

# Transfer Learning for Improving Results on Russian Sentiment Datasets

**Anton Golubev**

Bauman Moscow State  
Technical University  
Russia

antongolubev5@yandex.ru

**Natalia Loukachevitch**

Lomonosov Moscow State  
University  
Russia

louk\_nat@mail.ru

## Abstract

In this study, we test transfer learning approach on Russian sentiment benchmark datasets using additional train sample created with distant supervision technique. We compare several variants of combining additional data with benchmark train samples. The best results were achieved using three-step approach of sequential training on general, thematic and original train samples. For most datasets, the results were improved by more than 3% to the current state-of-the-art methods. The BERT-NLI model treating sentiment classification problem as a natural language inference task reached the human level of sentiment analysis on one of the datasets.

**Keywords:** Targeted sentiment analysis, distant supervision, transfer learning, BERT

**DOI:** 10.28995/2075-7182-2021-20-268-277

## 1 Introduction

Sentiment analysis or opinion mining is an important natural language processing task used to determine sentiment attitude of the text. One of its main business application is product monitoring consisting of studying customer feedback and needs. Nowadays most state-of-the-art results are obtained using deep learning models, which require training on specialized labeled data.

In recent years transfer learning has earned widespread popularity. This approach includes a pre-training step of learning general representations from a source task and an adaptation step of applying previously gained knowledge to a target task. In other words, deep learning model trained for a task is reused as the starting point for a model on a second task. Since there is a significant amount of text information nowadays, current state-of-the-art results can be possibly improved using transfer learning.

The most known Russian sentiment analysis datasets include ROMIP-2013 and SentiRuEval2015-2016 [4, 11, 12], which consist of annotated data on banks and telecom operators reviews from Twitter messages and news quotes. Current best results on these datasets were obtained using pre-trained RuBERT [20, 7] and conversational BERT model [23, 3] fine-tuned as architectures treating a sentiment classification task as a natural language inference (NLI) or question answering (QA) problem [7].

In this study, we introduce a method for automatic generation of annotated sample from a Russian news corpus using distant supervision technique. We compare different variants of combining additional data with original train samples and test the transfer learning approach based on several BERT models. For most datasets, the results were improved by more than 3% to the current state-of-the-art performance. On SentiRuEval-2015 Telecom Operators Dataset, the BERT-NLI model treating a sentiment classification problem as a natural language inference task, reached human level according to one of the metrics.

The contributions of this paper are presented below:

- we propose a new method of automatic generation of additional data for sentiment analysis tasks from raw texts using the distant supervision approach and a sentiment lexicon,
- we compare several variants of combining the additional dataset with original train samples and show that three-step approach of sequential training on general, thematic and benchmark train samples performs better,
- we renew the best results on five Russian sentiment analysis datasets using pre-trained BERT models combined with transfer learning approach,

- we show that BERT-NLI model treating sentiment classification problem as a natural language inference task reaches human level on one of the datasets.

This paper is structured as follows. In Section 2, we overview related methods applied to the considered task. Section 3 describes sentiment analysis datasets used in this paper. Section 4 represents a process of automatic generation of an annotated dataset using distant supervision approach. In section 5 and 6 we briefly cover main preprocessing steps and BERT models applied in the current study. Section 7 presents the full track of transfer learning study including comparison of different variants of constructing an additional datasets and several ways of their combining with benchmark train samples.

## 2 Related Work

Russian sentiment analysis datasets are based on different data sources [20], including reviews [19, 4], news stories [4], posts from social networks [16, 11, 17]. The best results on most available datasets are obtained using transfer learning approaches based on the BERT model [23], more specifically on RuBERT [3] and Russian variant of BERT [20, 7, 14, 1]. In [7], the authors tested several variants of RuBERT and different settings of its applications, and found that the best results on sentiment analysis tasks on several datasets were achieved using Conversational RuBERT trained on Russian social networks posts and comments. Among several architectures, the BERT-NLI model treating the sentiment classification problem as a natural language inference task usually has the highest results.

For automatic generation of annotated data for sentiment analysis task, researchers use so-called distant supervision approach, which exploits additional resources: users’ tags or manual lexicons [6, 16]. For Twitter sentiment analysis, users’ positive or negative emoticons or hashtags can be used [5, 16, 13]. Authors of [18] use the RuSentiFrames lexicon for creating a large automatically annotated dataset for recognition of sentiment relations between mentioned entities.

