Adv Syst Sci Appl 2023; **03**:108–126
Published online at https://ijassa.ipu.ru.

Exploratory Data Analysis and Natural Language Processing Model for Analysis and Identification of the Dynamics of COVID-19 Vaccine Opinions on Small Datasets

Alexander Chkhartishvili[1], Dmitry Gubanov[2], Vladislav Melnichuk[3,4*], Vladislav Sych[3]

[1]*Laboratory No. 57, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*

[2]*Laboratory No. 11, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*

[3]*Department of Applied Mathematics, Faculty of Fundamental Sciences,
Bauman Moscow State Technical University, Moscow, Russia*

[4]*Machine Learning Track, Technopark VK Education, Moscow, Russia*

Abstract: In this study, the successful implementation of an active learning algorithm on small-scale datasets is demonstrated. The study also examines the dynamics of public opinions on COVID-19 vaccinations using VK (social network) commentaries related to the COVID-19 vaccine and masks for opinion evaluation. The proposed methodology includes several stages such as natural language processing, classification with active learning, exploratory data analysis, and opinion dynamics. Natural language processing is used for text preprocessing, tokenization, and feature extraction. A machine learning model with active learning is employed to identify opinions as positive, negative, or neutral/unknown. The model includes classical machine learning, machine learning and deep learning models. The results show that the highest classification accuracy is 69.1% and 73.1% without and with the active learning algorithm, respectively. The experimental results suggest that classifiers using active learning perform better than simple natural language processing classifiers on small-scale datasets.

Keywords: active learning, deep neural network, opinion dynamics, opinion analysis, BERT, COVID-19 VK opinion classification

1. INTRODUCTION

The COVID-19 pandemic is a deadly virus that has affected many countries around the world, leading the World Health Organization to declare it a pandemic disease in March 2020 [40]. The virus has caused a great loss of life, and governments worldwide have employed various measures to combat it, including travel restrictions, vaccination, and facility closures. The COVID-19 vaccination is one of the most significant techniques applied by governments to control its spread, and many countries have reported fewer COVID cases due to a high percentage of vaccination. The Russian government has launched a widespread immunization effort, and the vaccine is now available to the general population after several stages of vaccine trials.

The success of any immunization campaign depends on its public approval rate and the speed of acceptance. However, there are many misconceptions and doubts about COVID-19 vaccinations among ordinary people. Therefore, it is necessary to understand public opinion dynamics to ensure an effective immunization company. This study aims to develop

[*] Corresponding author: vlamelni@gmail.com

an algorithm that can identify opinions and comprehend opinion dynamics from small-scale datasets with sufficient accuracy. Specifically, the study focuses on the opinion analysis of VK messages on the Russian language regarding the COVID-19 vaccine and the identification of their opinions as positive, negative, or neutral/unknown. The results of this work can comprehend in improving machine learning (ML) models via active learning and in formulating policies to address people's queries before mass vaccination [38]. Active learning in machine learning involves an interactive process where a learning algorithm seeks feedback from a user or another information source to assign desired outputs to new data points, which helps the algorithm improve its efficiency.

The article is organized as follows: Sections 2, 3, and 4 discuss opinion dynamics, specifically the dynamics of opinions on COVID-19. Section 5 and 6 detail the BERT-based algorithm and active learning, BERT is Bidirectional Encoder Representations from Transformers. Section 7 presents the experimental results, illustrations, and discussions, followed by the conclusions in Section 8.

1.1. Novelty of the proposed work

The current research presents the practical implementation of an algorithm for opinion analysis of Russian people's opinions about COVID-19 vaccines based on VK (social network) messages [13, 23] (see also the study on the dynamics of opinions regarding the wearing of medical masks [18, 19]). The methodology proposed consists of four main steps: natural language processing (NLP), opinion classification using active learning, exploratory data analysis (EDA), and analysis of opinion dynamics. The NLP step involves data preprocessing, including data cleaning, removal of irrelevant words, data normalization, and tokenization [14,20]. Further, preprocessed data is used as input in BERT-based model, which is trained using active learning after feature extraction and data labeling. The BERT-based model is utilized for opinion classification, with positive, negative, and neutral/unknown being the three categories. The EDA and opinion dynamics are conducted on a labeled dataset of two million using the BERT-based model with active learning. To conduct this study and build the BERT-based model, five machine learning techniques were utilized, such as logistic regression, gradient boosting, extreme gradient boosting (XGB), transformers, and active learning [10, 15]. The application of active learning for small datasets in opinion dynamics tasks was analyzed in this study.

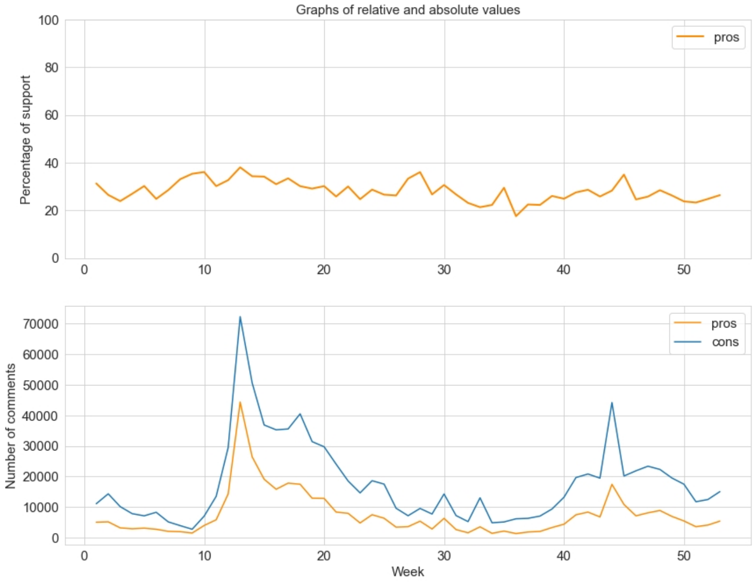


Fig. 1.1. The charts of relative and absolute values of ”for” and ”against” opinions

2. PRELIMINARY DATA ANALYSIS FOR OPINION DYNAMICS

In this work, data collected from the VK social network is used, and approximately 4GB of text data are used, containing users' commentaries, their identifiers, time, and source, user IDs and those with whom they interacted, information about likes on commentaries and posts. The data was collected from popular news communities from January 19, 2020, to January 19, 2021, reflecting the beginning of the COVID-19 pandemic and covering the first two waves of infections.

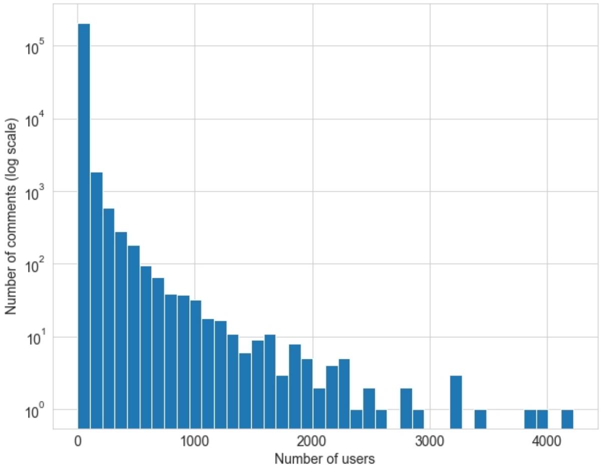


Fig. 2.2. A graph of the density distribution of commenting frequency for all users

First, each table is cleaned of junk data. For the commentaries tables, these were rows without the text of the commentary itself, and then they were checked for duplicates. For Table 7.3, uniqueness was checked by *h_id_comment*, but during encryption, some commentaries received the same *h_id_comment*, so it was necessary to change this hash to avoid throwing away false duplicates. As a result, after cleaning, 2.2 million records remained.

The first thing to look at is the opinions over the entire period, using Table 7.3 as a basis, which was labeled by the BERT-based model with an accuracy of 73.10%, trained on labeled samples. This will be described in the article below.

