

Analyzing sentiment dynamics from sparse text coronavirus disease-19 vaccination using natural language processing model

Jalaja Govindappa¹, Kavitha Channegowda²

¹Department of Computer Science and Engineering, Bhageerathi Bai Narayana Rao Maanay Institute of Technology, Visvesvaraya Technological University, Belagavi, India

²Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology Management, Bengaluru, India

Article Info

Article history:

Received Jul 26, 2021

Revised Mar 22, 2022

Accepted Apr 14, 2022

Keywords:

Coronavirus disease-19

Machine learning

Natural language processing

Sentiment analysis

Tweeter

Vaccination

ABSTRACT

Social media platforms enable people exchange their thoughts, reactions, emotions regarding all aspects of their lives. Therefore, sentiment analysis using textual data is widely practiced field. Due to large textual content available on social media, sentiment analysis is usually considered a text classification task. The high feature dimension is an important issue that needs to be resolved by examining text meaningfully. The proposed study considers a case study of coronavirus (COVID) vaccination to conclude public opinions about prospects for vaccination. Text corpus of tweets is collected, published between December 12, 2020, and July 13, 2021 is considered. The proposed model is developed considering phase-by-phase data analysis process, followed by an assessment of important information about the collected tweets on coronavirus disease (COVID-19) vaccine using two sentiment analyzer methods and probabilistic models for validation and knowledge analysis. The result indicated that public sentiment is more positive than negative. The study also presented statistics of trends in vaccination progress in the top countries from early 2021 to July 2021. The scope of study is enormous regarding sentiment analysis based on keyword and document modeling. The proposed work offers an effective mechanism for a decision-making system to understand public opinion and accordingly assists policymakers in health measures and vaccination campaigns.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Jalaja Govindappa

Department of Computer Science and Engineering, BNM Institute of Technology, Visvesvaraya Technological University

Belagavi, India

Email: jalajag.shaker@gmail.com

1. INTRODUCTION

Sentiment analysis is used for many purposes, such as determining the mood of social media users about a topic, their views on social events, and market price equilibrium [1], [2]. On the other hand, Twitter is widely preferred as a data source in sentiment analysis studies because it is a popular social network and is convenient for collecting data in different languages and content [3]. The coronavirus disease (COVID-19) is one of the trending topics on social media platforms, particularly on Twitter, since December 2019 and has it kept on to date. In recent months, the vaccine was introduced against coronavirus, and people have different opinions regarding vaccination. Still, most people are hesitating and making their personnel choice regarding vaccination. Many studies were conducted in recent years to understand possible aspects of the people's sentiments on the vaccination. However, considering sentiment analysis as a text classification problem and Twitter messages composed of short text, the dataset becomes sparse [4], [5]. This poses a significant

problem in terms of timing and performance, especially on large-sized data. For this reason, phase-by-phase text representation techniques are used in the proposed work to solve this performance problem arising from high feature space.

The proposed study presents a novel framework for sentiment analysis towards the COVID-19 vaccines based on the tweets. However, various research works on the similar context of sentiment analysis based on social media data are presented in the recent literature [6], [7]. In a similar direction, the authors in [8] collected the tweets over 3 weeks from the European continent to study the increasing impact of the coronavirus. A recent study presented in [9] focused on the temporal assessment of researches on coronavirus using different datasets with various machine learning (ML) and natural language processes (NLP). In [10], the researchers considered Twitter data for analyzing the trend of wearing a mask. The study has demonstrated that people are more serious about wearing masks on March 17, July 27, 2020. The researchers in [11] carried opinion mining using TextBlob sentiment analyzer on online knowledge delivery based on web scraping articles collected from blogging websites during the pandemic condition. The outcome demonstrated that most of the articles belong to positive sentiments than the news articles. The study on COVID-19 vaccination in Philippines is conducted in [12], where tweets in English and Filipino language with 81.77% accuracy. In [13], Chaudhri *et al.* have performed sentiment analysis to investigate whether the people accept the COVID-19 vaccine. The outcome of this study exhibited appositive attitude towards receiving the vaccine slots. However, this study has used very few samples of tweets and did not show the data collected for the analysis. A survey work of [14] mentioned effectiveness of sentiment analysis for different context needs large text corpus with the suitable techniques. Few research studies consider large text data for topic modeling to analyze various aspects of the COVID-19. Bai *et al.* [15] introduced a topic development study on COVID-19 news from Canada. Similarly, a topic modeling is carried to examine news media at an early stage of the coronavirus in China [16].

The research studies towards using the joint approach of topic modeling and sentiment analysis to investigate the impact of the COVID-19. Chandrasekaran *et al.* [17] and Xue *et al.* [18] presented study on topics and sentiments of user conversation on Twitter regarding COVID-19 pandemic. The study has adopted linear discriminant analysis and valence aware dictionary and sentiment reasoner (VADER) to perform sentiment classification. In the same way, Xue *et al.* [18] used latent dirichlet allocation (LDA) technique over tweets to understand to analyze public sentiment and mental status during COVID-19 pandemic. Recently, much deep learning is adopted for NLP tasks. One such approach is given by [19] where Yang *et al.* have developed a learning model the exhibits increase positive opinion over different timescale. In this study, a million multilingual tweets and manually annotated based on different fine-grained emotions. In [20] an analysis is made on the issue of fear and panic condition of people were considered based on tweet conversation. The authors have conducted comparative analysis which shows naïve Bayes is superior to logistic regression in the sentiment classification. In [21] presented a detailed design of tweet dataset, representing temporal and spatial dimensions for understanding the crisis due to pandemic. The authors have introduced network clusters with the identity of people from different regions. Chakraborty *et al.* in [22] applied Gaussian fuzzy classifier on social media data to assess the sentiments of the people during initial stage of COVID. The authors have mentioned how popularity negatively impacts the accuracy of the sentiment analysis [23]. An approach of the bidirectional encoder representations from transformers (BERT) model is adopted in [24], [25] to conduct opinion mining considering two different datasets, one dataset subjected to Indian tweets and the other based on the tweets collected from the overall world. However, this study has not shown any comparative analysis of their model.

