

5th International Workshop on Deep Learning in Computational Physics



National Research Centre “Kurchatov Institute”, Russia



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

Alexander Sboev, Ivan Moloshnikov,
Alexander Naumov, Anastasia
Levochkina, Roman Rybka

The study was supported by the Russian
Foundation for Basic Research project
№ 20-04-60528

Moscow 2021

Relevance



The world is currently experiencing an epidemic of coronavirus infection. It often takes a long time (up to 10 days) from the moment of infection to the receipt of an official diagnosis, which leads to a delay in the introduction of restrictive measures. For a timely response to the deteriorating epidemiological situation, accurate forecasting tools are needed.

Existing methods are based on data of the infected/recovered/deaths dynamics. The disadvantage of this approach is to work only with official statistics, while the number of unreported cases of the disease is much higher.

This work is aimed at creating tools for filtering Twitter posts (tweets) to obtain data relevant to the epidemic topic with an analysis of the dependencies of real dynamics and dynamics of discussions in Twitter.

Tasks



1. Collecting data from Twitter on COVID-19 topic.
2. Developing a robust tweet annotation scheme, i.e. instructions for annotators.
3. Checking and correcting annotations with quality assessment.
4. Development of a data classification model for the collected corpus using modern NLP methods.
5. Analysis based on the developed neural network model.

Dataset

Collecting data from Twitter



The data was collected by parsing search results from twitter.com using the open source library Twint *. Data were collected from **March 1, 2020 to March 1, 2021**.

The collection was carried out on the topic of coronavirus. The topic is described by a set of keywords: *ковид (covid), коронавирус (coronavirus), вирус (virus), эпидемия (epidemic), спутник (sputnik), ковивак (covivak), эпиваккорона (epivakcorona), вектор (vector), обоняние (smell), вкус (taste), дыхание (breath), кт (СТ), госпиталь (hospital), гибель (death), etc.*

The tweets were collected with reference to the region (indication in the text of the tweet/author's profile). The corpus for the Moscow region consists of 486 thousand tweets.

