Multi-Step Transfer Learning for Sentiment Analysis

Anton Golubev[1](✉) and Natalia Loukachevitch[2]

[1] Bauman Moscow State Technical University, Moscow, Russia
[2] Lomonosov Moscow State University, Moscow, Russia

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 obtained when the three-step approach is used where the model is iteratively trained 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

1 Introduction

Sentiment analysis or opinion mining is an important natural language processing task used to determine sentiment attitude of the text. Nowadays most state-of-the-art results are obtained using deep learning models, which require training on specialized labeled datasets. To improve the model performance, transfer learning approach can be used. 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.

The most known Russian sentiment analysis datasets include ROMIP-2013 and SentiRuEval2015-2016 [4,10,11] consisting of annotated data on banks and telecom operators reviews from Twitter posts and news quotes. Current best results on these datasets were obtained using pre-trained RuBERT [7,19] and conversational BERT model [3,5] 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.

E. Métais et al. (Eds.): NLDB 2021, LNCS 12801, pp. 209–217, 2021.
https://doi.org/10.1007/978-3-030-80599-9_19

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.

2 Related Work

Russian sentiment analysis datasets are based on different data sources [19], including reviews [4,18], news stories [4] and posts from social networks [10,14,15]. The best results on most available datasets are obtained using transfer learning approaches based on Russian BERT-based models [2,3,5,13,19]. 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, manual lexicons [6,15] and users' positive or negative emoticons in case of Twitter sentiment analysis task [12,15,17]. Authors of [16] use the RuSentiFrames lexicon for creating a large automatically annotated dataset for recognition of sentiment relations between mentioned entities.

3 Russian Sentiment Benchmark Datasets

In our study, we consider the following Russian datasets (benchmarks): news quotes from the ROMIP-2013 evaluation [4] and Twitter datasets from SentiRuEval 2015–2016 evaluations [10,11]. The collection of the news quotes contains opinions in direct or indirect speech extracted from news articles [4]. Twitter datasets from SentiRuEval-2015–2016 evaluations were annotated for the task of reputation monitoring [1,10], which means searching sentiment-oriented opinions about banks and telecom companies.

Table 1 presents the main characteristics of datasets including train and test sample sizes and sentiment classes distributions. It can be seen in Table 1 that the neutral class is prevailing in all Twitter datasets, while ROMIP-2013 data is rather balanced. 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. Insignificant part of samples contains two or more sentiment analysis objects, so these tweets are duplicated with corresponding attitude labels [11].

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

Dataset	Train sample				Test sample			
	Volume	Posit	Negat	Neutral	Volume	Posit	Negat	Neutral
ROMIP-2013[a]	4260	26	44	30	5500	32	41	27
SRE-2015 Banks[b]	6232	7	36	57	4612	8	14	78
SRE-2015 Telecom[b]	5241	19	34	47	4173	10	23	67
SRE-2016 Banks[c]	10725	7	26	67	3418	9	23	68
SRE-2016 Telecom[c]	9209	15	28	57	2460	10	47	43

[a] http://romip.ru/en/collections/sentiment-news-collection-2012.html
[b] https://drive.google.com/drive/folders/1bAxIDjVz_0UQn-iJwhnUwngjivS2kfM3
[c] https://drive.google.com/drive/folders/0BxlA8wH3PTUfV1F1UTBwVTJPd3c

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 [9], which includes general sentiment words of Russian language, slang words from Twitter and words with positive or negative associations (connotations) from the news corpus.

As a source for automatic dataset generation, we use 4 Gb Russian news corpus, collected from various sources and representing different themes, which is an important fact that the benchmarks under analysis cover several topics. For creation of the general part of annotated dataset, 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. Examples of such words are presented below (all further examples are 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:

- 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 relevant named entities depending on a task (banks or operators) using the named entity recognition model (NER) from DeepPavlov [3] cooccurred with sentiment words in the same sentences. To ensure that an attitude word refers to an entity, we restrict the distance between two words to be not more than four words:

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

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

Since the benchmarks contain also the neutral class, we extract sentences without sentiments by choosing among examples selected by NER those that do not contain any sentiment words from the lexicon:

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

To create an additional sample from the raw corpus, we divide raw articles into separate sentences using spaCy sentence splitter library [8]. Too short and long sentences, duplicate sentences (with similarity more than 0.8 cosine measure) were removed. We also take into account the distribution of sentiment words in the resulting sample, trying to bring it as close as possible to uniform. Since negative events are more often included in the news articles, there are much more sentences with a negative attitude in the initial raw corpus than with a positive one. We made automatically generated dataset and source code publicly available[1].

5 BERT Architectures

In our study, we consider three variants of fine-tuning BERT models [5] for sentiment analysis task. These architectures can be subdivided into the single-sentence approach using only initial text as an input and the two-sentence approach [7,20], 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 with an output dimension equal to the number of sentiment classes (3): 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.

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

[1] https://github.com/antongolubev5/Auto-Dataset-For-Transfer-Learning.

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?"*

In our study, we use pre-trained Conversational RuBERT[2] from DeepPavlov framework [3] trained on Russian social networks posts and comments which showed better results in preliminary study. We kept all hyperparameters used in [7] unchanged.

Table 2. 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	**80.56**	**71.68**	**86.14**	**86.19**
	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	**87.62**	**80.72**	**68.44**	**71.39**
	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	**78.14**	**70.03**	**64.53**	**68.29**
	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	**84.36**	**77.43**	**72.32**	**74.06**
	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	**77.59**	69.84	**68.11**	75.93
	Current SOTA	–	**70.68**	66.40	**76.71**

6 Experiments and Results

We consider fine-tuning strategies to represent training in several steps with intermediate freezing of the model weights and include two following variants:

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

- two-step approach: independent iterative training on additional dataset at the first step and on the benchmark training set at the second;
- three-step approach: independent iterative training in three steps using the general part from the additional dataset, the thematic examples from the additional dataset and the benchmark training sets.

