**Analyzing Public Sentiment Towards COVID-19 Vaccines: Insights From Twitter Opinions**

R P Sankararaman¹, S Kalaimaran2

1,2 Department of Computer Science Engineering, National Institute of Technology Puducherry, Karaikal, India​ - 609609

**Abstract**:The tumultuous COVID-19 pandemic led to rapid vaccine development and deployment. Understanding public opinion on these vaccines is crucial to public health as well as the existing policies. This study examines Twitter data to gain insights into the public opinion and sentiments regarding COVID-19 vaccines. We investigate a large volume of tweets concerning COVID-19 vaccines with sentiment analysis and natural language processing techniques, categorizing them as positive, neutral, or negative sentiments. We use different ML techniques including Logistic Regression, Naive Bayes, and Support Vector Machine to get the best possible results. The proposed system will show the changes in public opinion and sentiment over time. This study sheds light on public perception of COVID-19 vaccines, providing a data-driven foundation for informed decision-making and communication strategies during public health crises.

**Keywords:** Logistic Regression, Machine Learning, Naive Bayes, Natural Language Processing, Sentiment Analysis Technique, Support Vector Machine

**1. Introduction**

This study delves into the vast landscape of opinions on Twitter and sentiments surrounding COVID-19 vaccines, providing important details about prevalent attitudes and viewpoints.

We examine a large amount of data collected from Twitter related to COVID-19 vaccines using NLP and techniques of sentiment analysis. These tweets are systematically classified as positive sentiments, negative sentiments, or neutral sentiments, allowing for a comprehensive evaluation of the ever-changing public opinion on vaccines.

Machine Learning (ML) is commonly used to analyze data and build models to predict future outcomes. Basic concepts of ML include supervised learning and unsupervised learning, with the use of supervised machine learning in this paper. Furthermore, the capabilities of various techniques of machine learning, i.e. LR , SVM , and NB, are tested to extract deeper insights and improve our understanding of sentiment dynamics. The purpose of this project paper is to give people an insight about the sentiments of netizens (Twitter) on covid 19 vaccines using emerging machine learning algorithms. We have solved the problems faced by us while implementing the models which played a major role in increasing the accuracy of the predictions made by the models. The main features of the projects are analysis of sentiments of netizens during the pandemic into positive, negative and neutral and analysis of user influence which determines how many percent of the tweets are tweeted from user verified accounts as these verified accounts can only be seen with celebrities , Vips and visualizations of the results have been done and has been attached. This section provides an overview of the project. Section 2 contains a literature review of all reference papers. The methodology of the project is discussed in section 3. The project's outcomes are discussed in section 4. Section 5 discusses the project's conclusion as well as its future scope, and Section 6 contains the references that were used in the project.

**2. Literature Review**

This section provides us an overview of reference papers which have been used as reference in the project. We have discussed the Ml models which have been used by the authors of the respective papers and how they performed the analysis of sentiments. [[1]](https://doi.org/10.1007/s12652-022-03805-0) aims to analyse sentiments of different people throughout the pandemic by combining sentiment analysis and algorithms for NLP. The ML Models that were used to analyse the sentiments were (1) linear classifiers (SVM and NN ), probabilistic classifiers (NB (Naive bayes) and Bayesian Network), and decision tree classifiers(2) Lb (lexicon based) way which includes a corpus-based approach (Statistical and Semantic) and a dictionary-based approach(3) a hybrid way of classifier model which combines the previous two approaches[[1]](https://doi.org/10.1007/s12652-022-03805-0).

By Aasif Ahmad Mir and Rathinam Sevukan illustrates the Indian people's attitudes toward 'Covid-19 vaccines,' gains some insights into overall public communication about the topic, and adds to the existing literature. It can help health policymakers and administrators understand the polarity (positive, negative, and neutral) of Tweets about Covid-19 vaccines on Twitter in order to garner public awareness regarding vaccine misinformation and health concerns.