In contrast to previous work, in the current study we automatically create a dataset for targeted sentiment analysis, which extracts a sentiment attitude towards a specific entity. The use of an automatic dataset together with manually annotated data allows us to improve the state-of-the-art results.

Table 1: Benchmark sample sizes and sentiment class distributions (%).

Dataset	Train sample				Test sample			
	Vol.	Posit.	Negat.	Neutral	Vol.	Posit.	Negat.	Neutral
ROMIP-2013 <sup>3</sup>	4260	26	44	30	5500	32	41	27
SRE-2015 Banks <sup>4</sup>	6232	7	36	57	4612	8	14	78
SRE-2015 Telecom <sup>4</sup>	5241	19	34	47	4173	10	23	67
SRE-2016 Banks <sup>5</sup>	10725	7	26	67	3418	9	23	68
SRE-2016 Telecom <sup>5</sup>	9209	15	28	57	2460	10	47	43

## 3 Russian sentiment benchmark datasets

In our study, we consider the following Russian datasets (benchmarks) annotated for previous Russian sentiment shared tasks: news quotes from the ROMIP-2013 evaluation [4] and Twitter datasets from SentiRuEval 2015-2016 evaluations [11, 12]. Table 1 presents main characteristics of datasets including train and test sizes and distributions by sentiment classes. It can be seen in Table 1 that the neutral class is prevailing in all Twitter datasets. For this reason, along with the standard metrics of  $F_1$  macro and accuracy,  $F_1^{+-}$  macro and  $F_1^{+-}$  micro ignoring the neutral class were also calculated.

The collection of news quotes contains opinions in direct or indirect speech extracted from news articles [4]. The task of ROMIP-2013 evaluation was to distribute quotations between neutral, positive and negative classes depending on its sentiment. It can be seen in Table 1 that dataset is rather balanced.

<sup>3</sup><http://romip.ru/en/collections/sentiment-news-collection-2012.html>

<sup>4</sup>[https://drive.google.com/drive/folders/1bAxIDjVz\\\_0UQn-iJwhnUwngjivS2kfM3](https://drive.google.com/drive/folders/1bAxIDjVz\_0UQn-iJwhnUwngjivS2kfM3)

<sup>5</sup><https://drive.google.com/drive/folders/0BxlA8wH3PTUfV1F1UTBwVTJPd3c>

Twitter datasets from SentiRuEval-2015-2016 evaluations were annotated for the task of reputation monitoring [15, 11], which means searching sentiment-oriented opinions about banks and telecom companies. In such a way this task can be regarded as an entity-oriented sentiment analysis problem. Insignificant part of samples contains two or more sentiment analysis objects, so these tweets are duplicated with corresponding attitude labels. The SentiRuEval-2016 training datasets are much larger in size as they contain training and test samples of 2015 evaluation [12]. As it can be seen in Table 1, Twitter datasets are poorly balanced. This explains the choice of metrics considering only positive and negative classes.

#### 4 Automatic generation of annotated dataset

The main idea of automatic annotation of dataset for targeted sentiment analysis task is based on the use of a sentiment lexicon comprising negative and positive words and phrases with their sentiment scores. We utilize Russian sentiment lexicon RuSentiLex [10], which includes general sentiment words of Russian language, slang words from Twitter and words with positive or negative associations (connotations) from the news corpus. For ambiguous words, having several senses with different sentiment orientations, RuSentiLex describes senses with references to the concepts of RuThes thesaurus [9]. The current version of RuSentiLex contains 16445 senses.

As a source for automatic dataset generation, we use a Russian news corpus, collected from various sources and representing different topics, which is important in fact that the benchmarks under analysis cover several topics. The corpus was collected long before the evaluations, so there are no possible overlaps between additional and benchmark data. The volume of the original corpus was about 4 Gb of raw text, which implies more than 10 million sentences.

The automatically annotated dataset includes general and thematic parts. For creation of the general part, we select monosemous positive and negative nouns from the RuSentiLex lexicon, which can be used as references to people or companies, which are sentiment targets in the benchmarks. We construct positive and negative word lists and suppose that if a word from the list occurs in a sentence, it has a context of the same sentiment. The list of positive and negative references to people or companies (seed words) includes 822 negative references and 108 positive ones. Examples of such words are presented below (translated from Russian):

- positive: "*champion, hero, good-looker*", etc.;
- negative: "*outsider, swindler, liar, defrauder, deserter*", etc.