* <https://github.com/twintproject/twint>

Dataset

Annotation of tweets

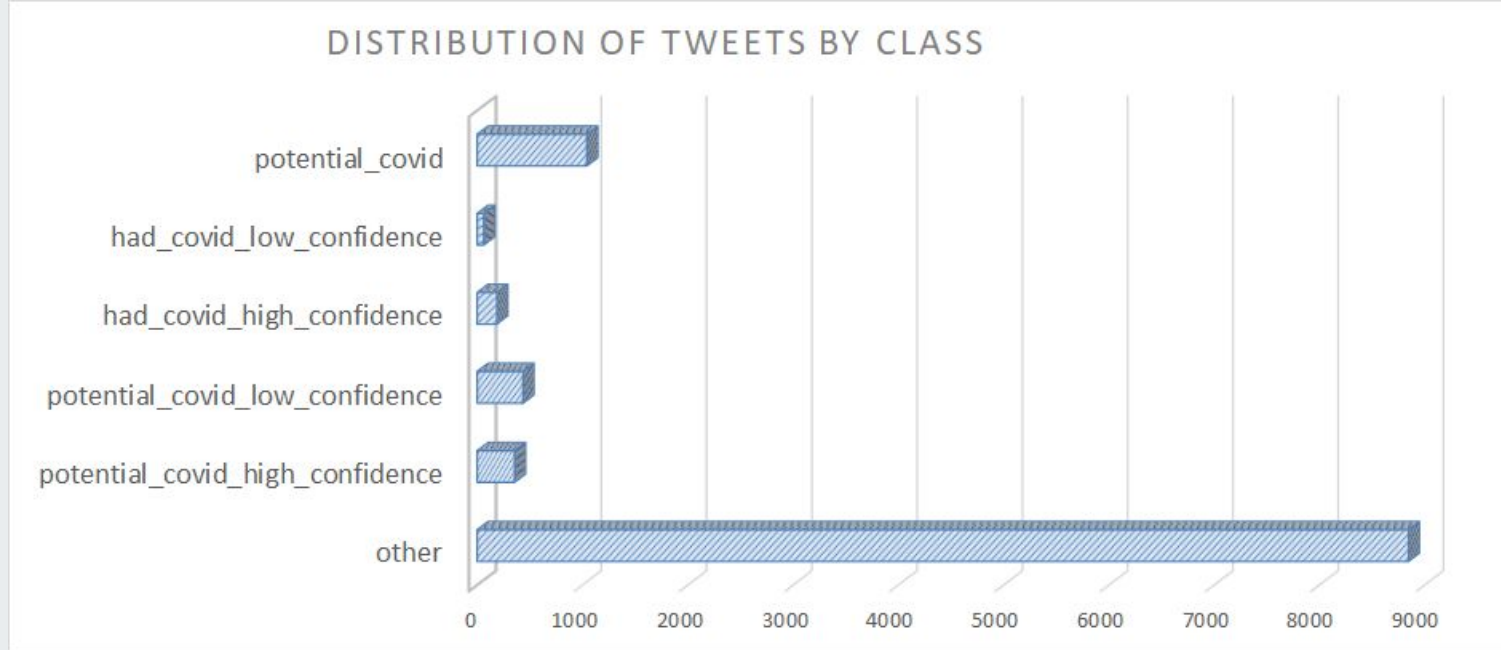


10,000 tweets were tagged by 4 annotators using annotation schemas agreed with machine learning experts. During the markup process, the annotator assigned one (or more) tags. Tag List:

- **potential_covid_high_confidence**
(high confidence that the author claims that he or his relatives are infected)
- **potential_covid_low_confidence**
(low confidence that the author claims that he or his relatives are infected)
- **had_covid_high_confidence**
(high confidence that the author claims that he or his relatives suffered from covid)
- **had_covid_low_confidence**
(low confidence that the author claims that he or his relatives suffered from covid)
- **other** (other tweets)

Dataset

Information about classes



Fleiss' kappa = 0.71

(!) potential_covid = sum for all classes except "other"

Classification model

Machine learning methods

1. Text vectorization based on frequency method: **TF-IDF** for n-gram of text characters, and fitting classification model Support Vector Machine with Linear kernel (**LinearSVC**)
2. Neural network model **RuDR-BERT** that based on Multilingual Cased BERT that pretrained on the top 104 languages with the largest Wikipedia. RuDR-BERT was achieved after training Multi-Bert on the **1.5M Russian reviews** about medications;
3. NN model **XLM-RoBERTa-sag** that based on XLM-RoBERTa model that pretrained on 2 TB of text data in 100 languages (including rare languages) from the CommonCrawl project. XLM-RoBERTa-sag was achieved after training XLM-RoBERTa on **2M Russian reviews** about medications.

Classification model

Estimations

Table 1. Estimation of models for the problem of classifying tweets into 5 classes.

model	micro_f1	macro_f1
tfidf+LinearSVC	0.90	0.33
RuDR-Bert	0.90	0.48
XLM-sag	0.92	0.60

Table 2. Estimation of models for the problem of classifying tweets into 2 classes.

model	micro_f1	macro_f1
tfidf+LinearSVC	0.92	0.73
RuDR-Bert	0.92	0.80
XLM-sag	0.94	0.85

Dataset was divided on 3 parts:

