



INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 7, Issue 5 - V7I5-1175)

Available online at: <https://www.ijariit.com>

Sentiment analysis of COVID-19 Vaccination Tweets on Twitter using Machine Learning Algorithm

Disha Jethva

dishaviniyom1310@gmail.com

Shah and Anchor Kutchhi Engineering College, Mumbai, Maharashtra

ABSTRACT

After Declaring pandemic in March 2020 public health prevention measures have proven to somewhat effective in limiting the spread of COVID-19. Protective immunity through vaccination will be great importance of in ending the pandemic. This work aims to identify the sentiments of the masses towards vaccination by analyzing the text tweets. 68,654 tweets are retrieved from twitter posted within the timeline from December 2020 to January 2021. Sentiment, polarity score and subjectivity score was computed and analyzed on the basis of text and date columns. According to the polarity and subjectivity scores tweets were classified as positive, negative and neutral using TextBlob library of natural language processing. Also, user's view on the vaccination was analyzed using machine learning algorithms such as Naïve Bayes (NB) and Logistic Regression (LR). The highest accuracy achieved was 90% by Logistic Regression (LR). It was observed from results that the sentiments towards the vaccine were positive in initial day but a shift to negative sentiments was observed. Later, the sentiment towards vaccine is again positive.

Keywords: COVID-19, Machine Learning, Naïve Bayes, Sentiment Analysis, Time Series Analysis, Vaccine.

1. INTRODUCTION

After Declaring pandemic in March 2020 public health prevention measures have proven to somewhat effective in limiting the spread of COVID-19 [7]. Protective immunity through vaccination will be great importance of in ending the pandemic [6]. COVID-19 has affected more than 90 million populations and three million deaths throughout the world [5]. The rest of the world has faced complete or partial lockdown from the year 2020; it has become economically not feasible to continue with it [6]. The vaccination started in the month of January 2021 with rate more than 90% has announced, the people reacted their views or opinion on twitter and other social media platforms about vaccination process as well as how vaccine will affect for different age of people [8]. It is estimated that at least of 70% population will need to be vaccinated at the end of the December 2021 to some level of herd immunity [7]. It has become mandatory to analyze what are the sentiments of people for COVID-19 vaccine [2]. Twitter was employed as the main source due to its popularity for discussion related to health information [7]. The main objectives were followed: (i) to find polarity and subjectivity according to text column. (ii) To find sentiments like positive, negative and neutral using date as well as text attributes. (iii) To compare the evaluation parameters [1]. A number of 2 349 659 tweets have been collected and a cleaned dataset containing 752 951 tweets has been extracted [3]. After extracting data we had performed stance detection, the performance for stance detection of several machine learning algorithms has been compared on the dataset [8]. The best performing algorithm has been selected and used for analyzing the sentiments, polarity and evaluation parameters [6]. Sentiment analysis is the process by which judgments and opinion are identified computationally analyzing a piece of text [1].

The remaining paper is organized as follows: Section 2 provides literature reviews in two main parts of natural language processing – focusing on sentiment analysis and stance detection from recent study on opinion of people based on COVID-19 extracted data from twitter. Section 3 gives the description of dataset of COVID-19 vaccination started in the month of January 2021 which was extracted from twitter. In next section i.e. Section 4 focus on the methodology used to find polarity and subjectivity scores, as well as sentiments like positive, negative and neutral. Comparison of different machine learning algorithm and find the evaluation parameters are also add in this section. The next Section 5 is about the analysis of result and visualization of the data. Section 6 contains conclusion and future scope of this sentiment analysis [8].

2. LITERATURE REVIEW

Numerous studies and research works have been conducted in the area of sentiment analysis. To find polarity of all tweets, sentiment polarity for tweets defines the expression as positive, negative and neutral. First create python script using TextBlob library to get sentiment polarity from the tweets. TextBlob provides an Application Programming Interface (API) for Natural Language Processing (NLP) tasks [5].

The sentiment and emotion analysis for the tweets were being done Syuzhet package version 1.0.1 was used to classify the tweets on the basis of sentiments (positive and negative) and also categorizes them into 8 emotions (fear, joy, anticipation, anger, disgust, sadness, surprise, trust). After the data cleaning, the NRC Emotion lexicon was applied using get_nrc_sentiment function to analyze the tweets [6].