First, relevant commentaries were taken from a large dictionary, and now the table has 400 000 records. Then, records with likes of these comments from Table 7.5 were added to this table of comments. The opinion of the liker is considered being like the opinion of the commentary on which the like was placed. As a result, the charts of relative and absolute values of ”for” and ”against” opinions are presented in Figure 1.1. The analysis was conducted without considering comments and likes with the opinion ”unknown = 0”.

According to the relative graph, we can see that the average ”for” (and ”against” as well) opinion remains unchanged and hovers around 30%. Peaks of activity are clearly visible on the absolute value graph. Comparing this graph with the 2020 disease incidence, these maximums coincide with the active phase of the first two waves of disease incidence growth. The plateau after the maximums coincides with the peak of disease incidence. The first wave was the most frightening for people, so the first peak is noticeably higher than the second, although in terms of infections, it was inferior to almost all the others, which indicates a trend towards a decrease in interest in the pandemic.

2.1. User commentaries distribution

The distribution of user comments shows that 211 000 unique users wrote 2.2 million commentaries throughout the year, half of whom wrote only one commentary. This is a bad sign for us because we cannot analyze their opinions over different periods of time, but we still need to monitor their influence on others. 90% of users wrote fewer than 15 comments in a year. A graph of the density distribution of commenting frequency for all users (on a logarithmic scale) is presented in Figure 2.2. We were able to verify that this is a power-law distribution with a heavy tail, $f(x) = 0.1357(\frac{x}{4222})^{-(0.1357+1)}$.

2.2. Distributions by groups and posts

Figure 2.3 shows a histogram for communities with over 1000 comments (20 in total). ”For” opinions range from 26–40% (Table 7.4 is used).

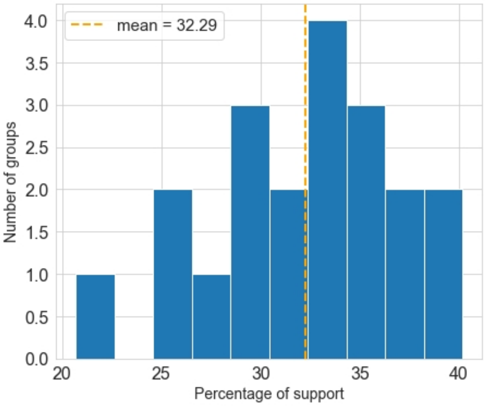


Fig. 2.3. Histogram for communities with over 1000 comments

Communities with polar opinions (names are hashed): *417*...: 26% support (20% if likes are considered), *366*...: 40% support (by comments only), f1d...: 40% support (by comments and likes). Next, we will analyze opinions on posts from the selected communities (20 out of 62) in Figures 2.4 and 2.5. Half of the posts have only a few commentaries, so it was decided not to consider posts with less than 10 commentaries to avoid getting anomalous average opinions ”for” and ”against”.

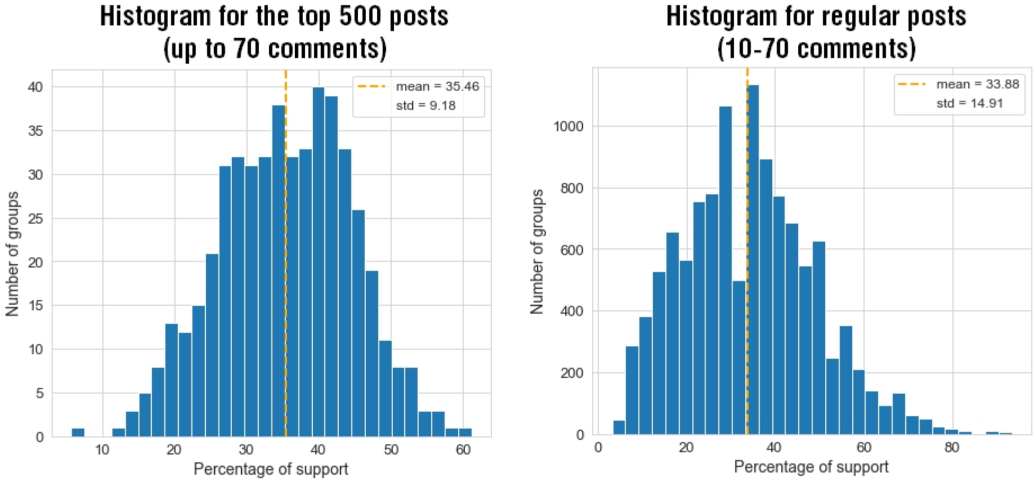


Fig. 2.4. Histograms without considering likes

In the first histogram, the density distribution is shown for very active posts with up to 70 commentaries, compared to the second histogram, which only includes posts with average activity. It is evident that the second graph is more biased towards lower support for COVID measures. When taking into account the likes, historgrams changes (Fig. 2.5).

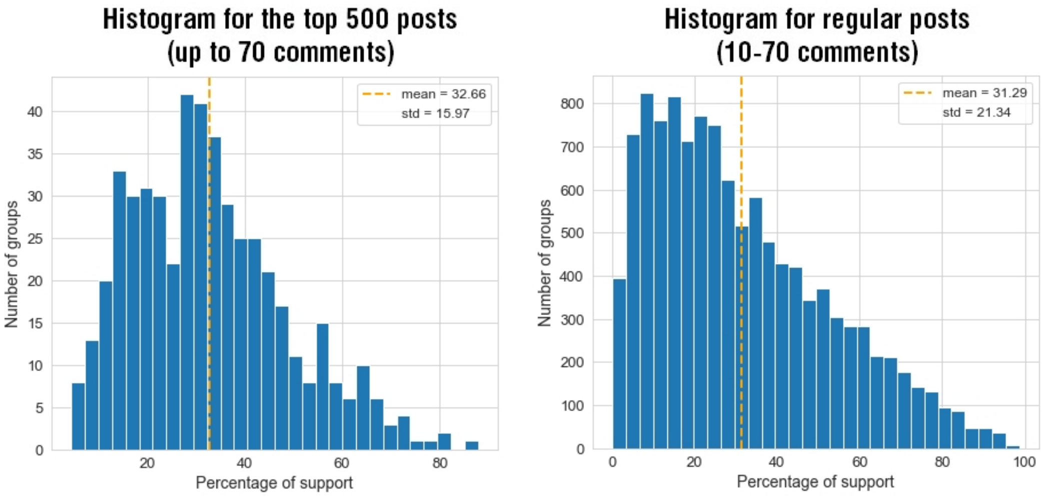


Fig. 2.5. Histograms with consideration of likes and commentaries

On the histogram for regular activity posts, it can be seen that support for polarizing opinions, especially negative ones, has increased [17]. This means that provocative comments are more often written in a negative tone. Posts with polarizing opinions are presented in Table 2.1.

Table 2.1. Posts with polarizing opinions

label	h_post_id	-1	0	1	pros
379	8bd...	732	187	34	4.43
419	d0d...	129	59	6	4.44
...	...	...	...	...	...
63	212...	53	52	240	81.91
210	6c0...	96	77	691	87.80

Examples of posts with negative reviews contain the following news:

1. It’s now almost official: Russia has the highest number of COVID-19 deaths per capita. This results from the actions of the authorities in the summer of 2020.
2. The last tests for a new vaccine against COVID-19 are underway in Russia. The vaccine will be offered first to doctors and pensioners, and then to everyone else.
3. The authorities plan to issue vaccination passports through the government services website.
4. ”Tsargrad” was the first media outlet to report that Sberbank’s subsidiary company, ”Immunotechnology”, which was created only in May of this year (it has no production or logistics facilities of its own), could become the sole supplier of vaccines to prevent COVID-19.

