PROCEEDINGS OF SCIENCE

The Russian language corpus and a neural network to analyse Internet tweet reports about Covid-19

Alexander Sboev,[a,b] **Ivan Moloshnikov,**[a] **Alexander Naumov,**[a,*] **Anastasia Levochkina**[a] **and Roman Rybka**[a]

[a] *National Research Center "Kurchatov Institute", 1, Akademika Kurchatova pl., Moscow, 123182, Russia*
[b] *National Research Nuclear University MEPHI, 31 Kashirskoe Shosse, Moscow, 115409, Russia*

E-mail: sag111@mail.ru

This work is aimed at creating a tool for filtering messages from Twitter users by the presence of mentions of coronavirus disease in them. For this purpose, a corpus of Russian-language tweets was created, which contains the part of 10 thousand tweets that are manually divided into several classes with different levels of confidence: potentially have covid, have covid now, other cases, and an unmarked part – 2 million tweets on the topic of the pandemic. The paper presents the process of creating a corpus of tweets from the stage of data collection, their preliminary filtering and subsequent annotation according to the presence of disease description. Machine learning methods were compared according to classification task on tweets. It is shown that a model based on the XLM-RoBERTa topology with additional training on corpus of unmarked tweets gives the F1 score of 0.85 on binary classification task ("potentially have covid have covid now" vs "other"). This is 12% higher relative to the simplest model using TF-IDF encoding and SVM classifier and 5% higher relative to the RuDR-BERT-based model. The created toolkit will expand the feature space of models for predicting the spread of coronavirus infection and other pandemics by adding the dynamics of discussion on social networks, which characterizes people's attitudes towards it.

The 5th International Workshop on Deep Learning in Computational Physics
28-29 June, 2021
Moscow, Russia

[*] Speaker

1. Introduction

The problem of forecasting the evolution of the Covid-19 pandemic is extremely relevant due to the need for hospital beds planning and implementation of containment policies. Given the limited time interval in datasets about Covid-19 evolution, it's reasonable to use machine learning algorithms with relatively low complexity, which efficiency depends mainly on the correct selection of significant features. One of the features we analyze is the number of tweets where Internet users report having Covid-19 symptoms. Extraction of such tweets from the Internet is complicated due to the lack of a Russian-language tweet dataset which is needed for training the machine learning models to automatically extract this category of tweets.

The detection of infectious disease outbreaks using data from Twitter has already been tested, for example, in the case of the Ebola virus [**?**] and the Zika virus [**?**]. A group of researchers in [**?**] proposes to use data from social networks to enhance disease surveillance by determining and predicting the prevalence of influenza on-site. The work [**?**] presents a corpus for the WNUT-2020 competition. It consists of annotated English-language tweets about the coronavirus. The size of the corpus is about 10 thousand messages with annotation for two classes: informative and non-informative tweets.

At the SMM4H 2021 [**?**] competition, two tracks for the analysis of tweets were presented: task 5 (Classification of tweets self-reporting potential cases of COVID-19) and task 6 (Classification of COVID19 tweets containing symptoms).

The existing datasets with annotation on the topic of coronavirus are mainly in English, so we focused on creating the first tool for the classification of the coronavirus mentioning in the texts of Russian-language tweets. Including the dynamics of social media discussions in the calculation of infection indices for SIS [**?**] and SIR [**?**] models can improve the quality of approximation and their predictive properties.

The paper contains a description of the collected and labeled data, as well as a comparison of the accuracy of various classification algorithms based on machine learning methods, including the latest language neural networks models that shows high efficiency in the similar task [**?**].

2. Dataset

2.1 Annotated part of the dataset

2.1.1 Data collecting

To form an annotated corpus, we collected Russian-language tweets containing the word "covid" that were published from 03/01/2020 to 03/01/2021. We randomly selected messages from collected tweets for further annotation. The keyword is frequently used in informal social media conversations related to the COVID-19 pandemic and coronavirus in general. Duplicate tweets, as well as tweets with a length of 10 characters or less, were not collected. In this way, various short tweets (for example, containing the same hashtags) that weren't related to our collection target were filtered out. In total there were ~295 thousands of unique tweets. Next, we randomly selected 10,000 messages from them for a manual annotation.