To assign a polarity of ‘positive’, ‘negative’ or ‘neutral’ to each tweet, it utilized Valence Aware Dictionary and sentiment Reasoner (VADER), a Python lexicon and rule-based sentiment analysis tool. VADER is designed to determine sentiments of social media posts based on individual words and sentences. Sentiments were assigned based on scores given by VADER [7].

These tweets are explored in three ways. Informal opinion mining reveals both concerns and support; the antivaxx community is vociferous in opposing the vaccine, spreading misinformation, spinning conspiracies and whipping hysteria. Significant hesitation about the safety of the Covid-19 vaccine is also expressed in particular because of its rapid deployment. The pro-vaxx community counters this opposition by pointing to prior successes of immunizations as well as by mocking the anti-vaxx attitudes [8].

Extracted data from twitter and explored in three ways. Opinion mining reveals both the concern and support. A comparison of the social features of the anti-vaxx and pro-vaxx tweets suggests that the anti-vaxx community has gained steam on social media platforms and is better connected than the pro-vaxx community. Identifying and labeling tweets that sow discordant information is one way to prevent their spread, which is facilitated by our classification framework that can distinguish between the anti-vaxx and provaxx tweets with an accuracy of over 80% [9].

The time period is of the essence because during this time, a measles outbreak was prevalent throughout the United States and a public debate was raging. Sentiment analysis is applied to the sample, clustering the data into topics using the term frequency–inverse document frequency (TF-IDF) technique. The analyses suggest that most (about 77%) of the tweets focused on the search for new/better vaccines for diseases such as the Ebola virus, human papillomavirus (HPV), and the flu [11].

3. METHODOLOGY

This methodology part was used a dataset from open-source which consist of all the posted tweets on twitter. The dataset had all the tweets about COVID-19 vaccine from all over the world. Each and every tweet has some sentiments; it may be Positive, Negative and Neutral [5]. For Sentiment analysis and time series analysis there are several approaches that can be apply. The first task was to find the polarity, subjectivity score and find the sentiments using text and date attributes [1]. This analysis helped us to get an idea of what impact has the COVID-19 vaccine has made in the world. As Fig. 1 show, it extracted the data from Kaggle called vaccination tweets. After that it created a python script to get the sentiment polarity from the tweets. TextBlob library has been used for the purpose of finding sentiments. TextBlob provides an Application Programming Interface (API) for Natural Language Processing (NLP) tasks [3]. This was used to achieve the sentiment polarity values. Now, the value of sentiment polarity may be zero, positive or negative number. For our project, we needed only positive or negative sentiments and hence the tweets that have zero sentiment polarity were not taken into consideration [4].

3.1 Time Series Analysis

For the next component of our paper is to find any trend with respect to time. That have been done by time series analysis to find how polarity scores and subjectivity score varied, on average, over time of 1 year. It can also hypothesize that the discourse on twitter optimistic prior to February, it has been used time series analysis to find tweet counts as well as sentiments of those tweets which was given by users over the period of time. Steps to analyze the tweets about vaccination process going on in our country as follows:

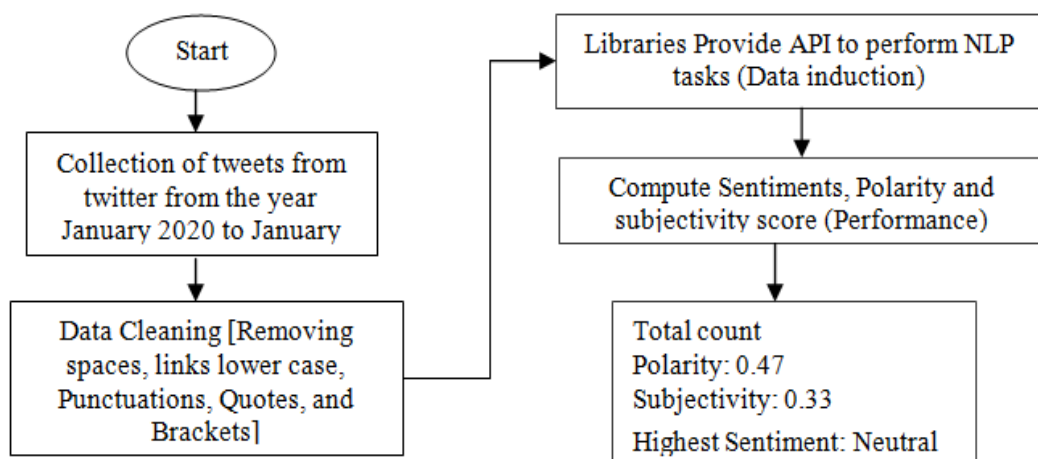


Fig. 1: Steps to compute Polarity and Subjectivity

Step1: Analysis of tweet count and plot a graph shows that how many tweets are commented about vaccination from January to July 2021.