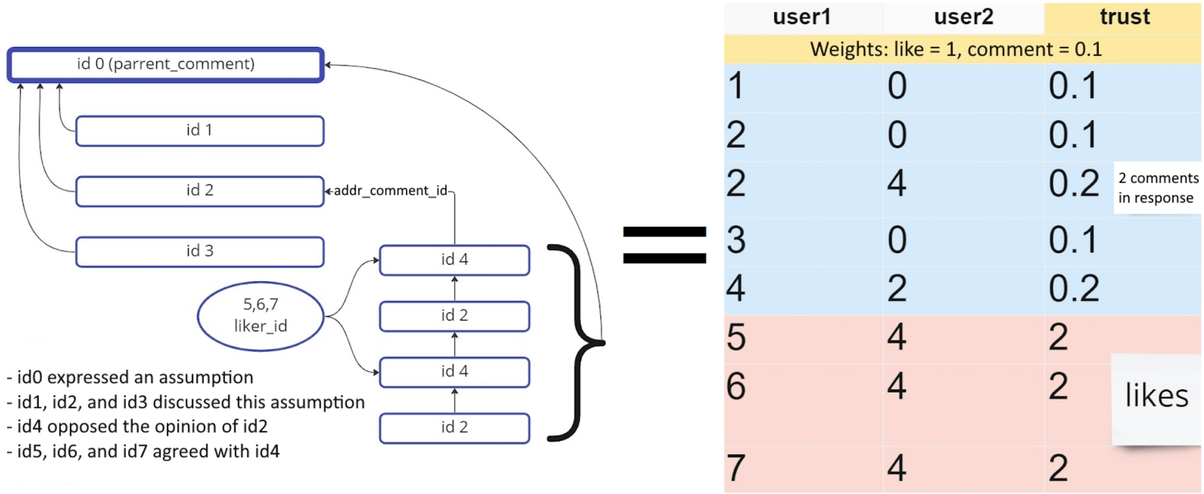
Examples of descriptions of posts with positive user ratings:

1. Mishustin contracted COVID-19.
2. According to most doctors, COVID-19 spreads fairly quickly but is not one of the most dangerous diseases - the fatality rate is 3.6%.
3. The number of COVID-19 deaths in Italy has set a daily record.
4. The Ministry of Internal Affairs named the regions with the highest number of violations of the isolation regime.

5. For the first time in Russia, over 15 000 new cases of COVID-19 were detected in one day.

3. IDENTIFYING A NETWORK ENVIRONMENT OF USERS

We need to determine the influence/trust of users (each to each) and predict their last opinions based on the cleaned data. Then, we can compare the results with the true and random outcomes. Figure 3.6 illustrates the structure of commentary threads and how we record interactions in a table.


id 0 (parrent_comment)
id 1
id 2
addr_comment_id
id 3
id 4
5,6,7
liker_id
id 2
id 4
id 2
- id0 expressed an assumption
- id1, id2, and id3 discussed this assumption
- id4 opposed the opinion of id2
- id5, id6, and id7 agreed with id4
=

user1	user2	trust	
Weights: like = 1, comment = 0.1			
1	0	0.1	
2	0	0.1	
2	4	0.2	2 comments in response
3	0	0.1	
4	2	0.2	
5	4	2	
6	4	2	likes
7	4	2	

Fig. 3.6. Method for obtaining the influence table from a commentary thread

The influence/trust table is compiled in several stages:counting comments within threads using (example: interaction between id2 and id4), counting replies to a parent commentary (example: id1, id2, id3), counting likes on comments (example: id5, id6, id7). The *trust* is calculated using the following formula [16]:

$$trust = w_{1k}c_{ij} + w_{2k}l_{ij} \tag{3.1}$$

where w_1 and w_2 are the weights assigned to comments and likes, c_{ij} and l_{ij} denote the number of comments and likes of agent $i \in N$ towards agent $j \in N$, respectively. The initial weights are set as $w1 = 0.1$ and $w2 = 1$. From the interaction table, a directed graph is constructed and the adjacency matrix is calculated using the *NetworkX* library. Let's inspect the interaction graph.

3.1. Visualization and characterization of the interaction graph

The visualization is created using the *Gephi* program with the *ForceAtlas2* layout. The colors in Figure 3.7 are determined by the modularity parameter, which measures the structure of the graph and indicates how well it is divided into dense communities. The modularity score ranges from –1 to 1, and in our case, the results for each time interval are around 0.7, indicating a clear division into communities.

When calculating modularity in *Gephy*, the algorithm groups nodes that are more densely connected to each other than to other nodes in the graph into the same community or cluster. This grouping leads to the division of the graph into subgraphs or clusters, each of which has numerous internal connections and relatively few external connections to other clusters.

In the graph, 10 clusters are clearly distinguished, which have been identified as subscribers to different VK communities. Some users are subscribed to several similar communities, while others may have a different position and gather in smaller isolated groups. The ”asteroid belt” around the clusters represents non-communicative users with at least one interaction.

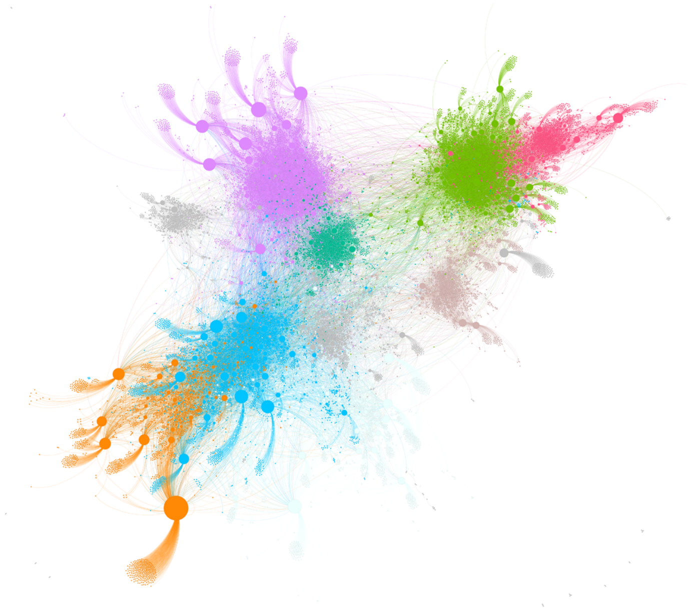


Fig. 3.7. The graph represents the interactions during the first week

The large nodes represent opinion leaders whose statements are actively discussed or liked, while the size of the nodes is proportional to their degree, i.e., the number of connections. The small branches of isolated user islands from opinion leaders represent likers. It is possible that these likes were artificially inflated to promote a particular opinion, but more likely, these are users who do not express their own opinions but only agree with others.

Another informative parameter is the degree of assortativity in the graph, which represents the preference of network nodes to connect to other nodes that are in some way similar to them. In our case, the assortativity parameter for each week was –0.1, indicating a negative correlation between nodes of different degrees.

Positive assortativity values indicate a correlation between nodes of similar degrees, while negative values indicate relationships between nodes of different degrees. When examining graphs over larger time intervals than current ones, the grouping in the center of the graph is blurred due to the fact that many users show activity in the same communities.

4. THE DYNAMICS OF OPINIONS ON COVID-19 TASK

In this part of the work describes the creation of a trust matrix from an interaction graph, a model, and an initial prediction. Formula (4.2) will be used for further prediction [16]:

$$b' = b * A \tag{4.2}$$

where b is a vector of initial average user opinions for the selected interval, which is multiplied by the trust matrix A to obtain the prediction b'. The graph described above is read as an adjacency matrix, which will be referred to as the trust/influence matrix ”each with each”. This is a sparse matrix, whose elements are taken as the trust value, considering the quantity and weights of interactions from (3.1).

The next step is to modify the matrix: add a constant term w_3 to the formula for matrix elements (3.1), add self-trust to the main diagonal with an initial weight of $w_4 = 0.5$, ensure stochasticity by making sure that $\sum_{j=1}^{n} a_{ij} = 1$ for each row, where users trust themselves with weight w_4 and everyone else with weight $1 - w_4$.

The resulting dynamics of the opinions of agent $i \in N$ (N - the set of users) is given by the equation [16]:

$$x_i^{'} = w_4 x_i + (1 - w_4) \sum_{j \in N \setminus \{i\}} a_{ij} x_j \tag{4.3}$$

where $a_{ij} \in [0; 1]$ represents the degree of trust that agent i has in agent j ($i, j \in N$).