To determine the Tweets’ sentiments, the 'Valence Aware Dictionary for sentiment Reasoning' (VADER) was used. Using this tool, each word in the lexicon is automatically classified as positive or negative or neutral[[2]](https://doi.org/10.1177/01655515221118049).

Used Pandas library to convert the dataset into data frames, simplifying data manipulation and analysis within the proposed system that also used the Nltk which is a natural language toolkit. This computational package focusing on linguistics offers simple interfaces to numerous lexical resources. NLTK was used for text pre-processing which includes stemming, stopword removal, lemmatization, tokenization, Lexicon-based Analysis of emotions , and analysis of sentiments (NLTK Sentiment Intensity Analyzer). Matplotlib was used to generate interactive and elaborate data visualisations[[3]](https://doi.org/10.1016/j.matpr.2022.04.809).

By Md Tarique Jamal Ansari and Naseem Ahmad Khan employs Tweepy, a Python tool that allows users to browse the Twitter API after creating a Twitter Developer profile and obtaining access credentials, was used in this research study. Sentiment Analysis was carried out using the TextBlob library after the data was available. As per the TextBlob documentation, TextBlob employs the Naive Bayes (NB) categorization model. On NLTK (Natural Language ToolKit), NB classification has been taught to recognize the valence of collected tweets. The data was classified as positive sentiment and negative sentiment[[4]](https://doi.org/10.29333/ejgm/11316).

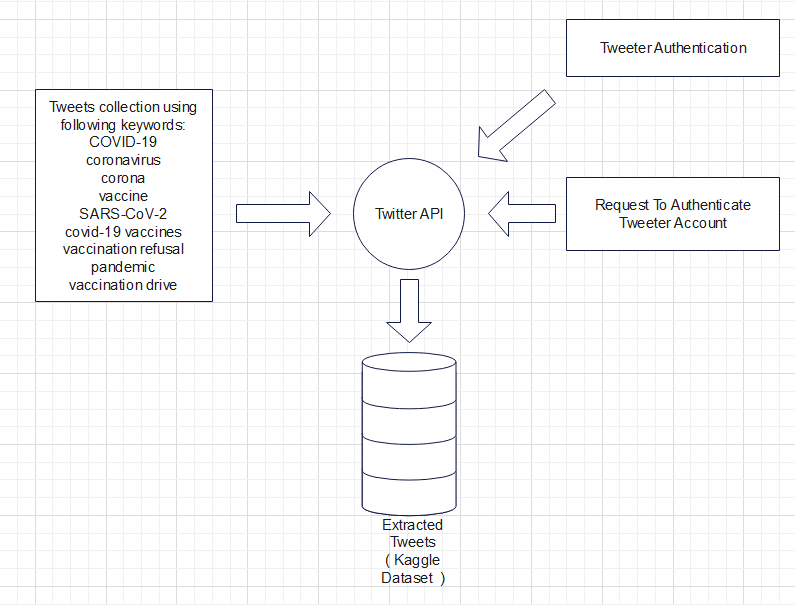
[5] used an approach involving purposive sampling to select all Tweets containing defined hashtags (e.g., #2019nCoV) related to COVID-19 on Twitter, observational study design Natural language processing methods to identify key COVID-19 topics and terms were used. The Twitter data mining strategy included data preparation as well as data analysis. Data preparation consisted of three steps: (1) sampling, (2) data collection, and (3) raw data pre-processing. The data analysis stage after pre-processing the raw dataset has (1) unsupervised machine learning, (2) qualitative methods, and (3) sentiment analysis[[5]](https://doi.org/10.1371/journal.pone.0239441).

By Teagen Nabity-Grover, Christy M.K. Cheung, Jason Bennett Thatcher [6] provides a report on the concept of online self-disclosure and summarises relevant literature for studying this construct during a pandemic. To explain self-disclosure online from the netizens during the pandemic, The concepts of self-focus and other focus has been used in this project[[6]](https://doi.org/10.1016/j.ijinfomgt.2020.102188).