2.1.2 Labeling

Annotators labeled messages, dividing them into 5 classes:

- potential_covid_high_confidence – potentially has coronavirus with high probability;
- potential_covid_low_confidence – potentially has coronavirus with low probability;
- had_covid_high_confidence – had coronavirus with high probability;
- had_covid_low_confidence – had coronavirus with low probability;
- other – mention coronavirus but don't refer to a potential case of the author or his\her relatives;

potential_covid_high_confidence and potential_covid_low_confidence – these labels mean that the tweet refers to the current potential case of coronavirus disease of the author or his\her relatives.

had_covid_high_confidence and had_covid_low_confidence – the author or someone close to him had coronavirus earlier with a certain probability.

The annotation took into account only the opinion of the author about the disease. Annotators could assign multiple classes to one tweet, for example "relapse" (classes potential_covid_high_confidence and had_covid_high_confidence).

2.1.3 The work of the annotators

The labeling was carried out by 4 annotators in two stages. At the first stage, several annotation iterations were made to generate a high-quality annotation guide. On each iteration we asked annotators to mark up from 100 to 1000 documents and write down all the questions and complex cases that were further analyzed to correct the annotation guide. When the corrections after the iteration became minor, the second stage of the annotation began. Messages from the first stage weren't included in the final version of labeled corpus because during their annotation the guide was constantly being changed and refined.

At the second stage, the annotators tagged the tweets that composed the corpus. The tweets of the corpus were divided into samples, each one contained from 1000 to 2000 unique examples, and 100 common for all samples. The common part allowed to check consistency between annotators for annotation quality control.

If one of the annotators strongly disagreed with others, we examined cases of discrepancy together with them. After the discussion, tweets with potential errors in labeling were re-annotated.

Also, after analyzing each part, complex cases (tweets in which the annotators found it difficult to choose a label) were deleted from the final corpus In general, about 10 000 tweets were labeled.

We selectively checked annotated texts for errors in label selection and corrected mistakes. We paid particular attention to the tweets with notes about the potential illness of the author or his/hers loved ones.

The statistics on the representatives of classes are given in the Table 1.

The annotators consistency was analyzed according to Cohen's kappa[**?**] metric. The calculations were conducted using python "Disagree library" [1].

The resulting agreement between annotators (see Table 2) is high and allows us to make a conclusion about the quality of the resulting annotation guide and the resulting annotated corpus.

[1] https://github.com/o-P-o/disagree

Table 1: Dataset statistics: number of samples in train, test, valid subparts for each class. Class names: p_hc - potential_covid_high_confidence, p_lc - potential_covid_low_confidence, h_hc - had_covid_high_confidence, h_lc - had_covid_low_confidence

part	task	other	p_hc	p_lc	h_hc	h_lc	all_about_covid (p_hc + p_lc + h_hc + h_lc)
train	all	5322	219	253	95	27	594
test	all	1769	67	85	42	9	203
valid	all	1757	69	88	38	20	215

Table 2: Pairwise consistency estimation of the annotators

Annotators ID	Cohen's kappa
1, 2	0.94
1, 3	0.92
1, 4	0.93
2, 3	0.94
2, 4	0.93
3, 4	0.92
Mean value:	0.73

2.2 Corpus of unlabeled tweets

In addition to the corpus of annotated tweets described above, we formed a corpus of unlabeled twitter messages to customize modern language models that demonstrate State-of-the-art results in the most NLP (natural language processing) tasks. The data on keyword "covid" was expanded with texts containing other words often occurred in hashtags on the Covid-19 pandemic: "covid", "stayhome", and "coronavirus" (hereinafter, these are translations of Russian words into English).

Separately, messages were collected from Twitter users from large regions of Russia. The search was provided using different word forms of 58 manually selected keywords on Russian related to the topic of coronavirus infection (including: "PCR", "pandemic", "self-isolation", etc.).

The affiliation of a tweet to one of the regions was determined in two ways:

1. the text of the message directly mentions the capital of the region, for example: "During the second wave of covid in Moscow there were three times more cases";
2. there is a capital of the region in the field "home region" of the author's account, for example, the author of the tweet "Tatarstan doctors explained why vaccination against covid was important and debunked the myths about vaccination..." the home region is indicated: "Kazan".

In total, the 15 largest regional center of Russia were used: Moscow, St. Petersburg, Novosibirsk, Yekaterinburg, Kazan, Chelyabinsk, Samara, Omsk, Ufa, Krasnoyarsk, Voronezh, Perm, Volgograd, Rostov-on-Don, and Nizhny Novgorod.

