning approaches based on the BERT model, more specifically on Russian variant of BERT (RuBERT) [8]. In [2], the authors compared several different configurations of RuBERT model and found Conversational RuBERT pre-trained on posts and comments from Russian segment of social networks to be the best model for sentiment analysis problem.

In [9], the authors suggested an approach for automatic creation of a labeled collection from a news corpus using the distant supervision technique and compared such technique to existing augmentation techniques. The authors tested different options of combining the additional data with several Russian sentiment analysis benchmarks and improved an existing state-of-the-art performance achieved by a vanilla BERT by more than 3% using BERT models in combination with the transfer learning approach.

In RuSentNe evaluation¹ dedicated to targeted sentiment analysis in Russian news texts, participants with top results applied ensembles of several BERT-like models such as RuRoBERTa, XLM-RoBERTa, RemBERT together with weighting or sampling techniques to overcome the prevalence of neutral class [10, 11].

III. METHODS

We are exploring different methods to optimize BERT architecture for targeted sentiment analysis problem. We categorize these methods into two approaches: the single-sentence approach, where only the original text is used as input, and the sentence-pair approach, where an additional sentence is added to the original text to transform the sentiment analysis task into a classification problem [13].

A. Sentence-single model

The sentence-single model is a basic BERT model with only one linear layer added on top. To enable classification, a special token *[CLS]* is inserted at the beginning of the input text. The sentence-pair architecture, on the other hand, includes an additional sentence in the input by separating them with a *[SEP]* token. The key distinction between the two models lies in the placement of the linear layer: in the sentence-pair model, it is added on top of the final hidden state of the *[CLS]* token, while in the sentence-single configuration it is added on top of the entire last layer. In case of targeted sentiment analysis, where there are labels for each object being analyzed, the labels can be substituted with a special token *[MASK]*.

B. Sentence-pair model

The sentence-pair model has two different architectures depending on the type of task: natural language inference (NLI) and question answering (QA) tasks. Each architecture has auxiliary sentences that are used for training process. Here are the examples of the auxiliary sentences for each model:

- pair-NLI: *"The sentiment of MASK is"*
- pair-QA: *"Which sentiment is MASK?"*

We use RuBERT [15] from DeepPavlov framework trained on posts and comments from Russian social networks segment as a pre-trained model.

IV. USE OF ADDITIONAL DATA

We suggest two methods of increasing training datasets since transfer learning approach performs better using more annotated data. This method includes a pre-training stage of learning general concepts of the source task and a stage of

adapting the application of previously acquired knowledge to the target task. In this work, we study augmentation and distance supervision techniques for the sentiment analysis task in Russian [16, 17].

The aim of augmentation is to alter an existing dataset in order to generate more data that resembles the original. It is extensively utilized in computer vision and audio processing, where its efficacy has been supported by numerous studies. However, applying this approach to the field of Natural Language Processing (NLP) is more challenging and has been less extensively researched due to the complexity of languages and tasks. The drawback of using augmentation for sentiment analysis is that it does not introduce new sentiment objects into the sentences, but merely adds additional noise to the original examples. We propose several alternative augmentation techniques.

SYNONYM REPLACEMENT. Synonym replacement consists of replacing n random words with their synonyms. The value of n is calculated based on the length of the sentence l with the following formula: $n = \alpha * l$, where α was experimentally set equal to 0.2.

RANDOM SWAP. Random swap involves exchanging the positions of two randomly selected words within the sentence. This swap is not limited to neighboring words; words that are further apart can also be swapped. In our approach, we define a word spacing threshold of 4, meaning that words within a distance of 4 (inclusive) can be swapped. The first word in the sentence, excluding the sentiment object, is randomly chosen using a uniform distribution. Next, a non-zero value was randomly selected from $[-4, 4]$ and the swap was performed.

RANDOM DELETION. The random deletion method involves removing a certain number of words from a sentence. The specific number of words to be deleted, denoted as n , is determined based on the length of the original sentence, denoted as l . This is done using the formula $n = \alpha * l$, where the value of α is experimentally set to 0.3. Similar to the synonym replacement method, we utilize a list of entities to ensure they are not removed from the augmented sentence during the deletion process.