Step 2: According to the tweet count we found sentiments like positive, negative and neutral. Also using total number of count and got polarity, subjectivity, how many re-tweets were there?

3.2 Sentiment Analysis Using ML Algorithms:

The first approach was used to found the polarity, subjectivity score, total number of tweet count and sentiments using time series analysis. Fig. 2 shows the workflow of proposed system using Machine Learning Algorithms [5]. We had used the train test split function, and split the data into train and test to 75% and 25% respectively. The extracted data was unstructured or raw data so it was mandatory to perform pre-processing before giving the dataset as an input to the ML algorithms [9]. The data consist of the text, hash_tags, and etc. columns which had some links, punctuations, brackets, extra spaces and more. This data was not required for the algorithm, so to clean all this symbols; it had to use the basic expressions that consisted of all the above mention garbage values [6]. To remove this garbage had to add some nltk packages to the text column of the dataset. After cleaning of the data. It has to made groups of words with semantic meaning and for this tokenization function was required. And also removed stop words using this tokenization process [7]. Stop words are the words which occur frequently in the tweets and have no meaning. After done with this entire thing we applied ML algorithms namely Multinomial Naïve Bayes (MNB), Logistic Regression (LR) and Support Vector Machine (SVM) [5].

3.2.1. Multinomial Naïve Bayes: Multinomial Naïve Bayes is an algorithm used in classification problems. It is formed by Bayes ‘Probability’ Theorem. This algorithm primary application will be text classification. It normally requires integer feature counts [2]. We had implemented Multinomial (MNB) for sentiment analysis. It had developed generic text for sentiment analysis and found number of examples and probability of those examples. Also got number of occurrences of each word and created a set for them [4]. There are many types of Multinomial Naïve Bayes algorithm namely Complement Naïve Bayes (CNB), Bernoulli Naïve Bayes (BNB), Gaussians Naïve Bayes (GNB). It had implemented Complement and Bernoulli and compare the evaluation parameters of this algorithm.

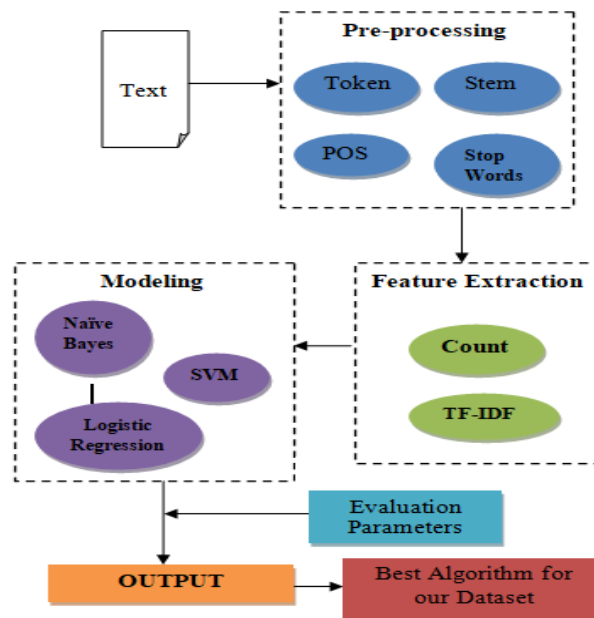


Fig. 2 Workflow Diagram of Proposed System using ML Algorithm

- (a) Complement Naïve Bayes: This approach is almost similar to multinomial, using this we counted frequently occurrence of word from the text of tweets of the COVID-19 Vaccination [5].
- (b) Bernoulli Naïve Bayes: This naïve bayes approach is different from other approaches. It takes Boolean value as an input instead of taking set of frequencies. It gives number of non-occurrence of the words (Words those are absents or not frequently occur) [5].
- (c) Gaussian Naïve Bayes: It is used to build on assumptions of normal distribution of probability. That means, spam or non-spam classes of tweets have frequencies distributed by the Gaussian principle [5].

3.2.2. Logistic Regression: This algorithm is used to find the relationship between binary variable and nominal variables. In our case the sentiments are the binary variable and text column is nominal attributes. LR gives predicted output, if we wanted sentiments like positive, negative or neutral. LR used the sigmoid curve which had a range from $-\infty$ to $+\infty$. LR took [6] the threshold value from the range and got the output as YES or NO. If the value of threshold is 0.5 it will give answer as YES, If the threshold value below 0.5 it will give answer NO.

