Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference

Kamila Alibaeva
Faculty of Computational Mathematics
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
camalibi@yandex.ru

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
louk_nat@mail.ru

Abstract

In this paper we present our approach for stance detection and premise classification from COVID-related messages developed for the RuArg-2022 evaluation. The methods are based on so-called NLI-setting (natural language inference) of BERT-based text classification (Sun et al., 2019), when the input of a model includes two sentences: a target sentence and a conclusion (for example, *positive to masks*). We also use translating Russian messages to English, which allows us to leverage COVID-trained BERT model. Besides, we use additional marking techniques of targeted entities. Our approach achieved the best results on both RuArg-2022 tasks. We also studied the contribution of marking techniques across datasets, tasks, models and languages of RuArg evaluation. We found that "<A:ASPECT> keyword </A:ASPECT>" gave the highest average increase over corresponding basic methods.

Keywords: stance detection, premise classification, natural language inference, BERT, RuArg-2022
DOI: 10.28995/2075-7182-2022-21-8-17

Автоматический анализ авторской позиции и аргументов с использованием архитектуры BERT на основе подхода вывода по тексту

Алибаева К.
МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва
camalibi@yandex.ru

Лукашевич Н.
МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва
louk_nat@mail.ru

Аннотация

В этой статье представлен подход к определению позиции автора и классификации доводов из сообщений, связанных с ковидной инфекцией, разработанный для тестирования RuArg-2022. Предложенные методы основаны на так называемом NLI-варианте (natural language inference, вывод по тексту) использования модели BERT для классификации текстов. При этом подходе на вход модели поступают два предложения: целевое предложение и заключение (например, *позитивно к маскам*). Для классификации также используется перевод сообщений с русского языка на английский, что позволяет использовать специализированную англоязычную модель BERT, дообученную на текстах, посвященных тематике обсуждений ковидной инфекции. Кроме того, мы исследуем дополнительное выделение целевых объектов. Предложенный подход показал наилучшие результаты в обеих задачах RuArg-2022.

Ключевые слова: определение позиции автора, классификация доводов, текстовый вывод, модель BERT, RuArg-2022

1 Introduction

Opinion mining is an an important task in natural language processing (Pang et al., 2008; Liu, 2012). This task was started from general sentiment analysis over texts or text fragments such as users' reviews, which should extract overall authors' sentiment conveyed in texts. More detailed analysis of opinions can be achieved via so-called targeted sentiment analysis, which determines the author's sentiment towards

specific entities or topics, discussed in texts. Sentiment expressed in relation to specific targets can be different from general sentiment of the text.

Targeted sentiment analysis comprises such tasks as aspect-based sentiment analysis (ABSA) (Pontiki et al., 2016), which determines sentiment in relations to aspects (parts or characteristics) of an entity, reputation monitoring of companies and organizations (Amigó et al., 2013; Loukachevitch and Rubtsova, 2015), extraction of an attitude towards some topics, so-called stance detection (Mohammad et al., 2016), or sentiment relations between entities (Rusnachenko et al., 2019).

Another significant direction of opinion mining is argument mining (Lawrence and Reed, 2020). Argument mining tasks are quite diverse, minimal tasks are detection of arguments and classifying them to "for" and "against" classes (so-called premise classification).

For Russian, various tasks of sentiment analysis such as general sentiment analysis, aspect-based sentiment analysis, reputation monitoring have been studied. But only a few works were devoted to stance detection and argument mining. This paper is devoted to description of an approach proposed for the RuArg-2022 evaluation (Kotelnikov et al., 2022), which is devoted to stance detection and premise classification for COVID-related topics discussed in users' comments. We study methods of text classification based on so-called NLI-setting (natural language inference) of BERT-based text classification (Sun et al., 2019; Golubev and Loukachevitch, 2020), for which the input of a model includes two sentences: a target sentence and a conclusion (for example, *positive to masks*. Besides, we use additional marking of targeted entities. Our approach achieves the best results on both RuArg-2022 tasks. We also study the contribution of marking techniques across datasets, tasks and models of RuArg evaluation. We found that marking 4 (<A:ASPECT> keyword </A:ASPECT>) gave the highest average increase over corresponding basic methods.