The trust matrix A is constructed as follows. First, we form a matrix A', where the elements are given by [16, 17]:

$$a_{ij}^{'} = \frac{w_1 c_{ij} + w_2 l_{ij} + w_3}{\sum_{j \in N \setminus \{i\}} (w_1 c_{ij} + w_2 l_{ij} + w_3)} \tag{4.4}$$

The denominator serves as a normalization factor to ensure stochasticity. We then account for self-trust ($w_4 \in [0; 1]$) [16]:

$$A = w_4 E + (1 - w_4) A' \tag{4.5}$$

Further model complexity is possible but is not practical due to incomplete information, as shown by optimizing weights using *scipy*'s Powell method to minimize the Jensen-Shannon divergence. Starting from the initial weights described above, the optimizer converged to the following weights: $w_1 = 7.88$, $w_2 = 5.48$, $w_3 = 6.25$, $w_4 = 0.99$.

The most significant result is the near-unity self-trust weight, indicating that each user's opinion is influenced only slightly by others. People are more likely to change their opinions by exploring news sources and discussing offline rather than engaging in online communication.

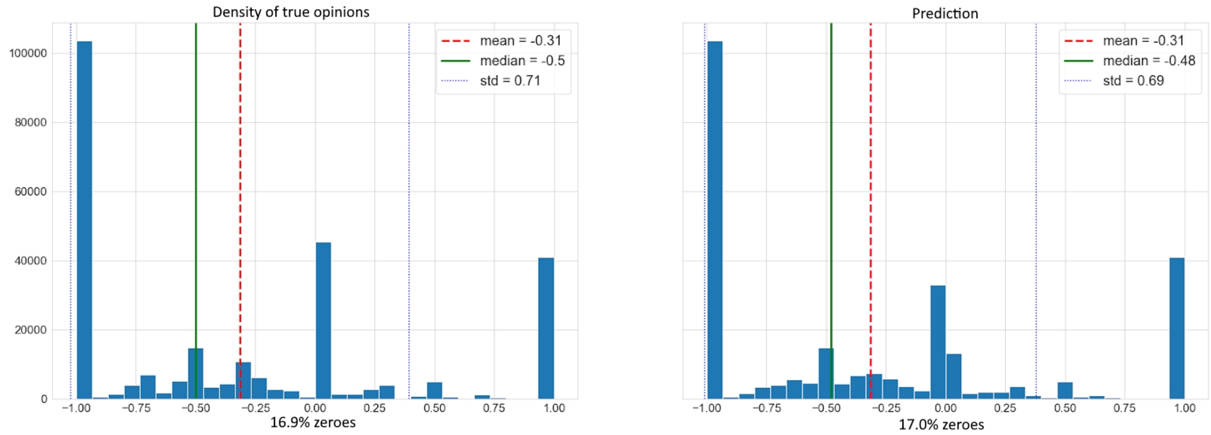


Fig. 4.8. Prediction of opinions on COVID-19

However, the optimized weights allow us to predict the percentage of negative opinions about vaccines, which we present in Figure 4.8. For the known period, the percentage is 61.85%, and for the forecast period, it is 73.72%.

5. IDENTIFYING OPINIONS ON VACCINATION

The proposed methodology in this study for opinion analysis, classification and opinion dynamics involves several steps:

1. Data gathering
2. Data preprocessing
3. Data labelling
4. Opinion analysis
5. Opinion classification using a BERT-based model without active learning
6. Opinion classification using the BERT-based model with active learning
7. Opinion dynamics research

The flowchart, encoder and decoder of the BERT-based model, along with the active learning algorithm, are shown in Fig. 5.9, Fig. 5.10, and Algorithm 1, respectively. The details of each step of the proposed methodology are described in subsequent subsections.

5.1. Data gathering

In an era marked by unprecedented challenges, social media platforms have emerged as crucial instruments for understanding public sentiment and opinion [1, 2, 6]. Among these platforms, VK, a popular social network, offers a rich dataset for analyzing public views on pressing issues such as COVID-19 vaccinations. This study leverages data from VK to conduct a comprehensive opinion analysis on the subject [3, 5].

5.1.1. Data source and origins The data for this study was sourced from VK, specifically from popular news communities. A total of 2.2 million commentaries related to ”COVID-19 vaccination” were collected using VK's API (Application Programming Interface) [24]. These commentaries were part of various discussions, posts, and threads within these communities. Importantly, the data spans a period from January 19, 2020, to January 19, 2021. This timeframe captures the beginning of the COVID-19 pandemic and includes the first two waves of infections, thereby providing a dynamic view of evolving public opinion.

5.1.2. Data anonymization and ethics To adhere to privacy and ethical standards, all collected commentaries were anonymized prior to preprocessing. Identifiable information such as usernames, profile pictures, and any other personal identifiers were removed to maintain the anonymity of the individuals involved.

5.1.3. Relevance to the study The VK dataset is particularly relevant to this study for several reasons:

1. Volume: The large number of commentaries provides a robust sample for analysis.
2. Diversity: The dataset includes opinions from individuals of various demographics, offering a more comprehensive view of public sentiment.
3. Timeliness: The data covers a critical period in the COVID-19 pandemic, making it highly current and relevant.

By leveraging this dataset, the study aims to offer a nuanced understanding of public opinions on COVID-19 vaccinations, thereby providing valuable insights for policymakers, healthcare providers, and the general public.

5.2. Data preprocessing

In this chapter, the data preprocessing process is explained, which includes four sub-steps: data cleaning, normalization of data, tokenization of commentaries, and vectorization. As the collected commentaries are noisy and unlabelled, only 7 000 commentaries out of 2.2

million were labelled manually by three linguistics and medical science experts. The data cleaning process is applied to eliminate noises such as unnecessary words, emojis, additional characters in sentences, blank spaces, stopping words, and punctuation. To normalize the data, stemming and lemmatization techniques are used from the WordNetLemmatize package of natural language tool kit (nltk) for logistic regression (LR), gradient boosting, and XGB models. The stemming technique is used to determine the root form by removing terminators from words, while lemmatization groups similar words in various forms to reduce dimensionality. Finally, the normalized dataset is tokenized by dividing the text into tokens and used as features for training, validation, and test datasets.