The final unlabeled corpus includes all unique Russian-language tweets from the collected data (>1M tweets). Since modern language models are usually multilingual, about 1M more tweets in other languages were added to this corpus using filtering procedures described above. Thus, in the unlabeled part of the collected data, there were about 2 million messages.

3. Models

Below there are the approaches that we used to obtain the baseline accuracy for the created corpus.

3.1 TF-IDF and SVM model

A model based on TF-IDF vectorization of the corpus and classification by a support vector machine (SVM) with a linear kernel was chosen as the baseline. Implementation of the scikit-learn [**?**] library was used. We used the TF-IDF method to vectorize the texts, with a document breakdown into n-grams of characters with an n-gram length from 3 to 8 and a minimum occurrence of n-grams in three documents. The SVM classifier was used with a linear kernel and parameter C equal to 10.

3.2 RuDR-BERT model

We used a model based on RuDR-BERT [**?**], implemented with the transformers [**?**] and simpletransformers [2] libraries. RuDR-BERT was fine-tuned for classification of tweets for 15 epochs, with learning rate equal to $4e^{-5}$, batch size 32; other parameters were standard for simpletransformers library.

3.3 XLM-RoBERTa model

We used a language model based on XLM-RoBERTA-large [**?**] from [**?**] (Model B). In this work, it was additionally trained as a masked language model on an unmarked part of the corpus (about 2M tweets on the topic of coronavirus). The training was carried out during 1 epoch with standard hyperparameters from the transformers library (the model denoted as “covid-twitter-xlm-roberta-large”).

Further, the model was fine-tuned for the classification task. The learning process and hyperparameters are similar to the RuDR-BERT model.

4. Evaluation

All models were evaluated using the F1-score metric, with macro averaging over all classes. This averaging was chosen because we consider all classes to be equally important, even thought they have different representations in the collected corpus. Two tasks were considered:

- full set of tags (all) – multilabel classification of tweets, while one tweet can be assigned to several tags;

[2] https://simpletransformers.ai

Table 3: F1-score of the classification models on "all" task. Class names in table: potential_hc - potential_covid_high_confidence, potential_lc - potential_covid_low_confidence, had_lc - had_covid_high_confidence, had_lc - had_covid_low_confidence.

model	ds part	other	potential_hc	potential_lc	had_hc	had_lc	macro_avg
tfidf_svm	train	0.96	0.36	0.39	0.20	0.25	0.43
tfidf_svm	test	0.94	0.09	0.11	0.08	0.00	0.25
tfidf_svm	valid	0.94	0.00	0.16	0.05	0.00	0.23
RuDR-BERT	train	1.00	1.00	1.00	1.00	0.97	0.99
RuDR-BERT	test	0.95	0.46	0.46	0.37	0.10	0.47
RuDR-BERT	valid	0.96	0.52	0.46	0.27	0.18	0.48
covid-twitter-xlm-roberta-large	train	1.00	1.00	1.00	0.99	1.00	1.00
covid-twitter-xlm-roberta-large	test	0.97	0.61	0.50	0.53	0.12	0.54
covid-twitter-xlm-roberta-large	valid	0.97	0.62	0.54	0.46	0.42	0.60

Table 4: F1-score of the classification models on binary task.

model	ds part	other	all_about_covid	macro_avg
tfidf_svm	train	1.00	1.00	1.00
tfidf_svm	test	0.96	0.54	0.75
tfidf_svm	valid	0.95	0.51	0.73
RuDR-BERT	train	1.00	1.00	1.00
RuDR-BERT	test	0.96	0.61	0.78
RuDR-BERT	valid	0.96	0.64	0.80
covid-twitter-xlm-roberta-large	train	1.00	1.00	1.00
covid-twitter-xlm-roberta-large	test	0.97	0.70	0.83
covid-twitter-xlm-roberta-large	valid	0.97	0.72	0.85

- binary classification (bin) – binary classification into two classes: other and all_about_covid, class all_about_covid - combines 4 classes:
  potential_covid_high_confidence, potential_covid_low_confidence
  and had_covid_high_confidence, had_covid_low_confidence;

For developing classification model we have divided the corpus into 3 parts:

- train - 60% of the corpus, is used to train the model;
- valid - 20% of the corpus, is used as a validation set for setting model hyperparameters, and early stopping;
- test - 20% of the corpus, is used as a test part for comparing different models;

Model comparison results are shown in Tables 3 and 4.

For the best model covid-twitter-xlm-roberta-large, we evaluated the effect of the size of the training set on accuracy. To do this, we took parts from a training set ranging from 10% to 100% of the training set with a 10% step. Results are shown in Table 5.