We were able to find a few problems while performing the literature review and those are Contextual Understanding: All three algorithms NB , SVM , and LR struggled to capture language and context nuances. Tweets frequently contain slang, sarcasm, and cultural references which could have been challenging for these models to interpret accurately, reducing sentiment analysis precision. Data Quality and Noise: Tweets can be noisy, with spelling mistakes, abbreviations, and incomplete sentences. NB, SVM, and LR struggled with noisy data, resulting in misclassification or biased sentiment analysis results. Generalization to New Hashtags: Because the algorithms were trained on specific COVID-19 vaccine-related hashtags, the classifier models may not generalize or scale well to new hashtags or emerging trends in vaccine-related discussions.

**3. Proposed Methodology**

*3.1. Objective-*

The primary aim of this study is to create a predictive model for categorizing tweets about COVID-19 vaccines into three major sentiment categories: positive, neutral, and negative. Through various methodologies and approaches, the study aims to enhance the accuracy of these sentiment predictions. The paper aims to contribute to more precise and reliable sentiment classification by addressing the challenges and issues associated with sentiment analysis in the context of covid 19 vaccine-related tweets. This proposal, shown in Figure 1, is compulsory for the better understanding of public sentiment, facilitating data-driven decision-making, and increasing the overall effectiveness of the sentiment analysis models which were used, ultimately assisting in the evaluation of public perceptions and attitudes regarding COVID-19 vaccines.

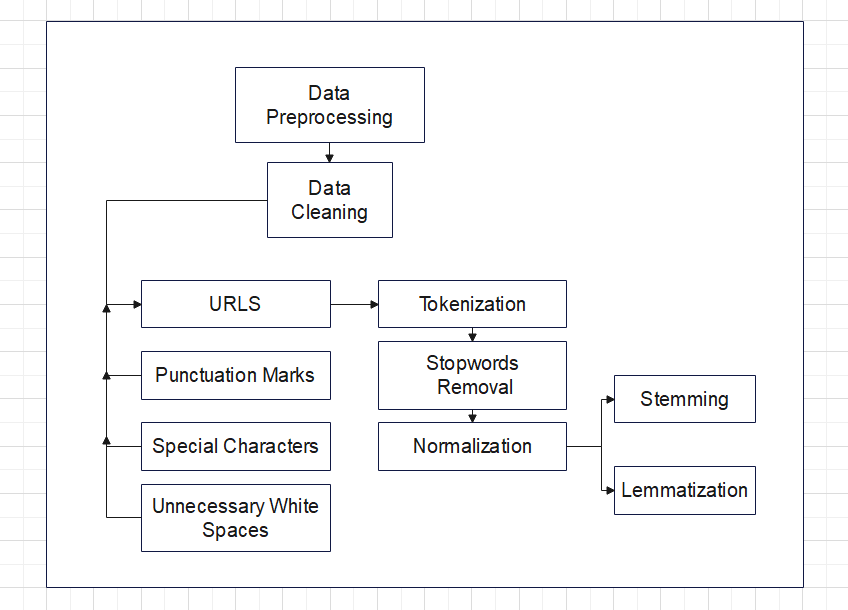
*Figure 1*: Flowchart of the process of data extraction through Tweepy.

*3.2. Data Collection-*

Tweets (opinions) about covid-19 vaccines were downloaded from kaggle. The dataset which was used was extracted for the project from twitter using tweepy (with Twitter API )[[4]](https://doi.org/10.29333/ejgm/11316). A set of hashtags were given to extract the tweets related to Covid-19. Then, Twitter authentication was requested and was given. And the required data was extracted and was stored as a dataset (Figure 1). The below given figure is flowchart of the process of data extraction through Tweepy:

*3.3. Data Preprocessing-*

The duplicated data of the opinions is being removed because it has an effect on the results. To remove unwanted and abnormal data, basic data cleaning is performed, which includes the removal of Urls, punctuation marks, special characters, and unnecessary white spaces. Tokenization, stopword removal, and normalization (stemming and lemmatization) are then carried out[[1]](https://doi.org/10.1007/s12652-022-03805-0) (Figure 2). The below given figure is a pictorial representation of the data preprocessing process:



*Figure 2*: Representation of the data preprocessing process.