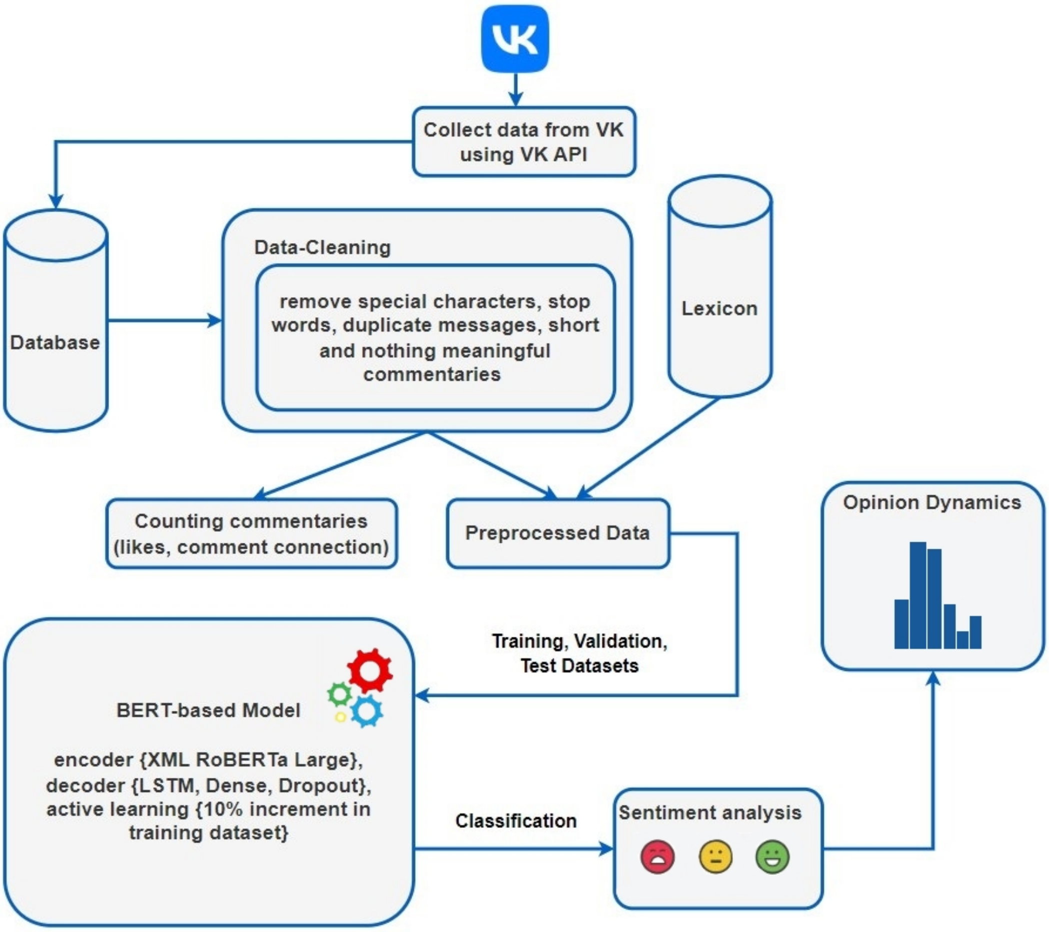

Collect data from VK using VK API
Database
Data-Cleaning
remove special characters, stop words, duplicate messages, short and nothing meaningful commentaries
Lexicon
Counting commentaries (likes, comment connection)
Preprocessed Data
Opinion Dynamics
Training, Validation, Test Datasets
BERT-based Model
encoder {XML RoBERTa Large}, decoder {LSTM, Dense, Dropout}, active learning {10% increment in training dataset}
Classification
Sentiment analysis

Fig. 5.9. Flowchart of the proposed research

5.3. *Data labelling*

Three experts in linguistics and medical science manually labelled a training dataset of sentences as negative, positive, or neutral/unknown, represented by the values 0, 1, and 2, respectively. The text data was then transformed into a vectorized form using the TF-IDF (term frequency–inverse document frequency) vectorization technique for LR. This technique extracts features from labelled data based on the frequency of words in the text.

The phrase frequency (TF) measures the frequency of a word in a document, while the inverse frequency of words (IDF) measures the frequency of a word in the document set:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}} \tag{5.6}$$

where f_t is the count of a term t in a commentary d.

$$IDF(t, D) = log\frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{5.7}$$

where D is infering to our document space (training dataset), $D = d_1, d_2, \ldots, d_n$, n is the number of commentaries in D, $|D| = n$. The denominator $|\{d \in D : t \in d\}|$ represents the total count of occurrences of the term t appeared in a commentary d.

The $TF - IDF$ score is calculated by multiplying the TF score with the IDF score:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \tag{5.8}$$

5.4. Opinion analysis

Different machine learning models including Logistic Regression, Gradient Boosting, Extreme Gradient Boosting (XGB), BERT models (transformers) and active learning are applied for opinion classification into positive, neutral, and negative/unknown categories. A brief description of each machine learning technique is provided in the following subsections.

5.4.1. Logistic Regression Machine learning applies a logistic regression classifier, a statistical model that establishes the relationship between independent and dependent variables [27]. The classifier uses a logistical function to determine the input, set of weighted functions, and the correlation between classes. Proper selection of features can improve the model's accuracy and generalizability. Feature vector i is classified as positive, neutral/unknown, or negative using a mathematical expression represented as:

$$S(f = 1|i) = l(i) = \frac{1}{1 + h^{zw_i}} \tag{5.9}$$

where, z means the feature weight, S is the probability of commentary i which belongs to class f.

5.4.2. Gradient Boosting In machine learning, the gradient boosting technique involves the combination of numerous weak models to form a robust predictive model that is suitable for categorizing extensive datasets [9]. This model is capable of reducing bias error, thus resulting in an accurate and efficient prediction model. By creating an approximation $\widehat{H}(b)$ of the function $H^*(b)$, the gradient boosting model maps the input instances b to their output values z. This function approximation $H^*(b)$ can be expressed as a weighted sum of functions, as shown by the mathematical expression:

$$H_c(b) = H_{b-1}(b) + p_c q_c(b) \tag{5.10}$$

where, p_c means the weight of the c-th function $q_c(b)$.

5.4.3. Extreme Gradient Boosting XGB is a set of gradient boosting techniques designed for current data science challenges [11, 12]. It uses an ensemble model of classification and regression tree sets (CART). XGB is known for its high scalability, parallelizability, speed, and regularized approach to control over-fitting. The mathematical representation of the XGB model is expressed as:

$$\widehat{A}_l = \sum_{p=1}^{P} q_p(m_l), q_p \in Q \tag{5.11}$$

where q_p for the p-th tree denotes a function in functional space Q, and P represents the total number of trees. The set of all possible CARTs is represented by Q.

5.5. *BERT-based model*

Applying an active learning algorithm to the BERT-based model can increase its accuracy for training and testing [7, 10, 15, 29]. By cyclically adding labeled data to the training dataset, the active learning algorithm can improve the model's predictive accuracy beyond that of a single BERT-based model.

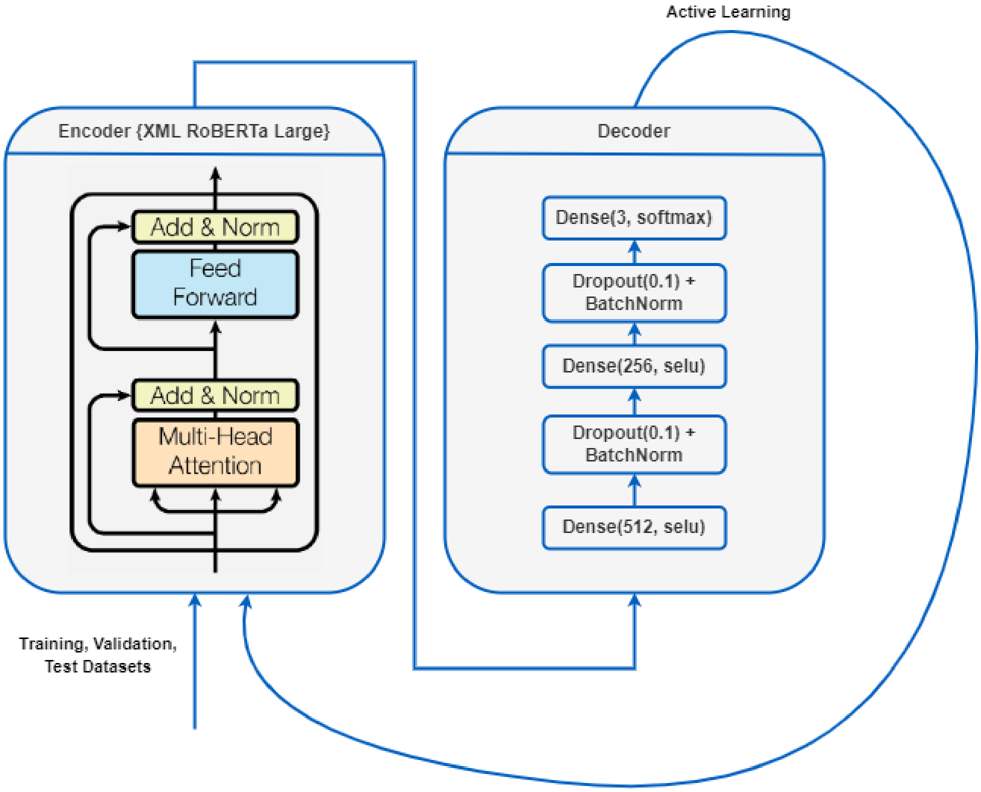
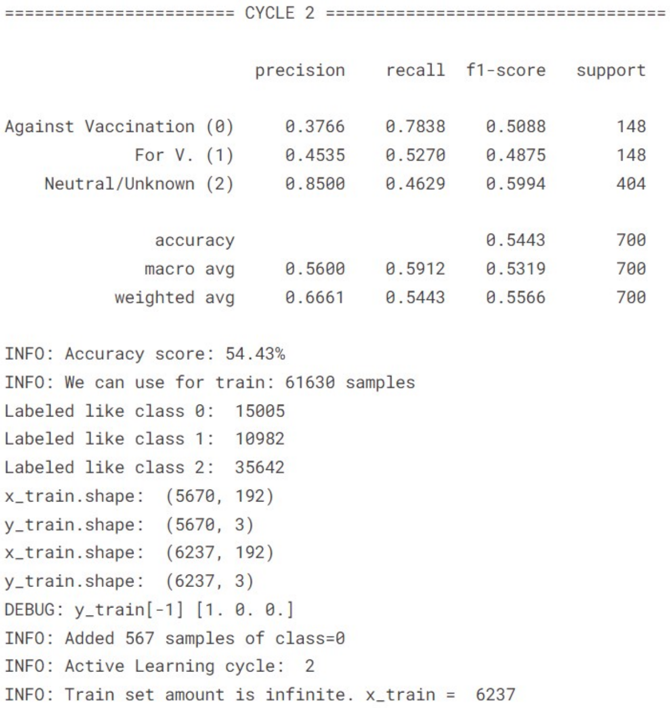


Fig. 5.10. BERT-based model with active learning cycle.