2. THE PROBLEM DESCRIPTION

Based on the literature analysis, it has been identified that there existing significant problems that need to be effectively resolved to perform sentiment classification and forecasting of other contexts. Existing approaches of knowledge extraction have shown a more extensive scale of dependencies towards using supervised learning approaches. The adoption of supervised learning approaches significantly offers computational complexity, and it offers accuracy at the cost of the larger size of training data. However, most of the existing approaches are applicable for offline analysis using a complex analytical strategy, which is not cost-effective. Data transformation is one of the primary steps that contribute to data accuracy in sentiment analysis. Existing approaches did not emphasize any transformation process and did not demonstrate how they have collected the dataset and what steps they have taken to analyze that effectiveness. In this work, all significant problems were considered, and effective modeling is carried for sentiment analysis. The proposed study presented a significant contribution to address the following challenges particularly: i) collecting and effectively analyzing collects a large amount of text corpus for the sentiment analysis, ii) an exploratory analysis of the dataset to understand the nature of data, and iii) sentiment analysis for vaccination progress in a different color in the fixed timeline for topmost countries. Hence, the adaptiveness of the proposed model.

The next subsection presents the proposed solution to address above mentioned challenges by introducing an effective and computational-efficient sentiment analysis model.

2.1. The proposed solution

The proposed study aims to enhance opinion mining using a supervised learning-based probabilistic model to understand better the public attitudes and emotions towards the viewpoint of the COVID-19 vaccination when it became widely available to most countries. This kind of analysis can be very operative for gaining insight into the status of the vaccination campaign and whether the people are aware of the situation. Another significant scope of the proposed study is that it can provide a very effective decision-making support system concerning health and life security. Therefore, the proposed work introduces a model for public opinion analysis based on their text conversation on social media platforms.

The study collects the data from a Twitter website that users posted to express their sentiments using emojis and #Tags. The tweets were collected in the English language then stored in the system database. The dataset consists of unstructured and irrelevant information that is not required in this analysis. Hence, preprocessing is carried out to remove unwanted data and perform tokenization. The labeling of the dataset is carried out based on the polarity computation and subjectivity analysis. Afterward, the dataset is split into two subcategories such as training dataset and testing dataset. Significant attributes term frequency-inverse document frequency (TF-IDF) is computed to perform sentiments analysis using a probabilistic model. The schematic architecture of the proposed framework for opinion mining-based tweet's analysis is depicted in Figure 1.

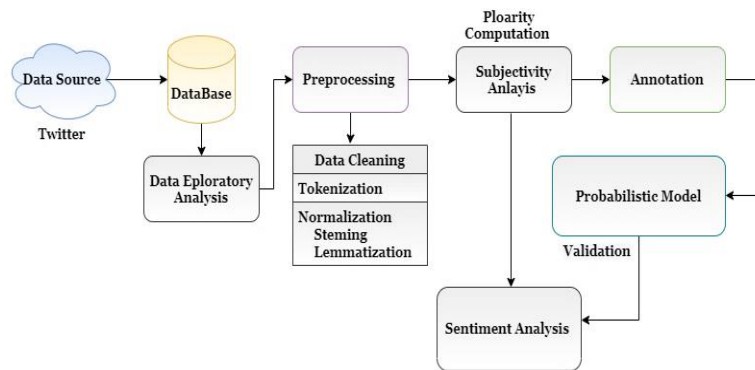


Figure 1. Illustration of the schematic architecture of the proposed system for public opinion mining

Sentiment analysis plays an important role to derives insights that help many sectors to grow and improve, like businesses, markets, healthcare, decision, and planning. Since the proposed study has considered a case study of COVID-19 vaccination, the sentiment analysis is usually observed as a text classification problem. High dimensional feature space is a significant issue that needs to be addressed for effective results. Therefore, the study presents an effective sentiment analysis system that does not suffer from features extraction problems and provides a compelling analysis of people's conversations to understand better their sentiments, mental ability, the trend of a topic, and needs in the context of the COVID-19 situation. The significant contribution of the proposed research word is discussed: i) the study extracted the large corpus of tweet texts on COVID-19 vaccine-related keywords and terms used globally; ii) the study conducted phase-wise data modeling for data preparation for the sentiment evaluation; iii) the study implemented two sentiment analysis models to explore their effectiveness regarding public emotion and sentiment on COVID-19 vaccine; iv) further, the supervised probabilistic approach of the classifier is implemented to validate the effectiveness of the sentiment analyzer; v) the study provides sentiment analysis in positive, negative, and neutral opinions about the COVID-19 vaccination. Also, time-series statistics of the vaccination process around the world are shown in the proposed work; and vi) the proposed study offers to provide meaningful and cross-cultural information on the trend of COVID-19 vaccination and public concern.

The remaining part of this paper is described: section 2 discussed a proposed method adopted in system designing and developing. This section highlights data collection process, dataset modelling and sentiment analysis. Section 3 discusses the probabilistic model adopted in the study for knowledge analysis and validation. Section 4 presents result analysis for the proposed system and finally section 5 concludes overall contribution of the proposed work.