Sentences may contain several seed words with different sentiments. In such cases, we duplicate sentences with labels in accordance with their attitudes. The examples of extracted sentences are as follows (all further examples are translated from Russian):

- positive: "*A MASK is one who, on a gratuitous basis, helps the development of science and art, provides them with material assistance from their own funds*";
- negative: "*Such irresponsibility — non-payments — hits not only the MASK himself, but also throughout the house in which he lives*".

To generate the thematic part of the automatic sample, we search for sentences that mention named entities depending on a task (banks or operators) using the named entity recognition model (NER) from DeepPavlov [3] co-occurred with sentiment words in the same sentences. We searched for sentences not only with organizations from benchmarks, but also with others companies from the relevant field. To ensure that a sentiment word refers to an entity, we restrict the distance between two words to be not more than four words.

We remove examples containing a particle "*not*" near sentiment word because it could change attitude of text in relation to target. Sentences with sentiment word located in quotation marks were also removed because they could distort the meaning of the sentence being a proper name. Examples of extracted thematic sentiment sentences are as follows:

- for banks (positive): "*MASK increased its net profit in November by 10.7%*"
- for mobile operators (negative): "*FAS suspects MASK of imposing paid services on subscribers.*"

Since the benchmarks contain also the neutral sentiment class, we need to extract sentences without

sentiments. For this task, we choose among examples selected by NER those that do not contain any sentiment words from the lexicon. Examples of extracted neutral sentences for both general and thematic parts are presented below:

- for persons: "MASK is already starting training with its new team."
- for banks: "On March 14, MASK announced that it was starting rebranding."
- for mobile operators: "MASK has offered its subscribers a new service."

While creating an additional dataset, we take into account the distribution of sentiment words in the resulting sample, trying to bring it as close as possible to uniform. A source corpus contains enough examples with a negative sentiment to form a balanced dataset, which can not be said about words with the positive sentiment. We made automatically generated dataset publicly available<sup>6</sup>.

Table 2: Results based on training on additional dataset only.

Dataset	Model	Accuracy	$F_1$ macro	$F_1^{+-}$ macro	$F_1^{+-}$ micro
ROMIP-2013	BERT-single	28.32	21.54	45.74	46.19
	BERT-pair-QA	28.04	21.32	45.35	45.78
	BERT-pair-NLI	27.76	20.89	45.12	45.68
SRE-2015 Banks	BERT-single	33.42	25.10	39.17	42.29
	BERT-pair-QA	33.19	25.56	38.98	42.31
	BERT-pair-NLI	32.56	24.87	38.63	41.87
SRE-2015 Telecom	BERT-single	26.11	19.12	33.56	34.21
	BERT-pair-QA	26.12	19.05	32.61	34.43
	BERT-pair-NLI	25.13	19.25	31.78	34.02
SRE-2016 Banks	BERT-single	28.91	22.14	36.45	38.88
	BERT-pair-QA	29.43	21.72	35.62	38.26
	BERT-pair-NLI	28.58	20.42	34.38	37.73
SRE-2016 Telecom	BERT-single	25.86	19.57	32.87	34.59
	BERT-pair-QA	25.27	18.76	32.09	33.65
	BERT-pair-NLI	24.14	18.23	31.06	33.28

## 5 Text preprocessing

To create an additional sample from the Russian news corpus, it was necessary to divide raw articles into separate sentences. For this task, we used rule-based sentence splitter from spaCy library [22], which is able to determine sentence boundaries automatically. This solution showed better quality in preliminary studies in comparison with NLTK variant [2] and simple splitter based on regular expressions.

In addition to conceptual steps of creating an automatic dataset described in previous chapter, a few cleaning measures were performed. In accordance with calculated quantiles of sentences from test samples, too short and long examples were removed from additional data. To remove duplicate sentences from different sources, we use the metrics of cosine similarity between pairs of tf-idf representations of examples. When the value of the specified boundary value was exceeded, one of the sentences was randomly removed. Conducting experiments with different thresholds and exploring resulting samples, we set value equal to 0.8.

After bringing the additional sample to the desired format, standard preprocessing track described in [7], including replacing similar text elements with appropriate tokens and removing special symbols was carried out for all datasets.