3.2.3. Support Vector Machine: Support vector Machine (SVM) is a supervised algorithm used in classification problems. It prefers many times as it produces significant accuracy with less computational power. The main purpose behind this algorithm is to create a boundary which is used to separate n-dimensional spaces into classes [5].

Evaluation Parameters: In the present paper, the performance of multiple classical machine learning and deep learning algorithms has been evaluated based on the following widely used metrics: Accuracy, Precision, Recall and F-score [8]. The total observation of tweets is defined as TP, TN, FP and FN refers to true positive, true negative, false positive and false negative respectively.

- Accuracy: Accuracy is defined as the percentage of tweets that are labeled correctly [9]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

- Precision: Precision measures the percentage of the tweets that are actually commented out of all the tweets that are predicted as commented [9]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall: Recall measures how many of the commented tweets are actually labeled as commented [9]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F-score: F-score seeks a balance between Precision and Recall [9]:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. EXPERIMENTAL SETUP

The data has been used is called ‘All COVID-19 Vaccine Tweets’ from kaggle dataset. The data was extracted using python library called ‘Tweepy’ which allows users to access the twitter API if they have successfully created twitter developer account and all the required credential with them. From dataset we can get the main attributes namely date, text, re-tweets, source, hash tags, favorites, etc. We first found polarity and subjectivity using python libraries and also compare the evaluation parameters using machine learning algorithms. We have only one table in this dataset. This table consists of records about the users who has account on twitter and express their feelings regarding vaccination process, as well as information about vaccines, tweets, comments, and date.

Experiment 1 was performed to determine the optimal value of Polarity and subjectivity scores from the tweets extracted from twitter using the libraries TF-IDF Vectorizer and TextBlob. The range of the score was 0 to ± 0.2 for both polarity as well as subjectivity.

Experiment 2 was performed to analyzed the sentiments using polarity and subjectivity scores using the text and date columns.

Experiment 3 was performed evaluate machine leaning algorithms, the highest accuracy was computed by the Logistic Regression (LR) as compare to all other algorithms which had used.

5. RESULTS AND DISCUSSION

This section represents results for the techniques used to find polarity, subjectivity score and some sentiments like positive, negative and neutral of the tweets which were commented by used about COVID-19 Vaccination process during the month from January 2021 to July 2021. Also we got the evaluation parameters using different types of machine learning algorithms namely Naïve Bayes, Logistic Regression and Support Vector Machine on the tweets from twitter.

Table-1 Performance of Polarity and Subjectivity

	Text	Polarity	Subjectivity
0	Fully vaccinated & it feels GREAT!!! 🍌🍌 #C...	1.0	0.75
1	I have been Modern-ised. Excellent organization...	1.0	1.00
2	Best Islamic Laws\n#MondayMotivation\n#CGBudge...	1.0	0.30
3	Best Islamic Laws\n#MondayMotivation\n#CGBudge...	1.0	0.30
4	very happy in this moment #moderna https://t.c...	1.0	1.00
5	The best illustration ever.\n#CyrilRamaphosa #...	1.0	0.30
6	This is my I've just had my COVID19 Vaccination...	1.0	1.00
7	All that you need to know about #Covaxin, an o...	1.0	0.78
8	Perfect 🍌 #corona #coronavirus #vaccine #Vaccine...	1.0	1.00
9	When AI and Blockchain meets for the best #IBM...	1.0	0.30

Table-1 show the polarity and subjectivity score for each and every tweet which was commented. Polarity and Subjectivity score for total tweet count represented below by using graph.