2 Related Work

Intensive study of the stance detection task in social networks began in 2016, when Mohammad et al. (Mohammad et al., 2016) created the SemEval-2016 dataset containing five independent topics, such as *legalization of abortion* or *Hillary Clinton*. Each of the tweet/topic pairs selected for annotation was annotated via the CrowdFlower crowdsourcing system by at least eight annotators. Sobbani et al. (Sobhani et al., 2019) presented the problem of stance detection on several topics (Multi-target) and created a dataset that consists of three sets of tweets corresponding to target pairs (US presidential candidates): Donald Trump and Hillary Clinton, Donald Trump and Ted Cruz, Hillary Clinton and Bernie Sanders. Tweets with hashtags related to two politicians were extracted to form the dataset. The task was to determine the position of the author (for, against or otherwise) to each of the politicians mentioned in the tweet. Last stance-oriented studies are devoted to extraction of stance towards various aspects of COVID epidemic(Glandt et al., 2021; Miao et al., 2020).

The best results for stance detection in the SemEval-2016 experiments were obtained using the SVM-ngrams classifier, which used word and symbol n-grams (Mohammad et al., 2016) as features. In recent works, it has been found that the best results in stance detection are achieved by approaches based on the BERT (Devlin et al., 2018) neural network model. Gnosh et al. (Ghosh et al., 2019) compared previous approaches and found that the BERT model is the best model for stance detection in the SemEval2016 dataset. The work (Glandt et al., 2021) compares three groups of methods for stance detection regarding aspects of COVID: based on LSTM and CNN networks, and also based on the BERT model. The best results with a large margin are given by models based on BERT.

For Russian, in 2015-2016 a shared task on targeted sentiment analysis was organized (Loukachevitch and Rubtsova, 2015). The participants should extract sentiments towards banks or mobile operators from tweets. Later, the results on these datasets were greatly improved by using BERT-based classifiers (Devlin et al., 2018) and automatically annotated additional data (Golubev and Loukachevitch, 2020; Golubev and Loukachevitch, 2021; Smetanin and Komarov, 2021). The best results were achieved using Russian BERT RuBERT (Kuratov and Arkhipov, 2019) and a sentence-pair classification task (such as natural language inference (NLI)), when auxiliary sentences are added to initial sentences (Sun et al., 2019). In (Nugamanov et al., 2021), a new Russian dataset annotated with stance in relation to

four COVID aspects (masks, quarantine, vaccination, government actions), was presented. For stance detection, classical machine methods, several BERT-based classifiers were used. The best results were obtained with the NLI setting of the BERT model.

In (Vychegzhanin and Kotelnikov, 2017), the authors study stance towards children vaccination. The dataset consists of messages from the social network "VKontakte" classified to two classes: "for" and "against". The best results (84.3 F-measure) were achieved by the SVM classifier with rbf kernel. In subsequent work (Vychegzhanin and Kotelnikov, 2019) additional two topics were considered: "unified state exam" and "human cloning". The best results were obtained using by majority voting based on various classifiers (kNN, SVM, Naive Bayes, etc.)

Etnicity-targeted sentiment analysis was considered in (Koltsova et al., 2020). The task was to determine hate-speech by classifying into three classes. The RuEthnoHate dataset containing 5,5K social media texts has been created. The best results were achieved by deep learning models despite a relatively small dataset size. The performance significantly benefit from a combination of linguistic and sentiment features with BERT pre-training and fine-tuning techniques.

3 Tasks and Data

RuArg-2022 evaluation (Kotelnikov et al., 2022) is devoted to analysis of COVID-related opinions and includes two tasks: stance detection and premise classification in relation to three topics: masks, vaccination, and quarantine during COVID epidemic. In the first task, it is required to determine the point of view (stance) of the text's author in relation to a given topic expressed in a given fragment. In the second task, is is necessary to determine if the text contains premises "for" or "against" to a given claim.

The dataset consists of single sentences. In total, 9,550 sentences were annotated by stance and premises for all three topics. Thus, each sentence has six labels. Each label can have one of the following values: "for", "against", "other", or "irrelevant". The difference between tasks can be explained as follows: the author of opinion can be positive to mask wearing, but does not explain why. In this case, stance to masks is "for", but argument for mask wearing is absent therefore the correct class for premise classification is "other".