Figure 5.10 shows a block diagram of the BERT-based model with active learning. The active learning cycle includes the Pre-Training phase and the Cycle phase. In the Pre-Training phase, the model is trained on a corpus of data with the original labeling. Then, in the Cycle phase, additional data is labeled and added to the training dataset. To create the BERT-based model, a range of machine learning algorithms were evaluated, including logistic regression, gradient boosting, extreme gradient boosting, and multiple BERT models, such as XLM RoBERTa Large, RoBERTa, ruRoBERTa, BERT large uncased, ruBERT, GPT (Generative Pre-trained Transformer), and DistillBERT (Distilled Bidirectional Encoder Representations from Transformers) [25, 35]. XLM-RoBERTa stands for "Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach," combining multilingual capabilities with optimized pretraining for robust performance. The "ru" prefix indicates that the model has been trained on a Russian-language data corpus. The BERT-based model constructed using XLM RoBERTa Large and Active Learning is shown in Figure 5.10 [36,39]. The BERT-based model operates as follows:

1. The dataset is split into three sets: training, validation, and testing.
2. The BERT-based model is trained on the trainset and validated on the validation set.
3. The accuracy of the model is evaluated on the testing dataset.
4. The resulting data labelled by The Active Learning Algorithm is added to the trainset.
5. Steps 1–4 are repeated until the accuracy stabilizes.

6. The final accuracy measurement is taken on the testing dataset.
7. 2.2 million commentaries are labeled by the trained model.



```
====================== CYCLE 2 =================================

                         precision    recall  f1-score   support

Against Vaccination (0)     0.3766    0.7838    0.5088       148
             For V. (1)     0.4535    0.5270    0.4875       148
    Neutral/Unknown (2)     0.8500    0.4629    0.5994       404

               accuracy                         0.5443       700
              macro avg     0.5600    0.5912    0.5319       700
           weighted avg     0.6661    0.5443    0.5566       700

INFO: Accuracy score: 54.43%
INFO: We can use for train: 61630 samples
Labeled like class 0:  15005
Labeled like class 1:  10982
Labeled like class 2:  35642
x_train.shape:  (5670, 192)
y_train.shape:  (5670, 3)
x_train.shape:  (6237, 192)
y_train.shape:  (6237, 3)
DEBUG: y_train[-1] [1. 0. 0.]
INFO: Added 567 samples of class=0
INFO: Active Learning cycle:  2
INFO: Train set amount is infinite. x_train =  6237
```

Fig. 5.11. Second Active Learning cycle

6. BERT-BASED CLASSIFICATION ALGORITHM WITH ACTIVE LEARNING

Algorithm 1. Active Learning A concept of active learning algorithm was implemented without a specific template [38]. The algorithm consists of two parts: Pretrain and Cycle. In the Pretrain part, the model is trained on a corpus of data with initial labeling. The corpus in this task consists of 7 000 instances, with 5 670 for training, 630 for validation, and 700 for testing. Pretrain can be considered as a well-trained version of the model for use in reference labeling.

Next, the Cycle part involves repeating a set of specific actions. An unlabeled corpus of commentaries is taken, which in this case is 63 000 instances, and labeled with the model trained in the Pretrain part (or in the previous step of the Cycle part). Then the recognition accuracy of each class is calculated from the test corpus (700 instances), and the class with the lowest recognition accuracy is selected. The reason for this is our objective to achieve average accuracy over all classes rather than focusing on accurate prediction of a single class [38]. Various methods of adding instances to the training corpus are available, such as fixed selection, percentage selection, taking a percentage of the current selection volume, and percentage selection with accuracy (taking only those selections that exceed the classification accuracy threshold). A comparison of the methods is presented in Table 6.2.

In the next step, only the necessary number of added elements for the class with the lowest accuracy is taken from the labeled corpus of 63 000 instances. Before this, our predictions are sorted in descending order of confidence in predictions by classes, and the necessary number of initial elements are taken to add to the training corpus. Also, the instances already taken are marked to avoid taking them again.

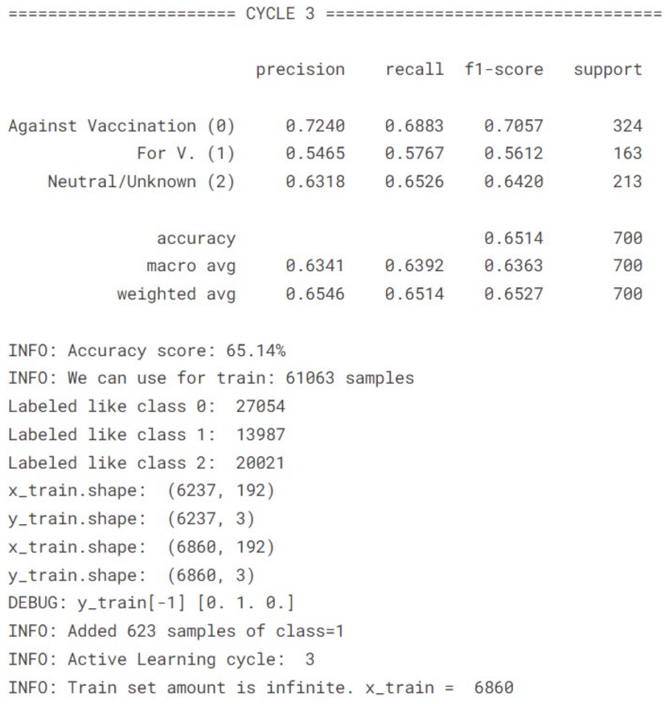
Table 6.2. Performance evaluation of different methods for augmenting the training corpus

Way of adding instances	Average precision score
percentage of the current selection volume	73.10
percentage selection with accuracy	73.02
percentage selection	72.78
fixed selection	72.15

If the number of instances is not sufficient, those available are taken, or none are taken, and the algorithm retrains the network to improve accuracy, giving the opportunity to take the necessary number of elements in the next iteration. While it is possible to retrain the network in each iteration of the Cycle part, it should be understood that this may lead to overfitting, which can result in worse network training outcomes.

Testing with training from scratch and continuous retraining of the model in the Cycle part showed that the model trained in the continuous retraining mode performed better despite overfitting [26]. It is recommended to create an additional test set for final evaluation at the end of the process to exclude any implicit overfitting. Active learning improves network accuracy by an average of 3–4% in this task. The Pretrain part is essential, and it is better to choose a larger number of epochs for training the model in this part, as the most accurate initial labeling has a greater impact than accurate labeling in the later iterations of the Cycle part (the accuracy is higher by 0.5% on average) [23].