BACK-TRANSLATION. Back-translation refers to the process of translating a sentence into another language and back into the original language. This method utilizes techniques that enable the paraphrasing of text while preserving its original meaning. Essentially, it maintains the same semantic content of the sentence but produces a different sentence structure. In this specific approach, the French language was chosen for the reverse translation due to its abundance of linguistic patterns, which helps in generating high-quality examples that do not alter the intended meaning.

V. DISTANT SUPERVISION TECHNIQUES

The primary concept behind utilizing the distant supervision approach is to generate an additional annotated sentiment dataset. This process relies on a sentiment dictionary that comprises both positive and negative words and phrases along with their corresponding sentiment scores. For our purposes, we employ the RuSentiLex Russian sentiment dictionary[12], which encompasses commonly used semantic words in the Russian language, slang expressions from Twitter, as well as words with

¹ <https://github.com/dialogue-evaluation/RuSentNE-evaluation>

positive or negative connotations derived from a news corpus. The current version of RuSentiLex consists of 16445 elements.

We utilize a 4 GB Russian news corpus sourced from various channels and covering diverse topics as our foundation for automatic dataset generation. This aspect is crucial as the benchmarks we analyze encompass multiple subject matters. The automatically annotated dataset consists of two sections: general and thematic. To establish the general part, we identify unambiguous positive and negative nouns from the RuSentiLex lexicon that can refer to individuals or companies (e.g. scammer, kidnapper) who serve as the subjects of sentiment in the evaluation datasets. We compile lists of positive and negative words and assume that if a sentence contains a word from either of these lists, it conveys the same sentiment. The roster of positive and negative mentions of individuals or companies (referred to as seed words) comprises 822 negative and 108 positive mentions. Sentences may include multiple seed words with different meanings. In such instances, we duplicate the sentences and assign labels based on their respective attitudes.

To construct the thematic portion of the automatic collection, we search for sentences that mention relevant named entities, such as banks or telecom operators, based on the specific task at hand. This is achieved by employing DeepPavlov's named entity recognition (NER) model and identifying instances where sentiment-related words cooccur with these named entities within the same sentences. Furthermore, to ensure that such sentiment-related word is in reference to the intended object, we establish a constraint where the distance between the two words should not exceed four words. As the test sets also encompass a category of neutral sentiments, it is necessary to extract sentences that do not convey any particular sentiment. To accomplish this, we scrutinize the examples identified by the NER and specifically select those which do not include any sentiment-related words from the lexicon.

A. General And Thematic Parts of Data

The distinction between the general and thematic supplementary samples can be explained as follows. The general dataset comprises sentences that involve commonly used positive or negative nouns, such as "hero" or "liar", which typically refer to individuals or organizations. On the other hand, the thematic cluster of the collection consists of sentences where a named entity from a specific domain appears alongside words that express a positive or negative connotation. It is important to note that only thematic sentences can be neutral.

When creating an additional dataset, we take into consideration the word distribution based on sentiment within the resulting sample and try to make it as balanced as possible. While the original corpus includes enough examples featuring negative sentiment to form a well-rounded collection, the same cannot be said for words expressing positive sentiment.

VI. ENCODING LEXICONS INTO NEURAL NETWORK

The next approach of using external data sources is to embed knowledge directly into the model architecture. For this task, we use a BiGRU model able to interact with

multiple embeddings from a language model and external knowledge sources (GRU-TSC) [5]. The model input contains three different types. The first one is an initial text constructed as a question-answering problem. Specifically, we concatenate the sentence and target mention and tokenize the two segments using the language model tokenizer and vocabulary. This step results in a text input sequence, consisting of n word pieces, where n is the manually defined maximum sequence length.

The second input is a feature representation of the sentence, which obtained using external knowledge sources (EKS), such as sentiment lexicon in our case. Given an EKS with d dimensions, EKS representation $E \in R^{n \times d}$ is constructed, where each vector e_i is a feature representation of the word piece i in the sentence. E matrix is utilized for creation learning associations between the token-based EKS representation and the language model vocabulary-based sequence T , so that it contains k repeated vectors for each token where k is the token's number of word pieces. Therefore, special characters, such as CLS , are considered.