## 6 BERT architectures

In our study, we consider three variants of fine-tuning BERT models [23] for sentiment analysis. These architectures can be subdivided into the single-sentence approach using only initial text as an input

<sup>6</sup><https://github.com/antongolubev5/Auto-Dataset-For-Transfer-Learning>

and the two-sentence approach [21, 7], which converts the sentiment analysis task into a sentence-pair classification task by appending an additional sentence to the initial text.

The sentence-single model represents a vanilla BERT with an additional single linear layer on the top. The unique token  $[CLS]$  is added for the classification task at the beginning of the sentence. The sentence-pair architecture adds an auxiliary sentence to the original input, inserting the  $[SEP]$  token between two sentences. The difference between two models is in addition of a linear layer: for the sentence-pair model it is added over the final hidden state of  $[CLS]$  token, while for the sentence-single variant it is added on the top of the entire last layer.

In our study, we use pre-trained Conversational RuBERT<sup>7</sup> from DeepPavlov framework [8] trained on Russian social networks posts and comments which showed better results in preliminary study.

Table 3: Results based on training on additional data mixed with benchmark train samples.

Dataset	Model	Accuracy	$F_1$ macro	$F_1^{+-}$ macro	$F_1^{+-}$ micro
ROMIP-2013	BERT-single	65.21	54.32	45.12	44.67
	BERT-pair-QA	65.53	54.68	45.73	45.14
	BERT-pair-NLI	65.45	54.93	45.52	44.89
SRE-2015 Banks	BERT-single	69.34	56.84	36.39	40.19
	BERT-pair-QA	70.21	57.25	36.83	40.79
	BERT-pair-NLI	69.54	57.06	36.65	40.31
SRE-2015 Telecom	BERT-single	66.43	53.19	33.41	37.71
	BERT-pair-QA	66.19	52.83	33.21	37.43
	BERT-pair-NLI	67.11	53.48	33.73	38.03
SRE-2016 Banks	BERT-single	67.71	54.76	33.61	37.85
	BERT-pair-QA	67.61	54.85	34.53	36.89
	BERT-pair-NLI	67.67	54.85	32.12	36.76
SRE-2016 Telecom	BERT-single	65.12	52.43	32.19	36.43
	BERT-pair-QA	64.76	52.06	32.28	36.12
	BERT-pair-NLI	65.21	52.27	32.49	36.51

For the targeted sentiment analysis task, there are labels for each object of attitude so they can be replaced by a special token  $[MASK]$ . Since general sentiment analysis problem has no certain attitude objects, token is assigned to the whole sentence and located at the beginning.

The sentence-pair model has two kind of architecture based on question answering (QA) and natural language inference (NLI) problems. The auxiliary sentences for each model are as follows:

- pair-NLI: "The sentiment polarity of  $MASK$  is"
- pair-QA: "What do you think about  $MASK$ ?"

## 7 Experiments and results

We consider different options of constructing pre-training samples from the collected data and combining the resulting additional dataset with benchmark train samples. Different constructing variants comprise the following options:

- training on the additional general and neutral thematic data only and studying dependence of the results on sentiment class distribution;
- training on the additional general and neutral thematic data mixed with the benchmark training set;
- training on the full generated data (the data of previous steps are extended with sentiment-oriented thematic examples) mixed with the benchmark training set;
- two-step approach: independent sequential training on additional dataset at the first step and on the benchmark training set at the second step;
- study of the dependence of the results on additional dataset size;

<sup>7</sup><http://docs.deeppavlov.ai/en/master/features/models/bert.html>

- three-step approach: independent sequential training in three stages using: the general data part from the additional dataset, the thematic examples from the additional dataset and the benchmark training sets.

All the results presented in the tables below are averaging over 3 experiments with different random initializations of models weights.

### 7.1 Mixing additional data with train samples

As a starting point for research, we train the models only on the automatically generated dataset (general and thematic neutral sentences). We compare two options of constructing the additional sample: uniform balancing between three sentiment classes and balancing in accordance with the average values of classes proportions for all datasets from Table 1.

For both options, the sample size was chosen equal to 15000. The results obtained with uniform balancing are 2-3 % higher and presented in Table 2. It can be seen, that performance is significantly lower than the current state-of-the-art results for all five benchmark datasets.