Table 5: F1 score dependency on the size of the training set for covid-twitter-xlm-roberta-large model trained for binary classification task.

train size, %	other	all_about_covid	macro_avg
10	0.95	0.54	0.74
20	0.96	0.62	0.79
30	0.96	0.66	0.81
40	0.97	0.73	0.85
50	0.97	0.70	0.83
60	0.97	0.70	0.84
70	0.96	0.66	0.81
80	0.96	0.64	0.80
90	0.97	0.71	0.84
100	0.97	0.70	0.83

The estimations from the table 5 show that 40% percent of the training set is sufficient to achieve good accuracy. The further increase in the size of the corpus doesn't generate much gain. It could be explained with high domain orientation of the deep neural network model.

5. Experiments

The created "covid-twitter-xlm-roberta-large" model was used to classify tweets from the unlabeled part of the corpus. We used tweets that contained keywords: "covid", "stayhome", "coronavirus", and "stay home" or tweets that had affiliation to one of the regions of Russia. The figure 1 shows the curves of numbers of infected people in whole Russia[3] and of tweets per day for period from 03/01/2020 to 03/01/2021. For each curve we standardized values by removing the mean and scaling to unit variance[4]: $z = (x - u)/s$, where x is the value, u is the mean of the curve values, s is the standard deviation of the curve values.

The daily dynamics of tweets about Covid filtered by the binary model to mentions of coronavirus disease are similar to the official statistics of COVID-19 cases in Russia for the same period. Thus the created model can be used for separating targeted tweets from the other tweets on the topic of COVID-19.

6. Conclusion

As a result, a classification toolkit was created to filter out Russian-language tweets for the presence of references to the coronavirus disease in them. The tool is based on the XLM-RoBERTa neural network language model that additionally trained on the texts of Russian-language medical reviews and Twitter messages on the topic of Covid-19 pandemic. This model allowed us to achieve the best accuracy among other machine learning methods: macro-averaged F1-score is 0.54 for all class determination, and 0.83 for a binary classification task. Model fine-tuning and comparative

[3] Data for all regions of Russia was collected from the Yandex DataLens platform

[4] StandardScaler function from scikit-learn library

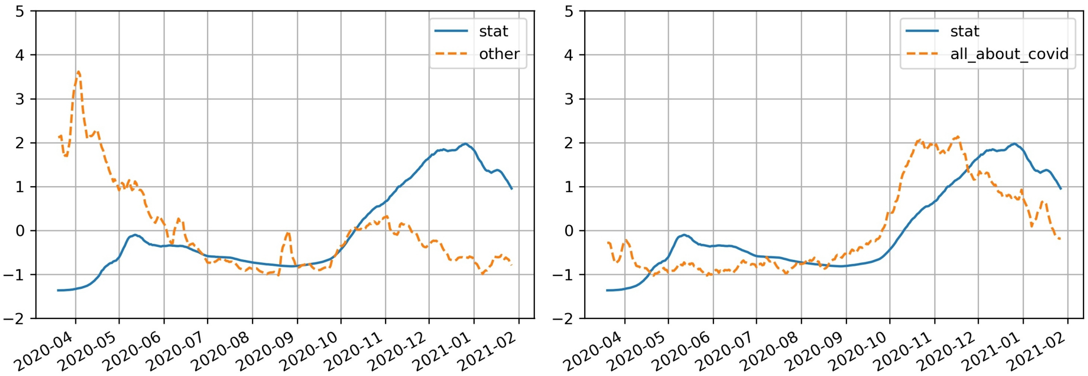

Figure 1: Official statistics of COVID-19 cases in Russia (stat) is presented on both left and right parts of the figure. The left one contains curve of tweets related to the class "other". The right part contains curve of tweets related to the class "all_about_covid".

analysis were conducted on a manually annotated corpus of Twitter posts. Created classification toolkit and the annotated corpus can be used to predict the dynamics of the epidemic development, as well as to identify new symptoms and side effects described by users of social networks. This is the direction of our further research.

Acknowledgments

This study was supported by the Russian Foundation for Basic Research project № 20-04-60528 and carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, http://ckp.nrcki.ru/.

A. Online Resources

Author‘s github:

- https://github.com/sag111/COVID-19-tweets-Russia – repository with code and additional materials for this paper;
- https://github.com/sag111/COVID-19-baselines-Russia – repository for COVID forecasting project;