3. PROPOSED METHOD

The proposed study describes the dataset collection procedure and dataset preprocessing operations. Also, this section presents learning models and implementation strategies adopted for the classification of sentiment analysis COVID-19 vaccination by discussing a mechanism for feature generation. The implementation method further discusses on word embedding for analyzing the text.

3.1. Data collection

The contextual information for data gathering is acquired from the Ritchie *et al.* [26], which provides important information such as which countries are using what vaccine and vaccination progress. The acquired information acts as a basis for collecting data subjected to the recent tweets. A Twitter account is created and linked with Twitter API using python library Tweepy to collect tweets regarding COVID-19 vaccination. The Tweepy library takes a parameter and provides tweets related to public opinion in return. This parameter includes usage of frequently used relevant terms (COVID-19, coronavirus, first dose, second dose, 1 dose, 2 doses) and a keyword concerning peoples thought on the vaccine (such as safe, harmful, health after vaccination) is considered in the search process with the help of Twitter API filter. The search process with Twitter API filter also considers different vaccines used in the entire world, such as BioNTech, Sinovac, AstraZeneca, Sinopharm, Moderna, Covaxin, and Sputnik. The tweets data are collected in the only English language published between December 12, 2020, to July 13, 2021. The retrieved tweets with different keywords and terms regarding public opinion on the COVID-19 vaccine were merged in the existing file and stored in the local database under various fields mentioned in Table 1 In total, 130,036 tweets and 113,743 hashtags were generated by the tweeps. The computing procedure for data collection using Twitter API is discussed in the algorithm 1:

Algorithm 1: Twitter Data Gathering

Input: Tweeps or user Profile (P), Tweepy API (T_{API}), Keyword (K) for Tweets (T)

Output: Tweet data fields (DF)

Start:

Foreach Tweets from K list do

Auth = tweepy.OAuthHandler(P_{key} , P_{secret})

Auth.set_access_token($access_{key}$, $access_{secret}$)

$T_{API} = f1(\text{Auth}, \text{with_on_rate_limit} = \text{True})$

Username $\in P$

Init max_T \in length of tweets

Foreach Tweets from

DF = $f2(T_{API}, \text{user_timeline}, \text{id} = \text{username}), \text{items}(\text{max_T})$ do

$K_{list} = [\text{List of Tweeter User Attributes}]$

Store_to_local_databasedo //using panda library

Tweets = $\text{pd.df}(K_{list}, \text{columns} = [\text{DF}])$

End

The above-mentioned pseudo construct demonstrates a computing step for Tweeter data collection using Twitter API. ' T_{API} ' linked with user Twitter account profile (P) uses two explicit functions $f1$ and $f2$. The function ' $f1$ ' refers to the Tweepy API, and the ' $f2$ ' refers to the function of python library for constructing collected data into a structured format. In Table 1, description of data collected or gathered is highlighted, including i) data fields (DF) consists of different attributes related to tweeps (T) such that $DF \in \{T_1, T_2, T_3 \dots T_{12}\}$, ii) counts (C) of samples of DF such that $C(DF) \in \mathbb{Z} + \{1, N\}$, iii) data type (DT) such that $DT \in \{\text{Int}, \text{float}, \text{String}, \text{Boolean}\}$, and iv) description $D \in \{\text{DF}\}$. The next sub-section discusses the process involved in the preprocessing of the acquired dataset.

Table 1. Data description

Data Fields	Counts	Data Type	Description
Id	130036	float	Id of the Tweeps (User)
name	130036	String	Name of Tweeps
Geography	96603	String	Location of Tweeps
Friends	130036	Int	Contains list of Tweep's friends
Followers	130036	Int	Contains list of Tweep's followers
Verified	130036	bool	Shows authenticity of Twitter account (True/false)
Text	130036	String	Tweets (Public opinion)
hashtags	113743	String	Presence of hashtags (#)
date	130036	String	Date of Tweets
UE_source	139630	String	Device information (Android, Apple, andWindows)
retweets	130036	Int	Shows number of re-posting of tweets
is_retweet	130036	bool	Shows presence of re-posting of tweets

3.2. Preprocessing

The proposed study has conducted an exploratory analysis to understand the preprocessing operation requirement over the collected Twitter dataset. Based on the exploratory analysis, it has been observed that the dataset is associated with irrelevant characteristics, punctuation, stop words, data fields, repetitive text, and many more, which are not significant in the sentiment analysis. Therefore, preprocessing operation is required to discard and clean such irrelevant data from the text field of the dataset. The entire preprocessing operation is carried out in the following manner discussed:

3.2.1. Omission of irrelevant data

In this step of preprocessing, removal of URLs, hashtags (#), special characters, quotes, empty spaces, repeating words punctuation are carried out. Here along with lowercasing, emoticons are converted to meaningful sentences. Duplications are also considered to make data more effective for preprocessing.

3.2.2. Tokenization

The tweet's text is split into smaller units (tokens). To extract this tweets tweepy is considered, a simple approach to utilize python library. Through this relevant tweets are extracted in a straight forward method. This process is carried out to provide a generalization ability and understanding to the model by interpreting the sequence of the text into smaller units.

3.2.3. Omission of stopwords

The stop words generally consist of less information in the text, such as 'and', 'but', 'the', 'like', and many more. To make extraction of text effectively, stop words which adds ambiguity to be removed. These words do not describe the meaning of contents; they can be ignored without sacrificing the contextual information of the Tweets text.