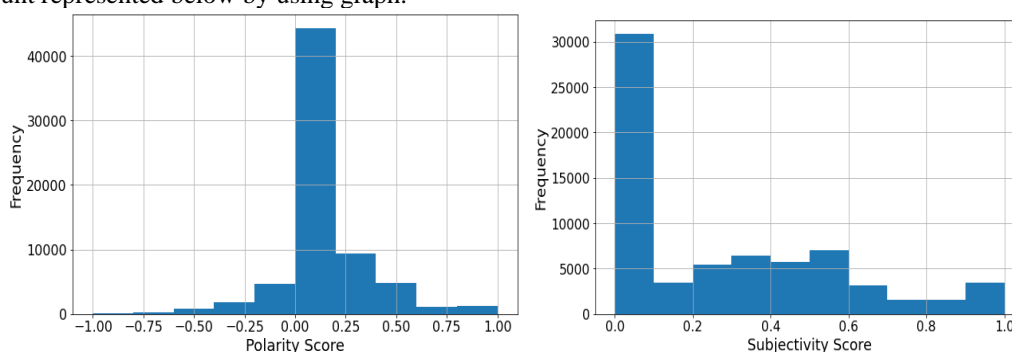


Fig. 3 Polarity and Subjectivity Score of total tweet count

Fig. 3 shows the representation of graph between polarity score and frequency i.e. total tweets counts. As per the figure the polarity score for the range 0 to 0.25 was the highest tweet counts. And it was showing the relationship between the subjectivity score on x-axis and frequency i.e. total tweets count and the highest subjectivity score for range 0 to 0.1 was the highest tweet count.

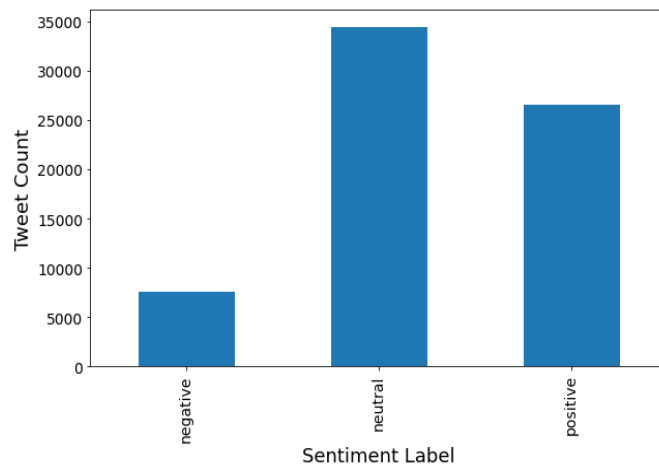


Fig. 4 Average sentiments count

Fig. 4 shows the analysis of total tweets counts with some sentiments like positive, negative and neutral. The representation shows total positive, negative and neutral sentiments according to the tweet counts from January 2021 to July 2021.

5.1 Sentiment Analysis Monthly

By observation of **Fig. 5** there was an analysis of positive and negative sentiments in the form of percentage from the month of December 2020 to June 2021. While observing it had clearly shown when the vaccination process was started people were excited about the vaccine, so positive sentiment was higher and negative sentiment was lower in the month of December, January. After that the trend of the positive tweets was decrease and negative trend was increased because people were scared about the outcomes after taking the vaccine. From April the positive tweets went raised gradually and negative sentiments felt. The highest percentage was 44.10% in the month of June 2021.

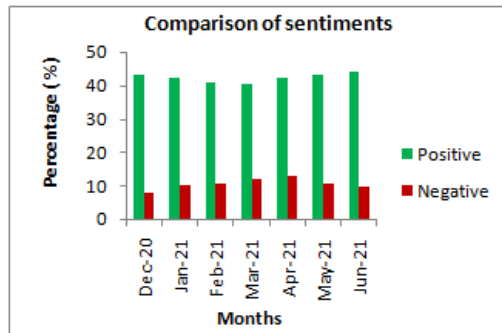


Fig. 5: Comparisons of Sentiments from Tweets

5.2 Sentiment Analysis using ML Algorithms

Also, it has used different ML algorithms for sentiment analysis. As shown in the **Fig 6**, it was found that LR model gives the highest accuracy of 90% among all other algorithms which was used. Closer to it was the SVM result with an accuracy of 89%. There was successfully found every model that accuracy above 75%. After calculating accuracy it has also found precision, recall and f1-score measurements of the algorithms. As shown in the figure it had found precision, recall and f1-score for the LR model was 89%, 55% and 89% respectively. The precision, recall and f1-score values for the SVM were almost similar to LR model i.e. 85%, 60% and 82% respectively.

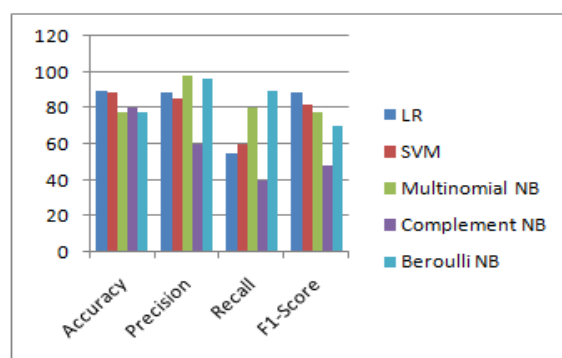


Fig. 6 Performances of Machine Learning Algorithms

It had been used three Naïve Bayes algorithms and found evaluation parameters for them. The accuracies of multinomial naïve bayes, complement naïve bayes and bernoulli naïve bayes were 78%, 80% and 78% respectively. Among all the algorithms of ML which we used Logistic regression proved a better ML algorithm.