At the end of the Cycle iteration, accuracy is calculated on the test corpus, and a training corpus is created with the added set of instances determined to be necessary for adding. The next iteration of the Cycle part starts again on the updated training dataset. The Active Learning algorithm's operation is presented in Fig. 5.11 and Fig. 6.12.



```
======================= CYCLE 3 ==================================

                          precision    recall  f1-score   support

Against Vaccination (0)      0.7240    0.6883    0.7057       324
             For V. (1)      0.5465    0.5767    0.5612       163
    Neutral/Unknown (2)      0.6318    0.6526    0.6420       213

               accuracy                          0.6514       700
              macro avg      0.6341    0.6392    0.6363       700
           weighted avg      0.6546    0.6514    0.6527       700

INFO: Accuracy score: 65.14%
INFO: We can use for train: 61063 samples
Labeled like class 0:  27054
Labeled like class 1:  13987
Labeled like class 2:  20021
x_train.shape:  (6237, 192)
y_train.shape:  (6237, 3)
x_train.shape:  (6860, 192)
y_train.shape:  (6860, 3)
DEBUG: y_train[-1] [0. 1. 0.]
INFO: Added 623 samples of class=1
INFO: Active Learning cycle:  3
INFO: Train set amount is infinite. x_train =  6860
```

Fig. 6.12. Third Active Learning cycle.

7. EXPERIMENTAL RESULTS AND PRACTICAL APPLICATION

This section covers the results of opinion analysis using natural language processing and the performance evaluation of machine learning techniques applied. The work collected a total of 2.4 million commentaries from Russian people in 2020 and 2021 for opinion analysis on vaccines using the VK API. Fifteen fields were extracted such as *post source id, post id, comment id, user id, parent comment id, liker id, user subscriptions, date, text, comment label*, which were shown in Tables 7.3–7.6.

Table 7.3. Unlabelled commentaries

h_owner_id	h_post_id	h_comment_id	h_commenter_id	...	h_addr_comment_id	date	text
3a4...	d8b...	9d4...	d41...	...	NaN	158...	uzhas [horror]

Table 7.4. Likes on posts and user commentaries

h_owner_id	h_post_id	h_comment_id	true_label	text
3a4...	d8b...	9d4...	0.0	Nuzhno lechit' koronovirus [Cure COVID]

Table 7.5. Information about posts

h_owner_id	h_post_id	h_comment_id	true_label	text
3a4...	d8b...	9d4...	0.0	Tol'ko 6% rossiyan... [Only 6% of Russians...]

Table 7.6. Commentaries on vaccination

h_owner_id	h_post_id	h_comment_id	true_label	text
3a4...	d8b...	9d4...	0.0	A gde budut vakcinirovat'... [Where will they vaccinate...]
0ef...	edc...	9b4...	1.0	Spasibo nashim uchyonym [Thanks to our scientists]
92e...	fe4...	576...	NaN	Oni vrut [They lie]
2e1...	e43...	76f...	2.0	Smekh,vakcina pomozhet [Jokes aside, the vaccine will help]

Table 7.7. Distribution of opinions in commentaries after data labelling

Total number of commentaries	Positive	Negative	Neutral/Unknown
2,285,620	911,269 (39.87%)	421,580 (18.44%)	952,771 (41.69%)

The trained classifier was used to label the data, and the resulting labels are added as an additional column *label* to Table 7.3. The distribution of opinion after labelling are shown in Table 7.7. The majority of input commentaries were analyzed as neutral/unknown positive and followed by negative commentaries. The results indicate that 41.69% of input commentaries were classified as neutral/unknown, 39.87% as positive, and 18.44% as negative. Inference (inf.) requires on average three times fewer RAM resources.

Table 7.8. Performance evaluation and resource utilization of different classifiers

Model	Precision score	Training time	RAM/HDD (train)	Inf. time	RAM/HDD (inf.)
BERT-based & Active Learning	73.10	2 hour(s)	15/32 GB	0.50 sec	5/8 GB
BERT-based Classifier	69.17	0.4 hour(s)	15/32 GB	0.50 sec	5/8 GB
Logistic Regression	56.14	0.1 hour(s)	5/32 GB	0.25 sec	3/8 GB
XGB	51.83	0.1 hour(s)	15/32 GB	0.25 sec	5/8 GB
Gradient Boosting	48.21	0.1 hour(s)	15/32 GB	0.25 sec	5/8 GB

7.1. Active learning

The precision score of different classifiers and BERT-based models were presented in Tables 7.8 and 7.9, with the BERT-based model first trained in the pretrain part of active learning followed by the cycle part. The study found that active learning increased accuracy by 3–4% on small datasets on average. The parameters values initialized for the BERT-based model and active learning are presented in source notebook on platform Kaggle [23].

Table 7.9. Performance evaluation and resource utilization of different classifiers

Model	Precision score	Training time
XLM RoBERTa Large with Active Learning	73.10	2 hrs
XLM RoBERTa Large	69.17	0.4 hrs
ruRoBERTa	65.23	0.1 hrs
RoBERTa	63.58	0.1 hrs
GPT	62.76	0.2 hrs
DistillBERT	57.33	0.2 hrs
ruBERT	52.40	0.1 hrs
BERT large uncased	41.72	0.1 hrs

8. CONCLUSION

This study analyzes the opinion of Russian individuals towards various COVID-19 vaccines. VK platform data was collected as a sample for analysis. The raw text messages received several natural language processing operations for text preprocessing. The opinion analysis classified 41.69%, 39.87%, and 18.44% of commentaries as neutral/unknown, positive, and negative, respectively. To classify the opinion, a proposed BERT-based model with active learning was used, achieving a precision of 73.10%. The results indicate that active learning can improve precision by 3–4% on small datasets. Future work will involve usage of Graph Neural Networks (GNN) and all available data, including user subscriptions [4, 28, 37].

The majority of people's opinions on social media regarding vaccinations and their effects are neutral or unknown. However, only 39.87% of people are optimistic, which is a worrying situation for policymakers and the government [8]. In order to make the vaccination program effective, the government must convince the majority of the population that the vaccine will have positive outcomes and consequences [30, 31]. Therefore, policymakers and the government should focus on reducing vaccine anxiety before starting mass vaccination.

According to recent research, although approximately 22 million Russians have been infected with COVID-19, a significant proportion of the population still believes that the pandemic has been exaggerated, based on our research [22, 32]. This has led to skepticism and opposition towards the COVID-19 vaccine, driven by concerns such as vaccine safety, skepticism towards pharmaceutical companies, and questions about the effectiveness of the vaccines available [33, 34]. Despite legitimate doubts, some unfounded conspiracy theories surrounding the virus have been dispelled.

The government, pharmaceutical companies, and Non-Governmental Organizations (NGOs) should make a significant effort to educate the general public about the vaccine program and the importance of returning to normal life. Special attention should be paid to addressing people's fears and misconceptions about the vaccine to encourage vaccine uptake.

Using various models such as Logistic regression, Gradient Boosting, XGB, and BERT-based models with active learning, the research categorizes people's feelings towards the vaccine as positive, neutral/unknown, or negative. The proposed BERT-based model achieved 73.10% accuracy, outperforming other classifiers in terms of classification accuracy [21].

An analysis was also conducted to identify the dynamics of opinions. The predicted opinions are moving towards the negative side, indicating that, without any new external factors, people's opinions about vaccines are deteriorating under the same news background.

ACKNOWLEDGEMENTS

The authors, Vladislav Melnichuk and Vladislav Sych, express their gratitude to their academic supervisors, Dmitry Gubanov and Alexander Chkhartishvili, for their support.

8.1. Funding Information

This research did not receive any specific grants from public, commercial, or non-profit funding agencies.

8.2. Data Availability Statement

The data can be viewed in the dataset uploaded to Kaggle [24].