For the next step, we mix the automatically annotated data with the benchmark training sets. We keep the balance of sentiment classes from the previous experiment. The results are presented in Table 3. For accuracy and  $F_1$  *macro* metrics, the results improve significantly but still did not reach state-of-the-art level. It could be probably explained by assumptions about different topics and styles of texts in additional and benchmark datasets and time dependence of automatically generated data (too many sentences about sports and New Year celebration).

Table 4: Results based on training on extended with sentiment thematic additional data mixed with the benchmark training sets.

Dataset	Model	Accuracy	$F_1$ <i>macro</i>	$F_1^{+-}$ <i>macro</i>	$F_1^{+-}$ <i>micro</i>
ROMIP-2013	BERT-single	66.78	62.44	71.49	70.61
	BERT-pair-QA	67.11	62.18	71.94	71.18
	BERT-pair-NLI	67.89	63.24	72.27	71.65
SRE-2015 Banks	BERT-single	70.54	66.18	67.31	66.59
	BERT-pair-QA	70.87	66.71	68.24	66.91
	BERT-pair-NLI	71.15	67.03	67.69	67.23
SRE-2015 Telecom	BERT-single	67.84	62.31	63.78	62.06
	BERT-pair-QA	68.35	62.44	64.21	62.51
	BERT-pair-NLI	68.89	62.71	65.02	63.12
SRE-2016 Banks	BERT-single	68.14	63.81	63.91	62.33
	BERT-pair-QA	68.81	64.42	65.43	64.16
	BERT-pair-NLI	69.21	65.02	65.76	65.59
SRE-2016 Telecom	BERT-single	67.31	62.15	63.28	61.68
	BERT-pair-QA	67.59	62.31	63.46	62.01
	BERT-pair-NLI	68.16	63.37	64.19	62.21

### 7.2 Extension of additional sample by thematic data

Analyzing low results of the previous experiment, we supposed it may be associated with topic differences between automatic and benchmark datasets, since at this stage an automatic sample was collected using personal descriptive words only. This way, we extend the additional dataset with sentiment thematic examples using the list of well-known organizations (banks and operators) and sentences obtained with NER from DeepPavlov, keeping sample size and sentiment class ratio unchanged.

The results are presented in Table 4. For all  $F_1$  metrics, the performance seems much better than in the previous experiment (mixed general additional sample and training benchmark datasets), but still worse than current state-of-the-art results.

### 7.3 Two-step transfer learning approach

The two-step transfer learning consists in the sequential training on two samples and differs from the previous one in that we do not mix automatically generated data with benchmarks train sets. At the first step, the models are trained on the additional data, then model weights are frozen and training continues on the training data from the benchmarks.

During the same experiment, we study the dependence between the results and size of additional dataset. It was found that with increasing sample size, the results improve too. The boundary between extension of additional dataset and increasing the results was set at a sample size of 27000 (9000 per each sentiment class). Using the two-step approach allows us to overcome the current best results for almost all datasets. The results of described experiment and comparison with the state-of-the-art results [20, 7] are presented in Table 5.

Table 5: Results based on using the two-step approach.

Dataset	Model	Accuracy	$F_1$ macro	$F_1^{+-}$ macro	$F_1^{+-}$ micro
ROMIP-2013	BERT-single	79.95	71.16	85.39	85.61
	BERT-pair-QA	80.21	71.29	85.72	85.93
	BERT-pair-NLI	<b>80.56</b>	<b>71.68</b>	<b>86.14</b>	<b>86.19</b>
	Current SOTA	80.28	70.62	85.52	85.68
SRE-2015 Banks	BERT-single	86.06	79.11	64.87	66.73
	BERT-pair-QA	86.34	79.58	65.29	67.02
	BERT-pair-NLI	<b>87.62</b>	<b>80.72</b>	<b>68.44</b>	<b>71.39</b>
	Current SOTA	86.88	79.51	67.44	70.09
SRE-2015 Telecom	BERT-single	77.11	69.76	61.89	66.95
	BERT-pair-QA	<b>78.14</b>	<b>70.03</b>	<b>64.53</b>	<b>68.29</b>
	BERT-pair-NLI	77.96	69.68	64.52	68.21
	Current SOTA	76.63	68.54	63.47	67.51
SRE-2016 Banks	BERT-single	81.94	74.08	67.24	70.68
	BERT-pair-QA	<b>84.36</b>	<b>77.43</b>	<b>72.32</b>	<b>74.06</b>
	BERT-pair-NLI	84.19	75.63	68.52	70.89
	Current SOTA	82.28	74.06	69.53	71.76
SRE-2016 Telecom	BERT-single	75.82	69.78	65.04	74.22
	BERT-pair-QA	77.25	69.71	67.35	76.22
	BERT-pair-NLI	<b>77.59</b>	69.84	<b>68.11</b>	75.93
	Current SOTA	–	<b>70.68</b>	66.40	<b>76.71</b>