Participating systems should automatically annotate each test sentence by stance and premises for each topics separately. In total, six labels (with one of four values) must be assigned to the sentence. For evaluating performance of systems, macro F-measure was used. It was calculated as averaging of F-measures of three relevant categories for each topic - $macroF_{rel}$ (Kotelnikov et al., 2022). In fact, for both tasks four-class classification is carried out, but irrelevant class is not significant for opinion mining therefore macro-averaging over three classes is performed.

4 BERT-based Natural Language Inference

The RuArg-2022 evaluation includes two tasks: stance detection and premise classification, which are both considered as four-class classification for each topic. Additionally, we can single out the relevance classification in both tasks, which separates irrelevant texts for each topic. Irrelevant texts are the same for both stance detection and premise classification. Therefore it is possible to consider a two-stage classification: extraction of relevant texts and then three-class classification for stance detection or premise classification.

Our approach to all tasks is based on NLI (Natural Language Inference) setting of the BERT model (Sun et al., 2019; Golubev and Loukachevitch, 2020). The NLI method adds to every input sentence an additional sentence, which can be an 'assumption' of the original sentence class. In the relevance classification task, the assumption sentence was the aspect itself ('Masks', 'Vaccination', 'Quarantine'). For other tasks, the assumption includes also a stance option (for the stance classification task) and a sentiment (for the premise classification task). In this way, the multi-label classification tasks are transformed into binary classifications for the model. Such a binary classification model for stance detection or premise classification can be trained and applied to all topics, it does not require training separate classifier for each topic. The NLI model was selected because it showed high performance in previous studies (Sun et al., 2019; Golubev and Loukachevitch, 2021; Nugamanov et al., 2021).

Table 1 shows examples of input for all tasks: relevance classification, stance detection, and premise classification.

We studied the following configurations based on the same NLI-BERT approach, where the RuBERT conversational model (Kuratov and Arkhipov, 2019) was used for Russian text representation:

- two-stage classification – three classifiers: relevance detection on the first stage, three-class classifiers for stance detection and premise classification applied to relevant sentences (**2stage3classf**),
- two four-class classifiers for stance detection and premise classification (**1stage2classf**),
- two-stage classification – two classifiers: relevance detection on the first stage, a single classifier for both tasks (stance detection and premise classification) (**2stage2classf**),
- three-stage processing – two classifiers using English translation: relevance detection on the first stage based on Russian texts, machine translation of Russian sentences into English, a single classifier for both tasks of translated texts using a specialized COVID-tuned BERT (**3stageEnglish2classf**) .

For the relevance task, no additional text preprocessing was applied, the input sentence pairs were in Russian in all settings. The relevance classifier was trained using entire training part of the dataset.

For the 3stageEnglish2classf method, all texts having at least one relevant topic were translated into English using Helsinki-NLP/opus-mt-ru-en model from HuggingFace Transformers library. Stance detection and premise classification tasks were combined into a single task via NLI method so every pair of the aspect and the current text that is relevant to this aspect was transformed into six input objects (three for each of the stance options and three for each of the sentiments for the premise classification). To run test cases, for every test text the relevant aspects set was obtained with the trained relevance classifier and then, as before, for each such aspect from the set six input examples were obtained. The classifier gives probabilities of answers "yes" (1) or "no" (0) for each label and topic in the stance detection and premise classification tasks. The final label is chosen according to the maximum soft-max BERT-classifier output value of the outputs for each label option.

The classifiers include the BERT models (different for each task) and the full-connected neural network. For all the above-described configurations, the same BERT parameters were used: learning rate = 0.000005, batch size = 16, epochs = 2. The parameters were not selected using validation with extra data because BERT learning is a very expensive and enduring process. The full-connected network consists of the dropout layer, the linear layer (BERT hidden size layer, 256) size, ReLU activation function, one more dropout layer and one more linear layer (256, number of classes) size, the softmax layer.

For the relevance task DeepPavlov/rubert-base-cased-conversational model was used that was trained on OpenSubtitles, Dirty, Pikabu, and a Social Media segment of Taiga corpus. For the stance detection and premise classification task specialized covid BERT model [1] was used which was pretrained on a corpus of English messages from Twitter about COVID-19.