REFERENCES

1. Alam K.N., Khan M.S., Dhruba A.R., Khan M.M., Al-Amri J.F., et. al. (2022). Deep learning-based sentiment analysis of COVID-19 vaccination responses from twitter data. *arXiv:* 2209.12604, [Online]. Available: https://arxiv.org/abs/2209.12604
2. Aygun I., Kaya B., & Kaya. M. (2022). Aspect based twitter sentiment analysis on vaccination and vaccine types in COVID-19 pandemic with deep learning. *IEEE Journal of Biomedical and Health Informatics*, **26**(5), 2362–2368.
3. Baker Q.B., Shatnawi F., Rawashdeh S., Al-Smadi M., & Jararweh Y. (2020). Detecting epidemic diseases using sentiment analysis of arabic tweets. *Journal of Universal Computer Science*, **26**(1), 50–70.
4. Best graph neural network architectures: Gcn, gat, mpnn and more (2022, September 23). *Theaisummer*. [Online]. Available: https://theaisummer.com/gnn-architectures/
5. Bonnevie E., Goldbarg J., Gallegos-Jeffrey A.K., Rosenberg S.D., Wartella E., et. al. (2020). Content themes and influential voices within vaccine opposition on twitter. *American Journal of Public Health*, **110**(3), S326—S330.
6. Brajawidagda U. & Chatfield A.T. (2012). Twitter tsunami early warning network: A social network analysis of twitter information flows. *ACIS*, **56**(1), 4-–7, https://aisel.aisnet.org/acis2012/56
7. Cambria E., Das D., Bandyopadhyay S., & Feraco A. (2017). Affective computing and sentiment analysis. *A practical guide to sentiment analysis* (pp. 103–105) Washington, D.C.: IEEE Intelligent Systems
8. Chakraborty K., Bhatia S., Bhattacharyya S., Platos J., Bag R., et. al. (2020). Sentiment analysis of COVID-19 tweets by deep learning classifiers. *Applied Soft Computing:* 106754, **97**(1), [Online]. Available: https://doi.org/10.1016/j.asoc.2020.106754
9. Cockett. J.R.B. (1987). Discrete decision theory: Manipulations. *Theoretical Computer Science*, **54**(1), 215-–236.
10. Devlin J., Chang M., Lee K., & Toutanova K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:* 1810.04805, [Online]. Available: https://arxiv.org/abs/1810.04805
11. Fayaz M., Khan A., Rahman J.U., Alharbi A., Uddin M.I., et. al. (2020). Ensemble machine learning model for classification of spam product reviews. *Complexity*, **2020**(1), 3–8.
12. Freund Y. & Schapire R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, **55**(1), 119–139.
13. Garcia K. & Berton L. (2021). Topic detection and sentiment analysis in twitter content related to COVID-19 from brazil and the usa. *Applied Soft Computing:* 107057, **101**(1), [Online]. Available: https://doi.org/10.1016/j.asoc.2020.107057

14. Ghosh M. & Sanyal G. (2019). Analysing sentiments based on multi feature combination with supervised learning. *International Journal of DataMining, Modelling and Management*, **11**(1), 391-–416.
15. Goyal N., Du J., Ott M., Anantharaman G., & Conneau A. (2021). Larger-scale transformers for multilingual masked language modeling. *arXiv:* 2105.00572, [Online]. Available: https://arxiv.org/abs/2105.00572
16. Gubanov, D.A., Novikov D.A., & Chkhartishvili A.G. (2019). Social Networks: Models of information influence, control and confrontation. *Springer International Publishing*, 158, Switzerland: Cham
17. Gubanov, D.A., Petrov I.V., & Chkhartishvili A.G. (2021). Multidimensional Model of Opinion Dynamics in Social Networks: Polarization Indices *Automation and Remote Control*, **82**(10), 1802-–1811.
18. Gubanov, D.A. & Novikov D.A. (2023). Models of Joint Dynamics of Opinions and Actions in Online Social Networks. Part I: Primary Data Analysis *Control Sciences*, **2**, 31–45.
19. Gubanov, D.A. & Novikov D.A. (2023). Models of Joint Dynamics of Opinions and Actions in Online Social Networks. Part II: Linear Models *Control Sciences*, **3**, 31–54.
20. Hou K., Hou T., & Cai L. (2021). Public attention about COVID-19 on social media: An investigation based on data mining and text analysis. *Personality and Individual Differences:* 110701, **175**(1).
21. Hung M., Lauren E., Hon E.S., Birmingham W.C., Xu J., et. al. (2020). Social network analysis of COVID-19 sentiments: Application of artificial intelligence. *J Med Internet Res:* e22590, **22**(1), [Online]. Available: http://www.jmir.org/2020/8/e22590/, https://doi.org/10.2196/22590
22. Interfax. (2022). *Coronavirus pandemic*. [Online]. Available: https://www.interfax.ru/chronicle/novyj-koronavirus-v-kitae.html
23. Kaggle source notebook with the current research (2022). *Opinion dynamics and bert COVID-19 assa*, [Online]. Available: https://www.kaggle.com/code/vladislavmelnichuk/opinion-dynamics-and-bert-COVID-19-assa
24. Kaggle source dataset (2022). *COVID-19 data vk commentaries*, [Online]. Available: https://www.kaggle.com/datasets/vladislavmelnichuk/COVID19-data-vk-comments
25. Kyriakides G. & Margaritis K. (2019) *Hands-On Ensemble Learning with Python: Build highly optimized ensemble machine learning models using scikit-learn and Keras*. Birmingham, United Kingdom: Packt Publishing, [Online]. Available: https://books.google.co.in/books?id=N4mkDwAAQBAJ
26. Lee H.D., Lee S., & Kang U. (2021). Auber: Automated bert regularization. *PLoS ONE:* e0253241, **16**(6), [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8238198/, https://doi.org/10.1371/journal.pone.0253241
27. Linear regression (machine learning) (2003). *University of Pittsburgh*. [Online]. Available: https://people.cs.pitt.edu/ milos/courses/cs2750-Spring03/lectures/class6.pdf
28. Liu J., Liu P., Zhu Z., Li X., & Xu G. (2021). Graph convolutional networks with bidirectional attention for aspect-based sentiment classification. *Applied Sciences*, **11**(4), 1528, 3–10.
29. Liu Y., Ott M., Goyal N., Du J., Joshi M., et. al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv:* 1907.11692, [Online]. Available: https://arxiv.org/abs/1907.11692
30. Malik A., McFadden S., Elharake J., & Omer S. (2020). Determinants of COVID-19 vaccine acceptance in the us. *EClinicalMedicine:* 100495, **26**(1).
31. Nezhad Z.B. & Deihimi M.A. (2022). Twitter sentiment analysis from iran about COVID 19 vaccine. *Diabetes&Metabolic Syndrome:* 102367, **16**(1).

32. Piedrahita-Valdes H., Piedrahita-Castillo D., Bermejo-Higuera J., Guillem-Saiz P., Bermejo-Higuera J.R., et. al. (2021). Vaccine hesitancy on social media: Sentiment analysis from june 2011 to april 2019. *Vaccines*, 9(1):28, [Online]. Available: https://doi.org/10.3390/vaccines9010028
33. Pogue K., Jensen J.L., Stancil C.K., Ferguson D.G., Hughes S.J., et. al. (2020). Influences on attitudes regarding potential COVID-19 vaccination in the united states. *Vaccines*, **8**(4), 582.
34. Praveen S., Ittamalla R. & Deepak G. (2021). Analyzing the attitude of indian citizens towards COVID-19 vaccine - a text analytics study. *Diabetes&Metabolic Syndrome: Clinical Research&Reviews*, **15**(2), 595-–599.
35. Samuel J., Ali G.G.M.N., Rahman M.M., Esawi E., & Samuel Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information*, **11**(6), [Online]. Available: https://www.mdpi.com/2078-2489/11/6/314, https://doi.org/10.3390/info11060314
36. Sarkar D. & Natarajan V. (2019) *Ensemble Machine Learning Cookbook: Over 35 practical recipes to explore ensemble machine learning techniques using Python*. Birmingham, United Kingdom: Packt Publishing, [Online]. Available: https://books.google.co.in/books?id=dCWGDwAAQBAJ
37. Scarselli F., Gori M., Tsoi A.C., Hagenbuchner M., Monfardini G., et. al. (2009) The graph neural network model. *IEEE Transactions on Neural Networks* , **20**(1): 61-–80. [Online]. Available: http://digital.library.wisc.edu/1793/60660
38. Settles B. (2009) Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences* , [Online]. Available: http://digital.library.wisc.edu/1793/60660
39. Shen F., Zhao X., Li Z., Li K., & Meng Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications:* 121073, **526**(1).
40. World Health Organization. (2020, March 15). *Global situation report-55*. [Online]. Available: https://www.who.int/publications/m/item/situation-report—55