The last input is a target mask $M \in R^n$, i.e., for each word piece i in the sentence that belongs to the target, $m_i = 1$. After passing T into the language model for getting a contextualized word embedding of shape $R^{n \times h}$, where h is the number of hidden states in the language model, E is fed into a randomly initialized matrix $W_E \in R^{d \times h}$ for creation of EKS embedding. The target mask M is repeated to be of shape $R^{n \times h}$. The same shape of all embedding types ensures a balanced influence of each input to the architecture's downstream components. After all preparations embeddings are stacked to form a matrix $TEM \in R^{n \times 3h}$.

All three different types of embeddings interact using a single-layer BiGRU which yields hidden states $H \in R^{n \times 6h} = TEM$. Since LSTM architecture is commonly used to learn a higher-level representation of word embeddings, BiGRU model is used. Three pooling techniques are employed to turn the interacted and sequenced representation H into a single vector. An element-wise mean and maximum over all hidden states H are calculated and the last hidden state h_{n-1} is retrieved. For creation of model output three vectors are stacked to P and fed into a fully connected layer.

As a result, let us summarize the list of approaches for sentiment analysis task compared in this work:

- a vanilla BERT architecture with question answering configuration (BERT-QA),
- a vanilla BERT architecture with natural language inference configuration (BERT-NLI),
- BERT-QA pre-trained on augmented dataset using transfer learning technique,
- BERT-QA pre-trained on automatically annotated dataset using distant supervision,
- GRU-TSC architecture with different variants of using external knowledge sources.

VII. DATASETS

We use five Russian datasets that have been annotated for past evaluations of Russian sentiment. These collections include news quotes from the ROMIP-2013 evaluations, as well as Twitter datasets from two SentiRuEval competitions conducted in 2015 and 2016. The evaluation datasets and

the sizes of their training and test portions are presented in Table I. Table II provides the distribution of the texts within the datasets according to their sentiment classes, where “-” represents the negative class, “0” represents the neutral class, and “+” represents the positive class.

TABLE I. EVALUATION DATASETS

Dataset	Train Volume	Test Volume
News Quotes ROMIP-2013	4260	5500
SentiRuEval-2015 Telecom	5241	4173
SentiRuEval-2015 Banks	6232	4612
SentiRuEval-2016 Telecom	9209	2460
SentiRuEval-2016 Banks	10725	3418

To compile a collection of news quotes, we extracted opinions expressed in direct or indirect speech from news articles. The objective was to classify these quotes as either neutral, positive, or negative comments made by the speaker regarding the topic of the quote. As shown in Table II, the distribution of classes in the dataset is imbalanced. The Twitter datasets were annotated for the purpose of monitoring reputation. The aim of SentiRuEval's sentiment analysis competition was to identify opinions or positive and negative facts related to two types of organizations: banks and telecommunications companies. This problem can be categorized as target-oriented sentiment analysis. Similar evaluations were conducted in both 2015 and 2016 years, resulting in the annotation of four target datasets. In 2016, the training datasets in both domains were created by combining the train and test sets from the 2015 evaluation, resulting in much larger datasets.

From Table II, it is evident that the neutral class dominates in sentiment distribution of datasets. Due to this, the primary quality measure used is the macro measure $F_1^\pm$. The macro measure is calculated by finding the average value between F_1 measure of positive and negative classes. The measure F_1 of the neutral class is disregarded because this category is typically not informative to consider.

VIII. RESULTS

This section includes the description of experiments conducted using the models described in the section III. All comparisons of results are presented only on SentiRuEval-2015 Telecom dataset to save space. During the study, it was revealed that the quality of classification does not depend on the selected data collection.

A. Experiments without Using Additional Data

As a first stage of the research, vanilla BERT architectures were compared in several variations: a model using one sentence as input and two in the natural language inference (NLI) and question answering (QA) settings. Additionally, a comparison was made between the use of several different language models: RuBERT and Conversational RuBERT (C) from DeepPavlov and ruRoberta-large trained by Sber AI team.