5 Additional Marking of Target Entities

Targeted opinion mining including stance detection and premise classification involves a target entity (topic). In this respect it is similar to relation extraction, which involves two entities. In previous works, several entity representation methods for relation extraction were proposed (Zhou and Chen, 2021), which we decided to compare in the RuArg evaluation. We evaluated the following entity representation techniques based on entity representations for relation extraction:

1. **Entity mask**. This technique introduces new special tokens [ASPECT] to mask the supposed entities in the original text, where ASPECT is substituted with one of the three topics studied in the evaluation; entity type,
2. **Entity marker**. This technique introduces a special tokens pair [E0], [/E0] to enclose the topic entity, therefore modifying the input text to the format of “[E0] keyword[/E0]”,
3. **Entity marker (punct)**. This technique is a variant of the previous technique that encloses entity spans using punctuation. In our case, it modifies the input text to “* keyword*. In contrast to the previous technique, this one does not introduce new special tokens into the model’s vocabulary,

[1] digitalepidemiologylab/covid-twitter-bert-v2

Task	Sentence	Aspect	Tokenized input
Relevance	('I don't get it. They said it was enough to wear a mask and gloves so they wouldn't get infected when you left the street.', 'Masks')	Masks	['[CLS]', 'i', 'don', "'", 't', 'get', 'it', '.', 'they', 'said', 'it', 'was', 'enough', 'to', 'wear', 'a', 'mask', 'and', 'gloves', 'so', 'they', 'wouldn', "'", 't', 'get', 'infected', 'when', 'you', 'left', 'the', 'street', '.', '[SEP]', 'masks', '[SEP]']
Relevance	('At a time when the time is right to introduce quarantine, it seems too early.', 'Vaccination')	Vaccination	['[CLS]', 'at', 'a', 'time', 'when', 'the', 'time', 'is', 'right', 'to', 'introduce', 'qu', 'aran', 'tine', ',', 'it', 'seems', 'too', 'early', '.', '[SEP]', 'va', 'cci', 'nation', '[SEP]']
Stance Detection	('Vacation would only give rise to the spread of the virus, and it was not the weekend that had to be declared but the quarantine.', 'Against Quarantine')	Quarantine	['[CLS]', 'vacation', 'would', 'only', 'give', 'rise', 'to', 'the', 'spread', 'of', 'the', 'virus', ',', 'and', 'it', 'was', 'not', 'the', 'weekend', 'that', 'had', 'to', 'be', 'declared', 'but', 'the', 'qu', '##aran', '##tine', '.', '[SEP]', 'against', 'qu', '##aran', '##tine', '[SEP]']
Stance Detection	('[USER], the virus is smaller than the mask cells, and the drops of water with which it flies are larger', 'None-stance Masks')	Masks	['[CLS]', '[', 'user', ']', ',', 'the', 'virus', 'is', 'smaller', 'than', 'the', 'mask', 'cells', ',', 'and', 'the', 'drops', 'of', 'water', 'with', 'which', 'it', 'flies', 'are', 'larger', '[SEP]', 'none', '-', 'stance', 'masks', '[SEP]']
Premise Classification	('When I started buying groceries and started wearing a mask, everyone laughed.', 'Neutral to masks')	Masks	['[CLS]', 'when', 'i', 'started', 'buying', 'groceries', 'and', 'started', 'wearing', 'a', 'mask', ',', 'everyone', 'laughed', '.', '[SEP]', 'neutral', 'to', 'masks', '[SEP]']
Premise Classification	('China, without any vaccine, managed the infection, the method of self-isolation.', 'Positive to quarantine')	Quarantine	['[CLS]', 'china', ',', 'without', 'any', 'vaccine', ',', 'managed', 'the', 'infection', ',', 'the', 'method', 'of', 'self', '-', 'isolation', '.', '[SEP]', 'positive', 'to', 'qu', '##aran', '##tine', '[SEP]']

Table 1: NLI-based BERT input representation.

4. **Typed entity marker**. This technique incorporates the stance topic types into entity markers. In our case, it introduces new special tokens "A:ASPECT", "/A:ASPECT", where ASPECT is the corresponding stance topic. The input text is accordingly modified to "<A:ASPECT> keyword </A:ASPECT>",
5. **Typed entity marker (punct)**. This variant marks the target span and target types without introducing new special tokens, which in our case looks as follows: * @ ASPECT @ keyword *.