### 7.4 Three-step transfer learning approach

For the final experiment of the study, we divide the first step of the previous experiment into two: sequential training on the general and thematic data. At first, the models are trained on the general data, then the weights are frozen and the training continues on the thematic examples retrieved with the list of organizations and NER from DeepPavlov. After the second weights freezing, the last stage of learning on the original training samples begins. Taken together, this sequence represents the three-step transfer learning approach.

During this experiment, we also changed the additional sample by adding sentiment examples to thematic part of additional sample. The logic consisted in the selection among thematic sentences, those which contain sentiment words. Thus, the first step sample contains 18000 general examples and the second sample consists of 9000 thematic examples (both samples are equally balanced across sentiment classes).

The use of three-step approach combined with addition of sentiment thematic contexts to the sample, improved the results by a few more points. New state-of-the-art results as well as comparison with manual labelling for SentiRuEval-2015 telecom dataset are presented in Table 6. According to the organizers of SentiRuEval-2016 evaluation, one participant sent the results of manual annotation of the test

set [12]. As it can be seen, BERT-pair-NLI model reaches human sentiment analysis level by  $F_1^{+-}$  micro metric.

Table 6: Results based on the three-step approach.

Dataset	Model	Accuracy	$F_1$ macro	$F_1^{+-}$ macro	$F_1^{+-}$ micro
ROMIP-2013	BERT-single	80.27	71.78	85.82	86.07
	BERT-pair-QA	80.78	72.09	86.14	86.42
	BERT-pair-NLI	<b>82.33</b>	<b>72.69</b>	<b>86.77</b>	<b>87.04</b>
	Current SOTA	80.28	70.62	85.52	85.68
SRE-2015 Banks	BERT-single	87.65	80.79	65.74	67.46
	BERT-pair-QA	87.92	81.12	66.47	68.55
	BERT-pair-NLI	<b>88.14</b>	<b>81.63</b>	<b>68.76</b>	<b>72.28</b>
	Current SOTA	86.88	79.51	67.44	70.09
SRE-2015 Telecom	BERT-single	77.85	70.42	62.29	67.38
	BERT-pair-QA	<b>79.21</b>	70.94	65.68	69.11
	BERT-pair-NLI	79.12	<b>71.16</b>	<b>65.71</b>	<b>70.65</b>
	Current SOTA	76.63	68.54	63.47	67.51
	Manual [12]	–	–	70.30	70.90
SRE-2016 Banks	BERT-single	83.21	75.31	68.45	71.69
	BERT-pair-QA	<b>85.59</b>	<b>78.93</b>	<b>74.05</b>	<b>75.12</b>
	BERT-pair-NLI	85.43	76.85	70.23	72.07
	Current SOTA	82.28	74.06	69.53	71.76
SRE-2016 Telecom	BERT-single	76.79	70.64	66.16	75.27
	BERT-pair-QA	78.42	70.54	<b>68.65</b>	<b>77.45</b>
	BERT-pair-NLI	<b>78.62</b>	<b>71.18</b>	69.36	76.85
	Current SOTA	–	70.68	66.40	76.71

## 8 Conclusion

In this study, we presented a method for automatic generation of annotated sample from a Russian news corpus using distant supervision technique. We compared different options of combining additional data with benchmark train samples and improved current state-of-the-art results by more than 3% using BERT models together with the transfer learning approach. The best variant was three-step approach of sequential training on general, thematic and benchmark train samples with intermediate freezing of the model weights. On one of benchmarks, the BERT-NLI model treating a sentiment classification problem as a natural language inference task, reached human level according to one of the metrics.

## Acknowledgments

The reported study was funded by RFBR according to the research project № 20-07-01059.