3.2.4. Normalization

In this process, the Tweets text is normalized to its base form using the stemming and lemmatization process. The stemming process reduces the text arguments to their stem. For example, 'tradition' and traditional have the same stem, 'tradi'. In lemmatization, text words are converted to their base form according to the part of speech. For example, the word 'changing' gets converted to base form 'change'. The computing procedure for Twitter dataset preprocessing is discussed in the algorithm 2:

Algorithm 2: Twitter Data Preprocessing

```

Input: DF
Output: Preprocessed Tweet_texts
Start:
Init DFp → []
                                Load → DF['Geography', 'date', 'text']
DF['date'] → convert to DateTime
DF['Text'] → Drop duplicate
                                def function: Text_preprocessing(text)
Text_lower = re.findall(text, '(.[a-z]|[A-Z])) do
                                Text = text.lower()
                                Text = text.replace argument(Text, '@\w+', '#', RT[\s]+, 'https?://\S+')
                                return = Preprocessed Tweet_texts
                                Foreach Tweet_text from DF do
DFp = Text_preprocessing(DF.text)
                                DFp = tokenizer(DFp)
                                DFp = stem(DFp)
End

```

The procedure mention in algorithm 2 describes the computing steps for performing the preprocessing operation over the Twitter dataset. In this process, initially, an empty vector is initialized as DF_p (preprocessed tweet data fields) that stores preprocessed data for further sentiment analysis. The data field 'date' is transformed to the standard representation that shows both date and time of tweet texts. Another data field, 'geography', is considered to analyze missing fields in the collected tweet data field (DF). Further, significant preprocessing is carried out on the tweet texts using a user-defined function Text_preprocessing(). This function takes the 'text' field of DF and applies to find, remove, and replace operation for the specific category of tweets contents. Table 2 presents a summarization of preprocessing actions taken for the text arguments (i.e., tweets contents). The proposed study also presented the word clouds in Figure 2 to visualize the frequency of the words after being preprocessed. The text cloud illustrates

Table 3. A sample representation of tweets dataset with sentiment score and label

Geography	Date	Text	Subjectivity	Polarity	Sentiments	Label
Bengaluru, India	2021-07-17 05:00:00	45 rural bengaluru covid vaccine availability	0.400000	0.20	Positive	1
Singapore	2021-10-07 01:22:00	done job Vaccination done	0.000000	0.000000	Neutral	0
Vancouver, Canada	2020-12-12 20:23:00	facts immutable	0.550000	-0.050000	Negative	2

Algorithm 3: Data Annotation for Opinion mining

```

Input: text
Output: sentiments, Labels
Start:
    Foreach Tweet_Text from DF do
        get_sub → f3(DF['text'])
        update_DF → DF.append get_sub do
            get_pol → f4(DF['text'])
            update_DF → DF.append [get_pol])
        def function: sentiment_analysis (pol_score)
            sentiment = positive
            else check pol_score == 0) do
                sentiment = neutral
            Otherwise sentiment = negative)
            Foreach tweet_text from DF do
                sentiment = sentiment_analysis(DF.Text[pol_score])
        update_DF → DF.append [sentiment]do
    Labelling → [positive: 1, negative: 2, neutral: 0]
End

```

The above-mention algorithm performs data annotation and labeling process based on the sentiments determined from subjectivity (sub) and polarity score (pol_score) using an unsupervised method by using two sentiment analysis functions $f3$ and $f4$. The function $f3$ is meant for subjectivity score and $f4$ for polarity score using Textblob. However, a similar computing procedure is also applicable for VADER. The scatter plot representing frequency distribution of subjectivity and polarity score of tweets text is given in Figure 3. The analysis obtained from Figure 3 provides a clear view of sentiments distribution using TextBlob, where it can be seen that dense tweet texts are aligned towards polarity score 0, which clearly indicates more people having a neutral opinion. Particularly, 52.15% of total tweets are subjected to neutral opinion, 11.73% of total tweets fall under the category of negative opinion, and 36.12% of total tweets are subjected to positive opinion. The statistics of sentiments count obtained from both methods is shown in Figure 4.