Used keywords are shown in Table 2. We gathered keywords for marking in the following way. The RuArg evaluation concerns three topics: masks, quarantine, and vaccines. Thus, we selected these words (*mask, quarantine, vaccination*) as initial keywords for marking. Besides, we added synonyms of initial keywords and morphologically related words, known examples of vaccines. Each marking method replaces a word from the keyword list in case if the current text is relevant to the corresponding topic of the current word. All the techniques of marking are illustrated in Table 3.

All keywords were prepared in Russian. To use keywords for the stance detection and premise classification tasks in the English-based 3stageEnglish2classf approach, Russian keywords were translated using the same translation model as for the dataset's texts. Original Russian lists are bigger than translated ones because some Russian words have the same translations into English. In Table 2 both original

Russian and auto-translated English variants are presented.

Aspect	Related words list (Russian)	Related words list (English)
Masks	'маска', 'масочный'	'mask'
Quarantine	'карантин', 'карантинный', 'локдаун'	'quarantine', 'lockdown'
Vaccination	'вакцина', 'вакцинный', 'вакцинация', 'иммунизация', 'вакцинировать', 'вакцинирование', 'прививка', 'прививать', 'прививочный', 'спутник', 'спутник v', 'модерна', 'pfizer', 'ковивак', 'эпиваккорона', 'astrazeneca'	'vaccine', 'immunization', 'vaccination', 'satellite', 'satellite v', 'moderna', 'pfizer', 'quivac', 'epivaccorone', 'astruseneca'

Table 2: Keyword lists used in marking methods.

No.	Marking rule	Text	Text with markers
1	[ASPECT]	'He certainly didn't make it, two weeks quarantine, and he went to work, healthy, not infected!'	'He certainly didn't make it, two weeks [QUARANTINE], and he went to work, healthy, not infected!'
2	[E0] keyword [/E0]	'After all, when a person wears a mask on his face and mouth, the perfect habitat for every micro-organ appears.'	'After all, when a person wears a [E0] mask [/E0] on his face and mouth, the perfect habitat for every mibaselinecro-organ appears.'
3	* keyword *	'The normal decline must be after vaccination, at least the people will get less sick.'	'The normal decline must be after * vaccination *, at least the people will get less sick.'
4	<A:ASPECT> keyword </A:ASPECT>	'Now the finals are accepted only by your citizens, and ours either fly directly to you or wait for the quarantine to be cancelled.'	'Now the finals are accepted only by your citizens, and ours either fly directly to you or wait for the <A:QUARANTINE> quarantine </A:QUARANTINE> to be cancelled.'
5	* @ ASPECT @ keyword *	'Then the academic epidemiologist Gundarov said everything about death masks and panic.'	'Then the academic epidemiologist Gundarov said everything about death * @ MASKS @ masks * and panic.'

Table 3: Marking methods.

6 Results on the RuArg Dataset

Table 4 shows results obtained with all approaches described in Section 4 on the validation part of the RuArg dataset. Approaches 2stage2classf and 3stageEnglish2classf were also applied with all the marking methods. It can be seen that all models obtained much better results than the baseline model provided by RuArg-2022 organizers. The baseline used " bert-base-cased" model from Hugging Face[2]. Three BERT models were trained separately for all three topics: "masks", "vaccines", "quarantine".

[2] https://huggingface.co/bert-large-cased

In our proposed approach, three best methods (3stageEnglish2classf, 3stageEnglish2classf + marking4, 3stageEnglish2classf + marking5) were applied to the test part of the RuArg dataset. The final RuArg leaderboard is shown in Table 5. The best results (camalibi) in the competition were obtained with the proposed model 3stageEnglish2classf + marking5, its scheme is shown in 1.

Approach	**Stance Detection**	**Premise Classification**
Baseline	39.24	45.17
2stage3classf	59.76	54.25
1stage2classf	61.74	61.05
2stage2classf	60.85	57.70
2stage2classf + marking1	60.64	61.14
2stage2classf + marking2	62.23	63.26
2stage2classf + marking3	60.29	58.89
2stage2classf + marking4	62.57	62.93
2stage2classf + marking5	58.59	58.72
3stageEnglish2classf	**69.81**	**67.81**
3stageEnglish2classf + marking1	68.82	66.64
3stageEnglish2classf + marking2	67.33	67.06
3stageEnglish2classf + marking3	67.54	**68.47**
3stageEnglish2classf + marking4	**71.30**	**67.37**
3stageEnglish2classf + marking5	**71.29**	66.55

Table 4: Results of the proposed methods on the validation set.