*3.4. Feature Selection-*

All the unique words which were used in the trained data, excluding stop words, are chosen as features to convert the opinions into numerical scores. Because the format is text data, considering the unique words as features is preferable. In order for the model to accurately predict the sentiment related to Twitter user’s opinions.

*3.5. Model Selection-*

The steps given below is to investigate various models for categorizing twitter opinions:

Step 1: Loading the dataset and importing the libraries required for classification using machine learning techniques.

Step 2: The pre-processed twitter opinion dataset is then read and classified.

Step 3: To accurately identify opinions belonging to a specific class, different ML classifiers are implemented on the text data.

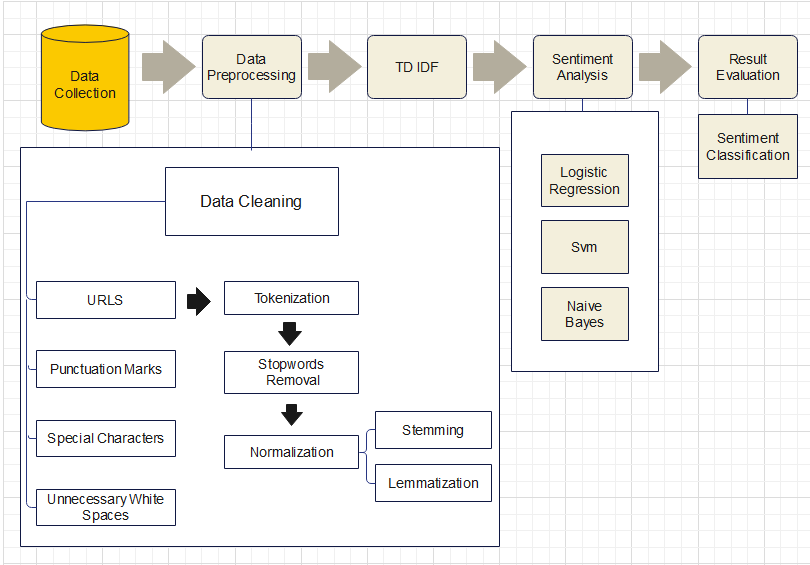
Step 4: To implement the classification models, the appropriate library modules are loaded.

Step 5: To ensure that the models are trained on a sufficient amount of data and evaluated on a different set of data, the dataset is divided into a training set (80%) and testing set (20%).

Step 6: The classifiers that were chosen are then applied to the dataset to classify opinions and obtain performance metrics such as F1 Score, recall, accuracy, and precision.

Step 7: The performance measures for each classifier are calculated and compared to determine the most effective approach for the analysis of sentiment..

'Logistic Regression'[[9]](https://doi.org/10.1016/j.eswa.2022.118715), 'Support Vector Machine'[[8]](https://doi.org/10.11591/eei.v12i3.4830) and 'Naive Bayes'[[7]](https://doi.org/10.1007/978-981-19-5443-6_27) models are employed in converting Twitter opinions into scores (sentimental analysis). To evaluate their performance in sentiment analysis, these models give us accuracies of each model and classification reports (Figure 3). The below given figure is the system architecture of the project, It explains the processes/methodology of the project:



*Figure 3:* Architectural view of the methodology.

*3.6. Problems Faced While Using The Classifier Models-*

Contextual Understanding: All three algorithms NB , SVM , and LR struggled to capture language and context nuances. Tweets frequently contain slang, sarcasm, and cultural references which could have been challenging for these models to interpret accurately, reducing sentiment analysis precision. Data Quality and Noise: Tweets can be noisy, with spelling mistakes, abbreviations, and incomplete sentences. NB, SVM, and LR struggled with noisy data, resulting in misclassification or biased sentiment analysis results. Generalization to New Hashtags: Because the algorithms were trained on specific COVID-19 vaccine-related hashtags, the classifier models may not generalize or scale well to new hashtags or emerging trends in vaccine-related discussions.