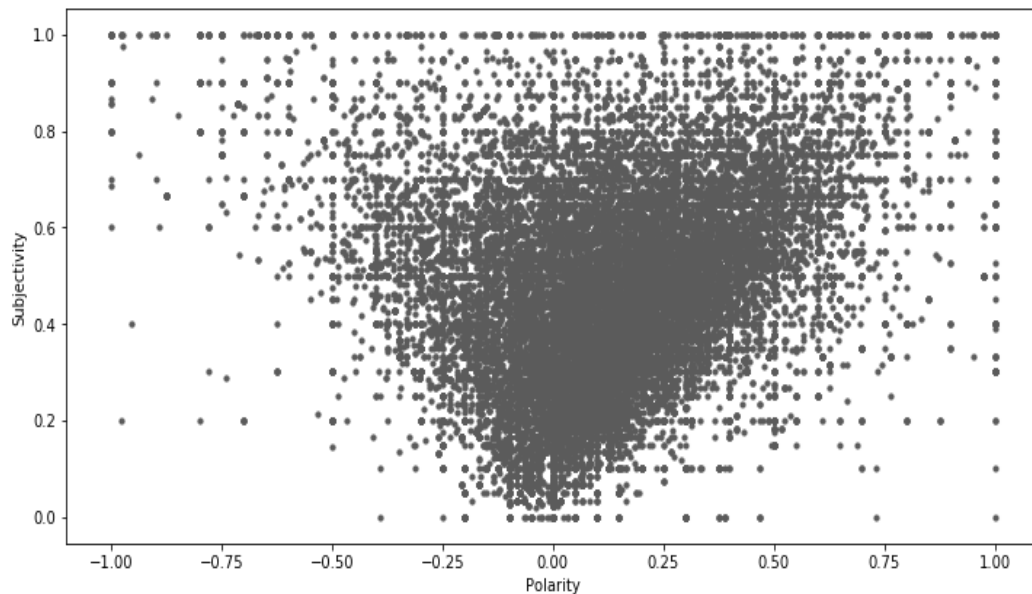


Figure 3. Scatter plot for subjectivity vs a polarity

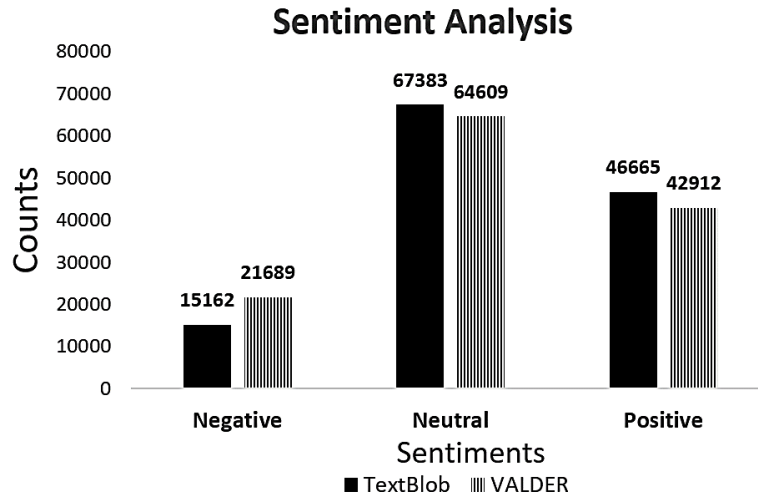


Figure 4. Statistics of sentiments regarding tweets texts

The statistics obtained from Figure 4 show the sentiments obtained from Textblob and VADER unsupervised methods. Based on the analysis, it has been observed that TextBlob has shown more positive and neutral sentiments compared to VEDAR. TextBlob 46,665 tweet texts reflect positive sentiments, 15,162 tweet texts indicate negative sentiment, and 67,383 tweet texts reflect the neutral attitude of people. In the case of VEDAR 21,689 tweets represents negative sentiment, 64,609 and 42,912 tweet text indicates neutral and positive sentiment. Both techniques have achieved similar performance in terms of positive and neutral sentiments. However, performance much varies for negative sentiment. The study also shows statistics demonstrating the current trend of the vaccination among different countries given in Figure 5.

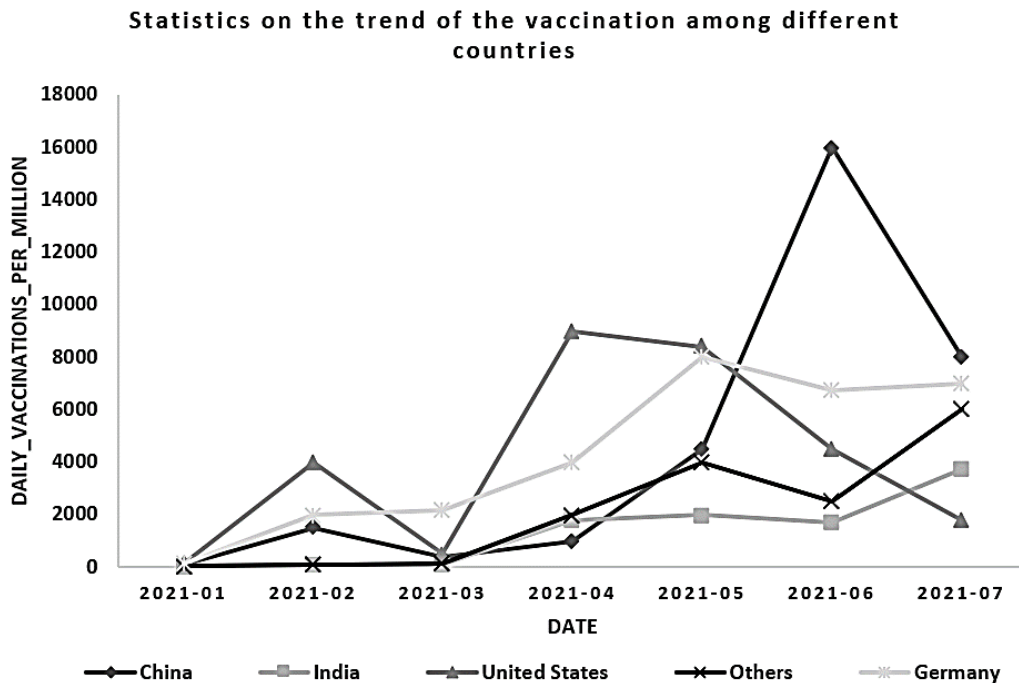


Figure 5. Statistics of vaccination progress

3.4. Validation with supervised learning mechanism using probabilistic model and SVM

Performing analysis of knowledge extracted plays the major role. Along with extraction opinions are also to be considered. In order to get knowledge analysis and more insight on the performance of both opinion

mining methods, i.e., TextBlob and VADER, this section presents an implementation of the supervised learning mechanism of naïve bayes and support vector machine (SVM) classifier.