Participant	**Stance Detection**	**Premise Classification**
camalibi	**69.68**	**74.04**
sevastyanm	68.15	72.35
iamdenay	66.76	65.55
ursdth	65.73	70.64
sopilnyak	56.03	43.38
kazzand	55.52	56.03
morty	53.53	54.53
invincible	52.86	54.28
dr	47.50	60.36
baseline	41.80	43.55

Table 5: Results on the RuArg test set. The best results (camalibi) in the competition were obtained with the proposed model 3stageEnglish2classf + marking5.

From Tables 4, 5, we can see that the contribution of marking methods may vary for different datasets and tasks. For example, marking 2 and 4 improve the basic method 2stage3classf in the stance detection task, marking techniques 1, 2, 4, 5 improve premise classification for the same basic method. The 3stageEnglish2classf basic method on the validation dataset can be improved using marking 4 and 5 for stance detection and marking 3 for premise classification.

To measure contribution of markers across tasks and datasets, we calculated the average improvement of marking techniques over a corresponding basic method for: two tasks of RuArg-2022, two datasets (validation and test) and two methods: Russian-based 2stage3classf and English-based 3stageEnglish2classf. The results of averaging are presented in Table 6. We can see that for stance detection the best techniques are markings 4, 5. For premise classification, the best marking technique is marking 4, and marking 5 is similar to a basic method on average. Thus, we can conclude that the marking method 4

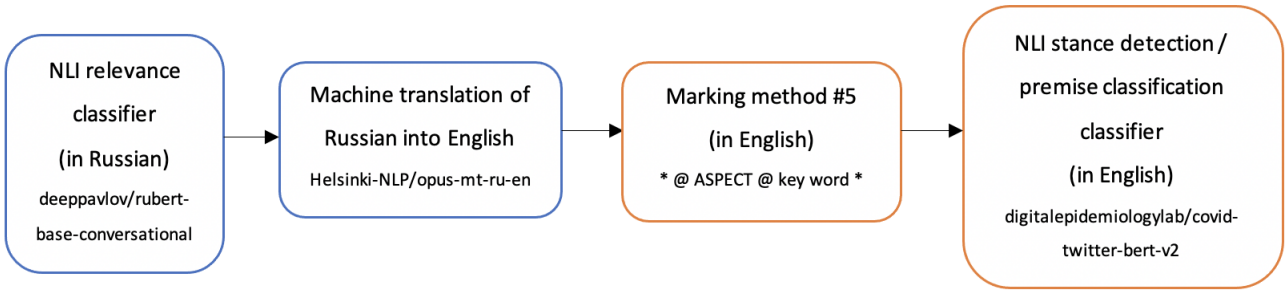

Figure 1: The pipeline scheme achieved the best results in the RuArg evaluation.

(<A:ASPECT> keyword </A:ASPECT>) was the best on average on the RuArg tasks. These results correlate with findings of (Zhou and Chen, 2021), which found that marking 4 (Typed entity markers) was best for BERT-based relation extraction.

Marking No.	Stance Detection	Premise Classification
1	0.46	-1.5
2	-0.71	-2.26
3	0.24	-0.61
4	**0.79**	**1.94**
5	1.65	-0.2

Table 6: Average score difference on marking methods.

7 Error Analysis

Tables 7 and 7 present confusion matrices for stance detection and premise classification. It can be seen that opposite labels are rarely mixed up. It is more difficult to distinguish between any polar opinion (stance or argument) and neutral one. Thus, the study should be continued to understand how best to find markers of difference between polar and neutral opinions.

Model prediction	Irrelevant	Against	Other	For
Irrelevant	2824	6	7	3
Against	0	144	46	5
Other	0	97	715	105
For	0	10	93	238

Table 7: Confusion matrix (stance classification).

Model prediction	Irrelevant	Against	Other	For
Irrelevant	2824	15	7	0
Against	0	72	54	4
Other	0	64	1087	47
For	0	7	37	81

Table 8: Confusion matrix (premise classification).