*3.7. Experimental Setup-*

The extracted Twitter data is converted into a sentimental score using pre-trained classifier algorithms such as Logistic regression, Svm, and Naive Bayes. The extracted data was cleaned up by removing urls, punctuation, special characters, and unnecessary white spaces, followed by tokenization, stopword removal, and normalization. The dataset is being divided into two sections 80% training and 20% testing sections. The model’s accuracy helped assess the prediction capability of the classifier models.

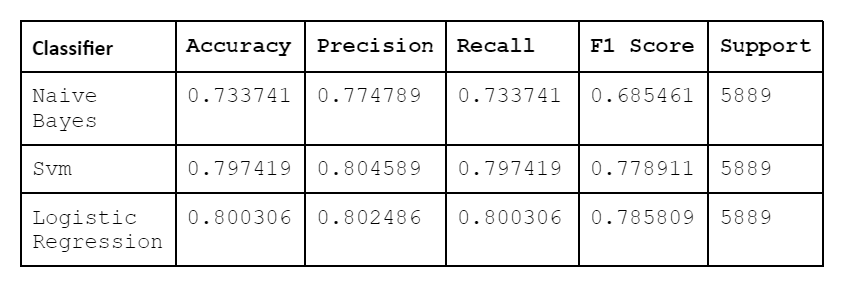
3.8. Time Frame-

Because simple ML algorithms are used, training time for all the models is nearly zero. The prediction of sentiments and the accuracies of all three classifier models took a maximum of 5 minutes.

**4. Results**

*4.1. Evaluation Metrics-*

If We Compare The Three Classifier Models We Can Conclude That ‘Logistic Regression’ Classifier Model Has The Highest Accuracy With Recall And F1 Score. But , It Has Precision Lesser Than SVM (Table 1). The below given table is the tabular visualization of evaluation metrics:



*Table 1:*  Tabular visualization of evaluation metrics.

Where ,

TP - True Positive.

TN - True Negative.

FP - False Positive.

FN - False Negative.

*4.2. Solutions For The Problems Given In Proposed Methodology-*

To tokenize text, remove stopwords, and preprocess tweets, we've introduced a custom text preprocessing function and a custom transformer class called TextPreprocessor, which can help handle slang and cultural references better. This text preprocessing is now used by the code to better handle noise in Twitter data. For each classifier, we build pipelines that combine text preprocessing, vectorization, and classification into a single process. Both the figures which are given below discuss the accuracy of the classification models which were used in the analysis of sentiments before the problems which were found were solved and after the problems which were found were solved(Figure 4 and Figure 5). The below given figures are the comparison of accuracies before the problems were solved and after the problems were solved:

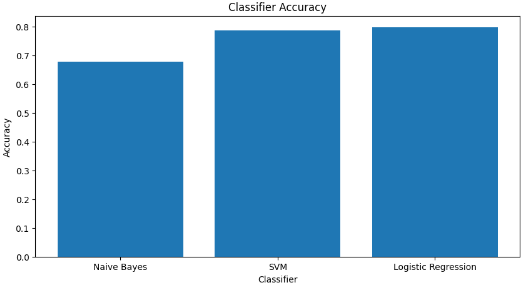


Figure 4: Accuracies before the problems were solved.