TABLE II. SENTIMENT CLASS DISTRIBUTION BY LABELS (%)

Collection	Train Dataset	Test Dataset
------------	---------------	--------------

	+	-	0	+	-	0
News Quotes ROMIP-2013	26	44	30	32	41	27
SentiRuEval-2015 Telecom	19	34	47	10	23	67
SentiRuEval-2015 Banks	7	36	57	8	14	78
SentiRuEval-2016 Telecom	15	28	57	10	47	43
SentiRuEval-2016 Banks	7	26	67	9	23	68

The results of the first stage experiments are presented in Table III. It is worth to note that one participant of the SentiRuEval-2015 uploaded manual annotation of the test telecom dataset and obtained the results described as Manual approach in table of methods. RuRoberta architecture [14] treating sentiment analysis task as a question answering problem showed the best results and was able to get closer to the result of human markings.

TABLE III. RESULTS WITHOUT USING ADDITIONAL DATA

Approach	$F_1^\pm$ macro	F_1 macro
BERT-single	58.43	67.04
BERT-pair-QA	58.15	67.83
BERT-pair-NLI	59.17	68.24
BERT-single (Conversational)	61.34	69.12
BERT-pair-QA (Conversational)	63.47	68.54
BERT-pair-NLI (Conversational)	63.14	68.83
RuRoberta-large-pair-QA	65.07	70.21
RuRoberta-large-pair-NLI	64.93	70.06
Manual approach	70.30	-

B. Experiments Using Transfer Learning

Through the utilization of a transfer learning method, we have incorporated several fine-tuning strategies that involve training in multiple stages with intermittent freezing of the model weights. The options are as following:

- Two-step approach: in the initial stage, the model undergoes independent sequential training on an additional dataset, followed by training on an evaluation train dataset in the second stage.

- Three-step approach: in this case, the model undergoes independent iterative training in three steps using the general and thematic clusters from the additional dataset, as well as the benchmark training sets.

During the experiment, we also investigated how the metrics were affected by the size of the additional dataset. It was observed that the point at which expanding the automatically generated data and improving the results occurred was when the sample size reached 27000 with uniform distribution across the sentiment classes. By utilizing the two-step approach, we were able to surpass level obtained on the previous stage.

TABLE IV. RESULTS USING TRANSFER LEARNING APPROACH

Approach	$F_1^\pm$ macro	F_1 macro
RuRoberta-large-pair-QA (2 step)	65.53	71.03

Approach	$F_1^{\pm}$ macro	F_1 macro
RuRoberta-large-pair-QA (2 step)	65.53	71.03
RuRoberta-large-pair-NLI (2 step)	65.31	70.87
RuRoberta-large-pair-QA (3 step)	66.76	73.44
RuRoberta-large-pair-NLI (3 step)	66.18	73.06
RuRoberta-large-pair-QA (SOTA)	65.07	70.21
Manual	70.30	-

To modify the previous experiment's three-step transfer learning approach, we divided the first step into two phases. In this manner, the models were initially trained on general data. Once this training phase was complete, the weights were frozen, and training continued using thematic examples obtained through a list of organizations and named entity recognitions (NERs) from DeepPavlov. After the second round of weight freezing, the models were trained using evaluation training sets. In this stage, we introduced additional examples of sentiments to the thematic section of the supplementary dataset by selecting thematic sentences that contained relation words. The first step sample included 18000 general examples, while the second sample consisted of 9000 thematic. Both sets were evenly balanced between the various sentiment classes.

The results for the experiments using transfer learning are presented in Table IV. All experiments were produced using the pre-trained RuRoBerta model, since it showed better results in the previous stage. RuRoBerta-pair-QA configuration improved the result by almost 3% F_1 macro and almost approached to manually labelled answers.

C. Experiments Using Encoded Lexicons

During the experiments with the GRU-TSC architecture, two different sentiment lexicons were used. RuSentiLex lexicon [12] is an alphabetically ordered list of words and expressions, containing the following types of Russian words, the meanings of which are related to sentiment connotation:

- words (expressions) of the literary Russian language for which at least one meaning has an evaluative component, which means that the word in this meaning either clearly expresses an attitude towards the object under discussion;
- words (expressions) that do not convey the author's evaluative attitude, but have a positive or negative association (connotation);
- slang and swear words from Twitter (not obscenities).

The RuSentiFrames lexicon [15] consists of descriptive words and expressions collected to so-called sentiment frames conveying several types of inferred information about relations and effects.