3.4.1. Naïve Bayes

The function of naïve Bayes is concerned with an approach of the probabilistic model, i.e., numerically expressed in (1):

$$P(M|N) = \frac{P(N|M)P(M)}{P(N)} \quad (1)$$

where, $P(M)$ denotes the prior probability being true, $P(N)$ refers to the probability of the data, $P(M|N)$ is the probability of presumption M for the available data N , and $P(N|M)$ denotes the probability of N given that presumption M is true. The proposed study explores public sentiments using text-based data. Therefore, for a collection of tweet text documents such that $T \in \{T1, T2, T3 \dots TN\}$ with a set of preassigned sentiment class $L \in \{1, 2, \text{ and } 3\}$, the task is to select a classification function f that produces the correct sentiment for each input document, such that $f(T_i) = L$. In this regard, it becomes vital to compute the probability of all possible values of M and the predicted labels with thoroughgoing probability, which can be numerically expressed in (2):

$$M = \operatorname{argmax}_M P(M) \prod_{i=1}^k P(N_i|M) \quad (2)$$

In order to train the classifier, the dataset is split within the ratio of 70% training and 30% testing set. Further, the study performs TF-IDF (terms frequency-inverse document frequency) vectorization to represent tweet text into word vector as a suitable input for the naïve Bayes classifier. The terms frequency can be computed using the following numerical (3):

$$TF(T, W) = \frac{N}{W} \quad (3)$$

where, T denotes the count of the word that appears, D represents the number of words in the text document. The term frequency (TF) deliberates all words equally significant, and inverse document frequency (IDF) only considers the unique terms, numerically can be represented in (4):

$$IDF(T) = \frac{N}{W} \quad (4)$$

The expression mentioned in (4) measures the significance of W , and N is the appearance of a number of words. However, W can be zero, in order to avoid division by zero, the following numerical expressions 5 and 6:

$$IDF(T) = \log\left(\frac{N}{DF+1}\right) \quad (5)$$

In (5) 1 is appended to the frequency of word in text document (DF). The final version of expression can be numerically expressed:

$$TF - IDF(T, W) = TF(T, W) \times IDF(T) \quad (6)$$

3.4.2. Support vector machine (SVM)

A SVM is most popular supervised learning mechanism based on the vector principle used to address classification and regression problem. The advantage of SVM is that it does not prone to overfitting problem unlike another machine learning classifier. Therefore, the proposed study implements SVM classifier in order to perform comparative analysis with the naïve Bayes technique. Since, the proposed system concern with multiple sentiment polarity regarding public opinion or attitude towards COVID vaccination. Therefore, the study considers implementation of multi-class SVM as shown in Figure 6.

In Figure 6 demonstration of multiclass SVM is given, where input tweets text such that $T \in \{T1, T2, T3 \dots TN\}$ is mapped to the output class such as positive (Pos), neutral (neu) and negative (Neg) using linear function such that $f(x) = (w \cdot \sigma(x)) + b$, where w and b are the weight and bias respectively and $\sigma(x)$ denotes mapping function of SVM kernel (K). In the implemented multiclass SVM classifier, the proposed study uses two SVM kernel (K1 and K2) connected in series, where each provides its output with

single polarity class based on the utmost approximated values. However, usage of multiple SVM kernel may pose computational overhead. In this regard, an approach of transition matrix which acts as an auxiliary mechanism to each SVM kernel towards processing text data from the observation state to the decision state. Also, feature learning performance in training phase depends on the kernel function used in the SVM classifier. Although, there are various kernel function available for the SVM, but based on the empirical analysis, radial basis kernel function is considered in the SVM implementation numerically given:

$$k(x, y) = \exp\left(-\frac{\|x-y\|^2}{\gamma}\right) \quad (7)$$

where, $\|x - y\|^2$ denotes measure of Euclidean space of length between two data points x and y , and γ refers to hyperparameters such that $\gamma = 2\sigma^2$, where σ refers to the variance. The core objective of this kernel function is to compute similarity or closeness of two data points towards each other. The next section discusses the performance analysis of implemented classifiers for public opinion analysis.

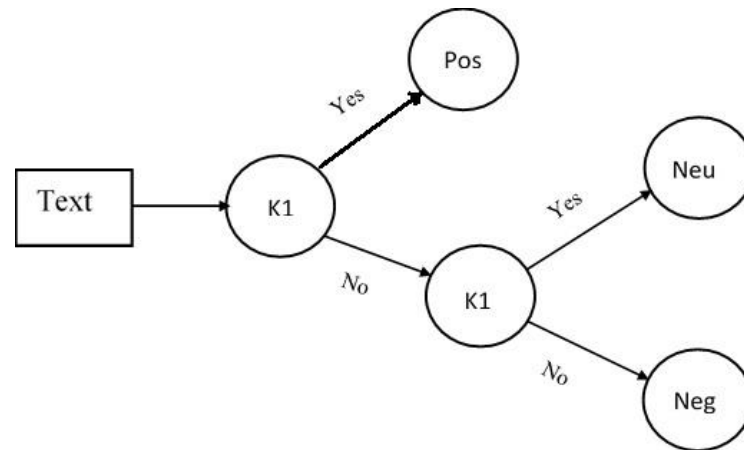


Figure 6. Illustration of classifier with multiple SVM Kernel

4. RESULTS ANALYSIS AND DISCUSSION

The entire modeling and development of the proposed system are carried out on the computing environment Anaconda using Python. Naïve Bayes and SVM classifier, a supervised learning mechanism, are considered for public opinion mining regarding COVID-19 vaccination. Multinomial naïve Bayes (MNB) classifier is selected as it can handle a large tweet text corpus and is suitable for sentiment prediction based on the text data. SVM is selected as it less prone to overfitting problems and suitable for better feature learning in high dimensional space. The training of the implemented classifier is carried on the preprocessed dataset, where 70% dataset, i.e., 90,447 samples, were selected out of a total of 129,210 samples remained after preprocessing operation. The classifier namely MNB is trained and also it is tuned by adjusting its hyperparameter i.e., alpha (α) that handles the problem of zero probability and performs smoothing in the training process. A grid search strategy is adopted to get optimal α values with cross validation approach to assess the model performance. In the case of SVM, the classifier is tuned considering hyperparameters namely C equal to 10 (acts as a regularizer to control error), gamma equal to 0.0001 and kernel equal to radial basis function (RBF). Since the dataset is preprocessed, it does not associate with any imbalance factor or skewness, and it has an equal number of sentiment labels. Therefore, the study considers accuracy as the primary metric for the performance assessment.