Tables 9 and 10 present examples that were misclassified by the best model.

Topic	Text	True label	Model prediction
vaccination	'Andrei will be first in line for the vaccine.'	For	Other
masks	'And only with masks do we want to stop this second wave?'	Other	For
quarantine	'It's funny– quarantine doesn't work, treatment doesn't work, prevention is funny– but!'	Other	Against

Table 9: Misclassification examples (stance detection)

Topic	Text	True label	Model prediction
vaccination	'Who will not be destroyed by the coronavirus, drugs and vaccines will kill'	Against	Other
masks	'In Japan and Korea and before the pandemic, the people wore masks, especially in transport.'	Other	For
quarantine	'[USER], quarantine is only a deterrent measure, not a neutralization of the virus.'	For	Against

Table 10: Misclassification examples (premise classification).

8 Conclusion

In this paper we presented our approach for stance detection and premise classification in argument mining from COVID-related messages developed for the RuArg-2022 evaluation. The proposed method is based on so-called NLI-setting (natural language inference) of BERT-based text classification, when the input of a model includes pair of sentences: a target sentence and a conclusion (for example, *positive to masks*) and should predict if a conclusion can be entailed from the target sentence. We also used translation of Russian messages to English, which allowed us to leverage a specialized BERT model pretrained on a text collection of COVID-related tweets. Besides, we used additional marking of targeted entities. Our approach achieved the best results on both RuArg-2022 tasks.

We also studied the contribution of marking techniques across datasets, tasks ans models of RuArg evaluation. We found that marking 4 (<A:ASPECT> keyword </A:ASPECT>) gave the highest average increase over corresponding basic methods. In the current evaluation, aspects for marking were very easy to determine. In future, we plan to integrate various techniques for aspect (topic) identification to use them for improving performance in opinion mining tasks.

Acknowledgements

The work is supported by the Russian Science Foundation, grant #21-71-30003.

References

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. // *International conference of the cross-language evaluation forum for european languages*, P 333–352. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. // *International Conference of the Cross-Language Evaluation Forum for European Languages*, P 75–87. Springer.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

Anton Golubev and Natalia Loukachevitch. 2020. Improving results on russian sentiment datasets. // *Conference on Artificial Intelligence and Natural Language*, P 109–121. Springer.

Anton Golubev and Natalia Loukachevitch. 2021. Multi-step transfer learning for sentiment analysis. // *International Conference on Applications of Natural Language to Information Systems*, P 209–217. Springer.

Olessia Koltsova, Svetlana Alexeeva, Sergei Pashakhin, and Sergei Koltsov. 2020. Polsentilex: Sentiment detection in socio-political discussions on russian social media. // *Conference on Artificial Intelligence and Natural Language*, P 1–16. Springer.

Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, and Alexander Panchenko. 2022. RuArg-2022: Argument Mining Evaluation. // *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Natalia Loukachevitch and Yuliya Rubtsova. 2015. Entity-oriented sentiment analysis of tweets: results and problems. // *International Conference on Text, Speech, and Dialogue*, P 551–559. Springer.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. // *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. // *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, P 31–41.

Eduard Nugamanov, Natalia Loukachevitch, and Boris Dobrov. 2021. Extracting sentiments towards covid-19 aspects. // *Supplementary 23rd International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2021*, P 299–312.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. *Proceedings of SemEval*, P 19–30.

Nicolay Rusnachenko, Natalia Loukachevitch, and Elena Tutubalina. 2019. Distant supervision for sentiment attitude extraction. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, P 1022–1030.

Sergey Smetanin and Mikhail Komarov. 2021. Deep transfer learning baselines for sentiment analysis in russian. *Information Processing & Management*, 58(3):102484.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2019. Exploring deep neural networks for multitarget stance detection. *Computational Intelligence*, 35(1):82–97.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. 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, Volume 1 (Long and Short Papers)*, P 380–385.

Sergey V Vychegzhanin and Evgeny V Kotelnikov. 2017. Stance detection in russian: a feature selection and machine learning based approach. // *AIST (Supplement)*, P 166–177.

Sergey V Vychegzhanin and Evgeny V Kotelnikov. 2019. Stance detection based on ensembles of classifiers. *Programming and Computer Software*, 45(5):228–240.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *CoRR*, abs/2102.01373.