During four experiments with RuRoberta-large combined with different combinations of sentiment lexicons, it was found that the results improve linearly with the addition of new external sources of knowledge. The results are presented in Table V. The approach using both dictionaries showed the best results and came close to human markup among all methods.

TABLE V. RESULTS USING EXTERNAL KNOWLEDGE SOURCES

Approach	$F_1^{\pm}$ macro	F_1 macro
GRU-TSC without EKS	64.35	70.11
GRU-TSC with RuSentiLex	65.54	71.23
GRU-TSC with RuSentiFrames	66.07	72.76
GRU-TSC with both lexicons	67.23	74.82
RuRoberta-large-pair-QA (3 step)	66.76	73.44
Manual	70.30	-

CONCLUSION

In this paper we considered the target-oriented sentiment analysis task for the Russian language. We compared several machine learning approaches based on transformer architecture: BERT, RuBERT, RuRoberta. The main focus of our studies was directed to the use of additional sources of knowledge stored in sentiment lexicons. Two Russian sentiment lexicons: RuSentiLex and RuSentiFrames were involved in our study.

We used the lexicons in two different ways. The first technique uses the RuSentiLex for automatic labeling of additional text data to enhance training dataset. Two variants of the lexicon-based annotated texts were obtained: general and domain-specific collections. Another technique creates embeddings for word labels in sentiment lexicons and, in such a way, integrate the lexicon knowledge into the whole process of neural network training. The combination of two Russian vocabularies and the GRU-TSC architecture achieved the state-of-the-art results on five evaluation datasets.

ACKNOWLEDGMENT

This research is partially supported by Russian Science Foundation (project 21-71-30003)

REFERENCES

- [1] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In Proc. NAACL-HLT, pp. 4171, 2019.
- [2] A. Golubev, N. Loukachevitch, "Improving results on Russian sentiment datasets. In Proc. of the Artificial Intelligence and Natural Language," AINL 2020, pp. 109–121, 2020.
- [3] S. Mohammad, S. Kiritchenko, Zhu X, "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets," Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 321–327, 2013.
- [4] T. Wilson, Z. Kozareva, P. Nakov, S. Rosenthal, V. Stoyanov, A. Ritter, "SemEval-2013 task 2: Sentiment analysis in Twitter," In Proc. of the International Workshop on Semantic Evaluation, 2013.
- [5] F. Hamborg, "NewsMTSC: a dataset for (multi-) target-dependent sentiment classification in political news articles," Association for Computational Linguistics (ACL), 2021.
- [6] M. Hosseini, E. Dragut, A. Mukherjee, "Stance Prediction for Contemporary Issues: Data and Experiments," In Proc. of the Eighth International Workshop on Natural Language Processing for Social Media, 2020.
- [7] Hu S., N. Ding, H. Wang, Zh. Liu, J. Wang, J. Li, W. Wu, M. Sun, "Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification," In Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2225–2240, 2022.
- [8] M. Burtsev.: "DeepPavlov: Open-Source Library for Dialogue Systems," In Proc. of ACL2018, System Demonstrations pp. 122–127, 2018.

- [9] A. Golubev, N. Loukachevitch, "Use of augmentation and distant supervision for sentiment analysis in Russian," In Proc. of Text, Speech, and Dialogue: 24th International Conference, TSD 2021, pp. 184–196, 2021.
- [10] P. Podberezko, A. Kaznacheev, S. Abdullaeva, "Half-MAsked Model for Named Entity Sentiment analysis," In Proc. of the International Conference Dialogue, 2023.
- [11] A. Glazkova, "Fine-tuning Text Classification Models for Named Entity Oriented Sentiment Analysis of Russian Texts," In Proc. of the International Conference Dialogue, 2023.
- [12] N. Loukachevitch, A. Levchik, "Creating a general russian sentiment lexicon," In Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1171–1176, 2016.
- [13] C. Sun, I. Huang, X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," In Proc. of North American Chapter of the Association for Computational Linguistics, pp. 380-385, 2018.
- [14] HuggingFace, ruRoBERTa-large, Available: <https://huggingface.co/ai-forever/ruRobertalarge>.
- [15] N. Loukachevitch, N. Rusnachenko, "Sentiment frames for attitude extraction in Russian," In Proc. Of Computational Linguistics and Intellectual Technologies, pp. 541-552, 2020.