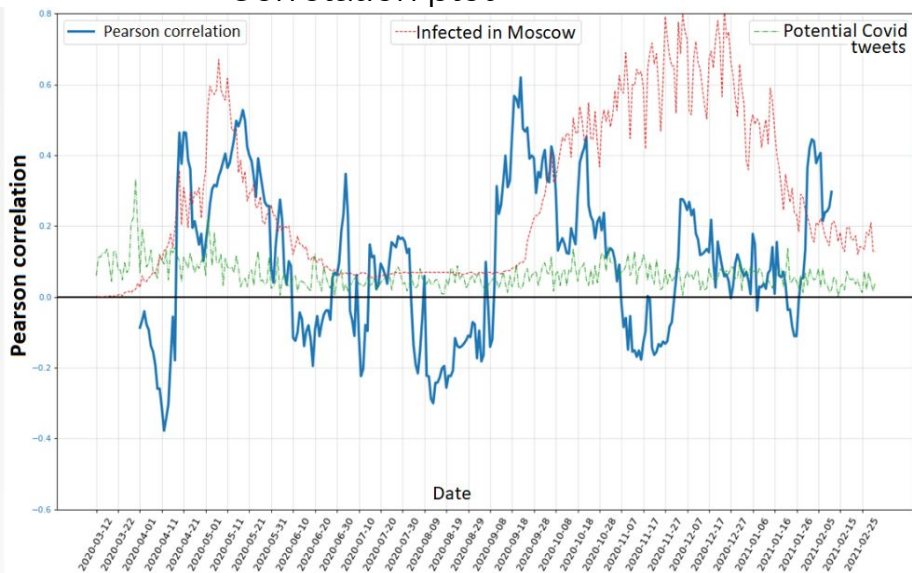
train part 60% of all data, validation part (20%), test part (20%)

Analysing unlabeled data

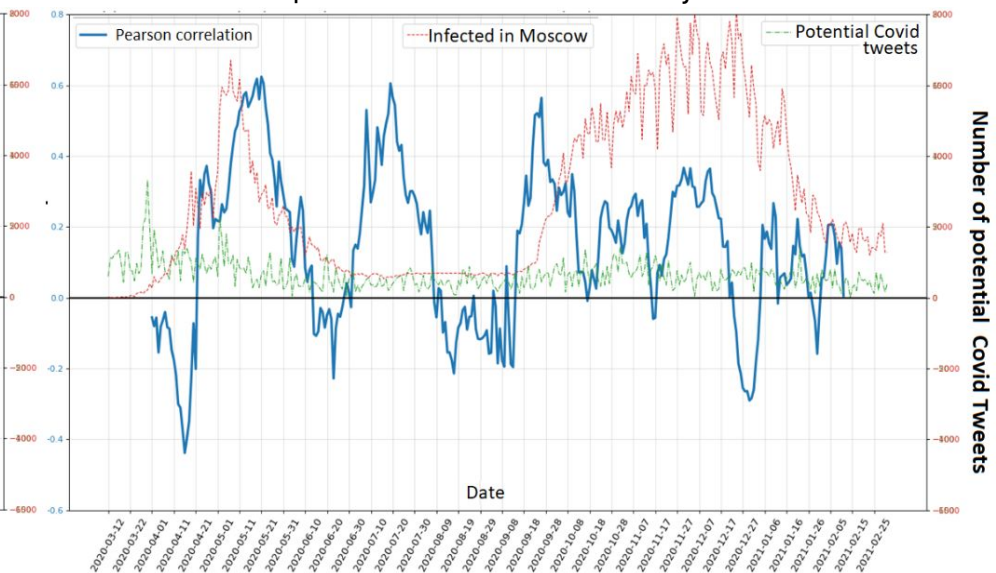
Moscow: real statistic (infected) and dynamics of discussions




Correlation plot



Correlation plot with a shift of 4 days to the



Results and Future plans

- 
1. The dataset of Covid tweets contained labeled (10`000 tweets) and unlabeled (486`000 tweets for Moscow region) parts was formed.
 2. Neural network classifier was trained to determine the class of tweets, where the authors mention cases of illness or damage from Covid in themselves or in the immediate environment
 3. A preliminary analysis of the applicability of Twitter data for inclusion in the general model for predicting the dynamics of the spread of coronavirus was carried out.

Further research will be aimed at finding more relevant signs from discussions on social networks and creating a general model for forecasting the epidemic based on statistical data and social media data.

Thanks for your attention!