4.1. Outcome analysis

In this section, the performance analysis is carried out for both supervised learning classifiers considering their output obtained based on the training dataset prepared from both TextBlob and VADER sentiment analyzer. Based on the analysis from Figure 7, it has been found that MNB has scored the highest accuracy (91.69%) for predicting sentiments of people regarding the COVID vaccine, whereas SVM has achieved a little less accuracy score (91.19%) compared to MNB in case of TextBlob sentiment analyzer. In the case of other sentiment analyzers, i.e., VADER, SVM outperforms MNB by achieving a higher accuracy score (i.e., 86.63%) than SVM (84.22%). Based on the overall analysis, it seems that TextBlob is better than VADER, and SVM also seems to be better than MNB.

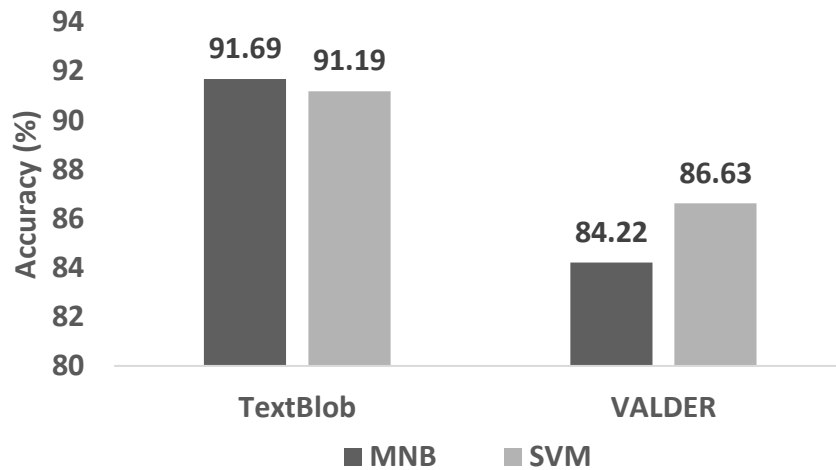


Figure 7. Sentiment classification accuracy (%) with the probabilistic model naïve bayes

4.2. Discussion

To get best results, proper and appropriate training is to be considered. Therefore, the classifiers used are needed to be trained. It is to be noted that both classifiers' models are trained over that dataset prepared based on the different sentiment analyzer, i.e., TextBlob and VADER. Therefore, the performance analysis is carried out regarding classifiers and carried out for both sentiment analyzers.

4.2.1. Analysis regarding sentiment analyzer

Both sentiment analyzers provide a wide range of features in the text classification problem. However, both techniques are associated with some advantages and disadvantages. The main advantage of using TextBlob is that it can efficiently handle many text data without posing much computational overhead. VEDAR sentiment analyzer has a wide range of features concerning sentiment lexicons. It doesn't have a large variety of features, and thus, users have to depend on some other libraries for advanced tasks. However, it poses a little higher computational overhead compared TextBlob. The VADER bets suit the text that consists of slang, and emojis, whereas TextBlob suits better with plain and formal text representation. The TextBlob sentiment analyzer is better comparatively than VADER because our dataset is preprocessed and does not consist of emoticons and slang in the text dataset.

4.2.2. Analysis of supervised classifiers

Therefore, it is quite difficult to say that which classifier among implemented SVM and MNB is better for text classification. However, a closer analysis shows that both classifiers have achieved similar performance to each other. However, SVM can be considered the most suitable as it is not susceptible to catastrophic failures. It better correlates the similarity in the data from the large text corpus, making its generalization and feature representation better and more accurate, leading to better performance in sentiment polarity score prediction. However, in the case of the outcome are not found to be consistent enough. Based on the overall analysis, the TextBlob and SVM are much better than the VADER and MNB in analyzing people's opinions or sentiments regarding the COVID-19 vaccination.

4.2.3. Findings and scope of the study

The proposed work's major contribution is that it provides a steppingstone to explore and analyze public sentiment based on their conversations about COVID-19 vaccination on social media platforms. The study findings suggest that public sentiment is more positive than negative sentiments. However, most sentiments are observed to be neutral. It also has been found that peoples are conscious about their health and lifestyle after vaccination with positive sentiments. However, the good news is that under any circumstances, negative sentiment does not exceed positive sentiments. The study also mentioned the time series statistics of the trend of vaccination for topmost countries between January 2021 to July 2021. This clearly shows how the peoples from different countries have welcomed the vaccine against novel coronavirus. Apart from this, the proposed system also has another scope, like analyzing other topics like wearing a mask, social distancing, traveling, popular vaccines, and many more. The proposed study provides meaningful and cross-cultural information on the trend of COVID-19 vaccination and public concern.

5. CONCLUSION

This paper has presented a sentiment forecast model for analyzing people's attitudes regarding the COVID-19 vaccination. The design and development of the proposed model are carried out considering phase-wise data modeling based on the collection of large tweet text corpus, thereafter, assessing sentiments of people regarding COVID-19 vaccine using two sentiment lexicon methods and probabilistic model for knowledge analysis. This work is focused on facilitating a model that can perform sentiment analysis of public opinion mining and provide an effective decision-making process in the context of healthcare and a healthy lifestyle. The study validated the adopted sentiment methods with two supervised learning approaches, such as MNB and SVM. Also, it showed a comparison based on the obtained results so that a suitable mechanism can be adopted for sentiment analysis. Based on the comparative analysis, SVM is superior to the MNB, and TextBlob is better than the VADER sentiment analyzer. It also serves as a guideline to help public health and policymakers to provide the public with necessary services and resources. It also provides effective information to government or health officials to better understand vaccination activities. In future work, the proposed work may be extended to propose more robust designs that support deep learning mechanisms or artificial intelligence and focus on the stability and security aspects to be suitable for real-time deployment scenarios.