Sr. No.	Algorithm	Accuracy	Precision	Recall	F1-Score
1	LR	90.00%	0.89	0.55	0.89
2	SVM	89.00%	0.85	0.6	0.82
3	Multinom	78.00%	0.98	0.8	0.78
4	Complem	80.00%	0.6	0.4	0.48
5	Bernoulli	78.00%	0.96	0.9	0.7

Fig. 7 Comparison table of Machine Learning Algorithms

Fig. 7 shows the comparison table of evaluation parameters of all machine learning algorithms and highest value of the parameters highlighted by blue color of any three algorithms.

6. CONCLUSIONS

We present the study that done the work on sentiment analysis of COVID-19 tweets on vaccine from twitter using TextBlob library and machine learning algorithms. We have analyzed the sentiments of people and found polarity and subjectivity score since the vaccination process was initiated. For this purpose, we used python library to find sentiments, polarity and subjectivity of all the tweets. Also we have used different ML algorithms LR, MNB, CNB, BNB and SVM. The LR model had given highest accuracy of 90%, after LR model SVM is the model which had given accuracy of 89%. And the naïve bayes algorithms namely Multinomial Naïve bayes, Complement Naïve Bayes and Bernoulli Naïve Bayes have given the accuracy of 78%, 80% and 78% respectively. For this vaccination data LR model was the great model which had given highest accuracy as well as other parameters. It conclude that LR model will give good results in future for any predictions. Also comparison has done between positive and negative sentiments of total tweet count. In the month of December, January the vaccination process was started and people were excited to take vacchin, After users view of the tweet count went towards negative sentiment because the outcome of the vaccine. In the month of May, June were the trend of total tweet count again goes high towards the positive side. 44.10 % was the tweet count of Positive sentiment in the month of June 2021 which was the highest percentage of positive tweet count of vaccination.

7. RESEARCH GAP

In future attempt, instead of using these algorithm there are many algorithm which gives better accuracy as well as success rate of the vaccination process and sentiment analysis of the country for different vaccines. And find sentiments, scores of polarity and subjectivity of tweets about donation for vaccination from Celebrities, government, different organizations and commomn people.

8. REFERENCES

- [1] A. Khakharia, V. Shah, and P. Gupta, "Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning," *SSRN Electronic Journal*, 2021.
- [2] A. D. Dubey, "Public Sentiment Analysis of COVID-19 Vaccination Drive in India," *SSRN Electronic Journal*, 2021.
- [3] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, 2021.
- [4] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanas, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month Following the First Vaccine Announcement," *IEEE Access*, vol. 9, pp. 33203–33223, 2021.
- [5] N. Paul and S. S. Gokhale, "Analysis and Classification of Vaccine Dialogue in the Coronavirus Era," *2020 IEEE International Conference on Big Data (Big Data)*, 2020.
- [6] R. F. Sear, N. Velasquez, R. Leahy, N. Johnson Restrepo, S. E. Oud, N. Gabriel, Y. Lupu, and N. F. Johnson. "Quantifying Covid-19 Content in the Online Health Opinion War using Machine Learning". *IEEE Access*, 8:91886–91893, May 2020.
- [7] V. Raghupathi, J. Ren, and W. Raghupathi, "Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, p. 3464, 2020.
- [8] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanas, D. S. Gherai, and F. Tajariol, "The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month Following the First Vaccine Announcement," *IEEE Access*, vol. 9, pp. 33203–33223, 2021.
- [9] N. Paul and S. S. Gokhale, "Analysis and Classification of Vaccine Dialogue in the Coronavirus Era," *2020 IEEE International Conference on Big Data (Big Data)*, 2020.
- [10] R. F. Sear, N. Velasquez, R. Leahy, N. Johnson Restrepo, S. E. Oud, N. Gabriel, Y. Lupu, and N. F. Johnson. "Quantifying Covid-19 Content in the Online Health Opinion War using Machine Learning". *IEEE Access*, 8:91886–91893, May 2020.