During this experiment, we also studied the dependence between the results and the size of additional dataset. It was found that the boundary between extension of automatically generated data and increasing the results was set at a sample size of 27000 (9000 per each sentiment class). Using the two-step approach allowed us to overcome the current best results [7,19] for almost all benchmarks (Table 2).

Table 3. Results based on using 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	**82.33**	**72.69**	**86.77**	**87.04**
	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	**88.14**	**81.63**	**68.76**	**72.28**
	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	**79.21**	70.94	65.68	69.11
	BERT-pair-NLI	79.12	**71.16**	**65.71**	**70.65**
	Current SOTA	76.63	68.54	63.47	67.51
	Manual	–	–	70.30	70.90
SRE-2016 Banks	BERT-single	83.21	75.31	68.45	71.69
	BERT-pair-QA	**85.59**	**78.93**	**74.05**	**75.12**
	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	**68.65**	**77.45**
	BERT-pair-NLI	**78.62**	**71.18**	69.36	76.85
	Current SOTA	–	70.68	66.40	76.71

For a three-step transfer learning approach, we divided the first step of the previous experiment into two. Thus, 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, models are trained on the benchmark training sets.

At this stage we also added sentiment examples to the thematic part of the additional sample via selection thematic sentences containing attitude words. 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 the three-step approach combined with an extension of thematic part of the additional dataset improved the results by a few more points (Table 3). One participant of SentiRuEval-2015 evaluation sent the results of manual annotation of the test sample [11]. As it can be seen, BERT-pair-NLI model reaches human sentiment analysis level by $F_1^{+-} micro$.

Some examples are still difficult for the improved models. For example, the following negative sarcastic examples were erroneously classified by all models as neutral:

- *"Sberbank of Russia – 170 years on the queue market!"*;
- *"While we are waiting for a Sberbank employee, I could have gone to lunch 3 times"*.

In the following example with different sentiments towards two mobile operators, the models could not detect the positive attitude towards the Beeline operator:

- *"MTS does not work! Forever out of reach. The connection is constantly interrupted. We transfer the whole family to Beeline."*

7 Conclusion

In this study, we presented a method for automatic generation of an annotated sample from a news corpus using the distant supervision technique. We compared different options of combining the additional data with several Russian sentiment analysis benchmarks 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 the three-step approach of iterative 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.

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

References

1. Amigó, E., et al.: Overview of RepLab 2013: evaluating online reputation monitoring systems. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 333–352. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40802-1_31

2. Baymurzina, D., Kuznetsov, D., Burtsev, M.: Language model embeddings improve sentiment analysis in Russian. In: Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 53–62 (2019)
3. Burtsev, M.: DeepPavlov: open-source library for dialogue systems. In: Proceedings of ACL 2018, System Demonstrations, pp. 122–127 (2018)
4. Chetviorkin, I., Loukachevitch, N.: Evaluating sentiment analysis systems in Russian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pp. 12–17 (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford **1**(12), 2009 (2009)
7. Golubev, A., Loukachevitch, N.: Improving results on Russian sentiment datasets. In: Filchenkov, A., Kauttonen, J., Pivovarova, L. (eds.) AINL 2020. CCIS, vol. 1292, pp. 109–121. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59082-6_8
8. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python. Zenodo (2020). https://doi.org/10.5281/zenodo.1212303
9. Loukachevitch, N., Levchik, A.: Creating a general Russian sentiment lexicon. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1171–1176 (2016)
10. Loukachevitch, N., Rubtsova, Y.: Entity-oriented sentiment analysis of tweets: results and problems. In: Král, P., Matoušek, V. (eds.) TSD 2015. LNCS (LNAI), vol. 9302, pp. 551–559. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24033-6_62
11. Loukachevitch, N., Rubtsova, Y.: SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. In: Proceedings of International Conference Dialog-2016 (2016)
12. Mohammad, S., Salameh, M., Kiritchenko, S.: Sentiment lexicons for Arabic social media. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 33–37 (2016)
13. Moshkin, V., Konstantinov, A., Yarushkina, N.: Application of the bert language model for sentiment analysis of social network posts. In: Russian Conference on Artificial Intelligence. pp. 274–283. Springer (2020)
14. Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., Gribov, A.: Rusentiment: an enriched sentiment analysis dataset for social media in Russian. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 755–763 (2018)
15. Rubtsova, Y.: Constructing a corpus for sentiment classification training. Softw. Syst. **109**, 72–78 (2015)
16. Rusnachenko, N., Loukachevitch, N., Tutubalina, E.: Distant supervision for sentiment attitude extraction. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 1022–1030 (2019)
17. Sahni, T., Chandak, C., Chedeti, N.R., Singh, M.: Efficient Twitter sentiment classification using subjective distant supervision. In: 2017 9th International Conference on Communication Systems and Networks (COMSNETS), pp. 548–553 (2017)

18. Smetanin, S., Komarov, M.: Sentiment analysis of product reviews in Russian using convolutional neural networks. In: 2019 IEEE 21st Conference on Business Informatics (CBI), vol. 1, pp. 482–486 (2019)
19. Smetanin, S., Komarov, M.: Deep transfer learning baselines for sentiment analysis in Russian. Inf. Process. Manag. **58**(3) (2021)
20. Sun, C., Huang, L., Qiu, X.: Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 380–385 (2019)