REFERENCES

- [1] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, Nov. 2015, doi: 10.1016/j.knosys.2015.06.015.
- [2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
- [3] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *AI and SOCIETY*, vol. 30, no. 1, pp. 89–116, Feb. 2015, doi: 10.1007/s00146-014-0549-4.
- [4] V. A. and S. S. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 5–15, Apr. 2016, doi: 10.5120/ijca2016908625.
- [5] S. Joshi and D. Deshpande, "Twitter sentiment analysis system," *arXiv preprint arXiv:1807.07752*, Jul. 2018, doi: 10.5120/ijca2018917319.
- [6] D. D. Albesta, M. L. Jonathan, M. Jawad, O. Hardiawan, and D. Suhartono, "The impact of sentiment analysis from user on Facebook to enhanced the service quality," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3424–3433, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3424-3433.
- [7] A. Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa, and J. Baldwin, "Sentiment analysis of comments in social media," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 5917–5922, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5917-5922.
- [8] A. D. Dubey, "Twitter sentiment analysis during COVID19 outbreak," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3572023.
- [9] A. Ebadi, P. Xi, S. Tremblay, B. Spencer, R. Pall, and A. Wong, "Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing," *Scientometrics*, vol. 126, no. 1, pp. 725–739, Jan. 2021, doi: 10.1007/s11192-020-03744-7.
- [10] A. C. Sanders *et al.*, "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse," *medRxiv*, Sep. 2020, doi: 10.1101/2020.08.28.20183863.
- [11] K. K. Bhagat, S. Mishra, A. Dixit, and C.-Y. Chang, "Public opinions about online learning during COVID-19: a sentiment analysis approach," *Sustainability*, vol. 13, no. 6, Mar. 2021, doi: 10.3390/su13063346.
- [12] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Twitter sentiment analysis towards COVID-19 vaccines in the Philippines using naïve bayes," *Information*, vol. 12, no. 5, May 2021, doi: 10.3390/info12050204.
- [13] A. A. Chaudhri, S. S. Saranya, and S. Dubey, "Implementation paper on analyzing COVID-19 vaccines on twitter dataset using tweepy and text blob," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 3, pp. 8393–8396, 2021.
- [14] P. P. Rokade and A. K. D., "Business intelligence analytics using sentiment analysis-a survey," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 613–620, Feb. 2019, doi: 10.11591/ijece.v9i1.pp613-620.
- [15] Y. Bai, S. Jia, and L. Chen, "Topic evolution analysis of COVID-19 news articles," *Journal of Physics: Conference Series*, vol. 1601, no. 5, Aug. 2020, doi: 10.1088/1742-6596/1601/5/052009.
- [16] Q. Liu *et al.*, "Health communication through news media during the early stage of the COVID-19 outbreak in china: digital topic modeling approach," *Journal of Medical Internet Research*, vol. 22, no. 4, Apr. 2020, doi: 10.2196/19118.
- [17] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: temporal in-foveillance study," *Journal of Medical Internet Research*, vol. 22, no. 10, Oct. 2020, doi: 10.2196/22624.
- [18] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter," *PLOS ONE*, vol. 15, no. 9, Sep. 2020, doi: 10.1371/journal.pone.0239441.
- [19] Q. Yang *et al.*, "SenWave: monitoring the global sentiments under the COVID-19 pandemic," *arXiv preprint arXiv:2006.10842*, Jun. 2020.
- [20] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, Jun. 2020, doi: 10.3390/info11060314.
- [21] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, May 2021, doi: 10.1007/s10489-020-02029-z.
- [22] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment analysis of COVID-19 tweets by deep learning classifiers-a study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2020.106754.
- [23] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Network Analysis and Mining*, vol. 11, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00737-z.




- [24] V. Bonta, N. Kumares, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, Mar. 2019, doi: 10.51983/ajcst-2019.8.S2.2037.
- [25] A. Shrivatava, S. Mayor, and B. Pant, "Opinion mining of real time Twitter tweets," *International Journal of Computer Applications*, vol. 100, no. 19, 2014.
- [26] H. Ritchie *et al.*, "Coronavirus (COVID-19) testing," *Our World in Data*, 2020. <https://ourworldindata.org/coronavirus-testing> (accessed Jul. 13, 2021).
- [27] S. Loria, "Textblob documentation (release 0.16.0)," 2020. <https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf> (accessed Jul. 13, 2021).

BIOGRAPHIES OF AUTHORS



Jalaja Govindappa    received B.E. from Kuvempu University, MTech. and Ph.D. from Visvesvaraya Technological University. She has guided more than 15 Post Graduate students. Her research interest includes Text Mining, Internet of Things, Machine Learning. She is working as an Associate Professor in the Department of CSE, BNMIT. She has several publications, international Journals, and Conferences. She can be contacted at email: jalajag.shaker@gmail.com.



Kavitha Channegowda    received B.E., M.E. from Bangalore University, Ph.D. from Visvesvaraya Technological University. Under her able guidance, three students have received a Doctoral degree, and currently, three students are pursuing research. Her research interest includes Wireless Sensor Networks, Internet of Things, Machine Learning, Mining. She is working as a Professor in the Department of CSE, DSATM. She has several publications in national and international Journals. She can be contacted at email: kavitha_prasanna@yahoo.com.