## References

- [1] Baymurzina DR, Kuznetsov DP, Burtsev MS. Language model embeddings improve sentiment analysis in Russian // *Komp’juternaja Lingvistika i Intellektual’nye Tehnologii*. — 2019. — P. 53–62.
- [2] Bird Steven Edward Loper, Klein Ewan. *Natural Language Processing with Python*. O’Reilly Media Inc. — 2009.
- [3] Burtsev M. DeepPavlov: Open-Source Library for Dialogue Systems // *Proceedings of ACL 2018, System Demonstrations*. — 2018. — P. 122–127.
- [4] Chetviorkin Iliia, Loukachevitch Natalia. Evaluating sentiment analysis systems in Russian // *Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing*. — 2013. — P. 12–17.

- [5] Efficient Twitter sentiment classification using subjective distant supervision / Tapan Sahni, Chinmay Chandak, Naveen Reddy Chedeti, Manish Singh // 2017 9th International Conference on Communication Systems and Networks (COMSNETS) / IEEE. — 2017. — P. 548–553.
- [6] Go Alec, Bhayani Richa, Huang Lei. Twitter sentiment classification using distant supervision // CS224N project report, Stanford. — 2009. — Vol. 1, no. 12. — P. 2009.
- [7] Golubev Anton, Loukachevitch Natalia. Improving Results on Russian Sentiment Datasets // Proc. of the Artificial Intelligence and Natural Language (AINL 2020). — 2020. — P. 109–121.
- [8] Kuratov Yu. Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. — 2019. — P. 122–127.
- [9] Loukachevitch Natalia, Dobrov Boris V. RuThes linguistic ontology vs. Russian wordnets // Proceedings of the seventh global wordnet conference. — 2014. — P. 154–162.
- [10] Loukachevitch Natalia, Levchik Anatolii. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). — 2016. — P. 1171–1176.
- [11] Loukachevitch Natalia, Rubtsova Yuliya. Entity-oriented sentiment analysis of tweets: results and problems // International Conference on Text, Speech, and Dialogue / Springer. — 2015. — P. 551–559.
- [12] Loukachevitch Natalia, Rubtsova Yuliya. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // Proceedings of International Conference Dialog-2016//Proceedings of International Conference Dialog-2016. — 2016.
- [13] Mohammad Saif, Salameh Mohammad, Kiritchenko Svetlana. Sentiment lexicons for Arabic social media // Proceedings of the tenth international conference on language resources and evaluation (LREC’16). — 2016. — P. 33–37.
- [14] Moshkin Vadim, Konstantinov Andrey, Yarushkina Nadezhda. Application of the BERT Language Model for Sentiment Analysis of Social Network Posts // Russian Conference on Artificial Intelligence / Springer. — 2020. — P. 274–283.
- [15] Overview of replab 2013: Evaluating online reputation monitoring systems / Enrique Amigó, Jorge Carrillo De Albornoz, Irina Chugur et al. // International conference of the cross-language evaluation forum for european languages / Springer. — 2013. — P. 333–352.
- [16] Rubtsova Y. Constructing a corpus for sentiment classification training // Software and Systems. — 2015. — no. 109. — P. 72–78.
- [17] Rusentiment: An enriched sentiment analysis dataset for social media in russian / Anna Rogers, Alexey Romanov, Anna Rumshisky et al. // Proceedings of the 27th International Conference on Computational Linguistics. — 2018. — P. 755–763.
- [18] Rusnachenko Nicolay, Loukachevitch Natalia, Tutubalina Elena. Distant supervision for sentiment attitude extraction // Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — 2019. — P. 1022–1030.
- [19] Smetanin Sergey, Komarov Mikhail. Sentiment analysis of product reviews in Russian using convolutional neural networks // 2019 IEEE 21st Conference on Business Informatics (CBI) / IEEE. — Vol. 1. — 2019. — P. 482–486.
- [20] Smetanin Sergey, Komarov Mikhail. Deep transfer learning baselines for sentiment analysis in Russian // Information Processing & Management. — 2021. — Vol. 58, no. 3. — P. 102484.
- [21] Sun Chi, Huang Luyao, Qiu Xipeng. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, V. 1. — 2019. — P. 380–385.

- [22] spaCy: Industrial-strength Natural Language Processing in Python / Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd. — Zenodo, 2020.
- [23] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.