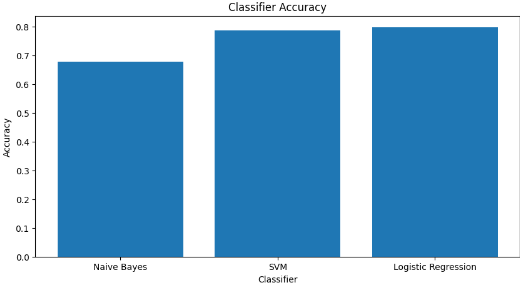
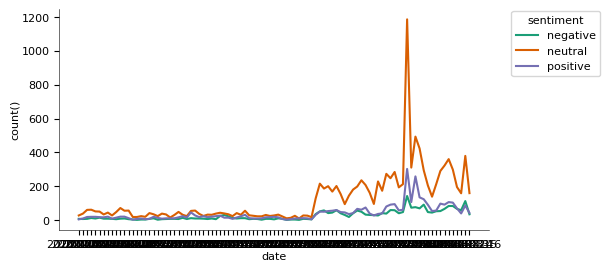


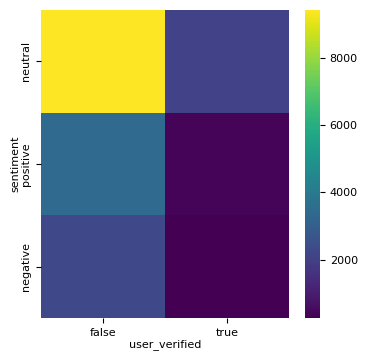
Figure 5: Accuracies before the problems were solved.

*4.3. Results-*

The output of this proposed model of analysis is the predicted sentiments for the twitter opinions. The accuracy for the predicted sentiments is also found. When the three classifier models are compared, we can say that the 'Logistic Regression' classifier model’s accuracy is the highest in relation to recall and F1 score. However, its precision is lower than that of the SVM (Figure 6 and Figure 7). The below given figures represent the sentiments of the opinions of unverified users and verified users*.* In Figure 6 , the purple line depicts the positive sentiments , the orange line depicts the neutral sentiments and the green line depicts the negative sentiments and In Figure 7 , the positive sentiment , neutral sentiment and negative sentiments are represented with various colors based on the counts of sentiment.

**

*Figure 6: Sentiments Of unverified users in twitter.*

**

*Figure 7: Sentiments of verified users in twitter.*

**5. Conclusion\**

Within this paper, We used three distinct classifier models to perform sentiment analysis of twitter opinions (covid-19 vaccines). Logistic Regression was the most effective classifier model as the accuracy of sentiment analysis of logistic regression was the highest. The project shows that many people across the globe have shown a neutral sentiment, many people across the globe have shown positive sentiment and few people have shown a negative sentiment towards the covid-19 vaccines. This project can be helpful as a reference in future if we are in a pandemic situation again as many people around the globe are showing hesitation towards vaccines.

**References**

[1] Sarirete, A. Sentiment analysis tracking of COVID-19 vaccine through tweets. *J Ambient Intell Human Comput* (2022). <https://doi.org/10.1007/s12652-022-03805-0>

[2] Mir, A. A., & Sevukan, R. (2022). Sentiment analysis of Indian Tweets about Covid-19 vaccines. Journal of Information Science, 0(0). <https://doi.org/10.1177/01655515221118049>

[3] P. Chinnasamy, V. Suresh, K. Ramprathap, B. Jency A. Jebamani, K. Srinivas Rao, M. Shiva Kranthi, COVID-19 vaccine sentiment analysis using public opinions on Twitter ,Materials Today:Proceedings,Volume 64, Part 1,2022

<https://doi.org/10.1016/j.matpr.2022.04.809>.

[4] Ansari MTJ, Khan NA. Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content. Electron J Gen Med. 2021;18(6):em329. <https://doi.org/10.29333/ejgm/11316>

[5] Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T (2020) Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLoS ONE 15(9): e0239441. <https://doi.org/10.1371/journal.pone.0239441>

[6] Teagen Nabity-Grover, Christy M.K. Cheung, Jason Bennett Thatcher,Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media,International Journal of Information Management,Volume 55,2020

<https://doi.org/10.1016/j.ijinfomgt.2020.102188>.

[7] Gaur, P., Vashistha, S., Jha, P. (2023). Twitter Sentiment Analysis Using Naive Bayes-Based Machine Learning Technique. In: Shakya, S., Du, KComputingL., Ntalianis, K. (eds) Sentiment Analysis and Deep Learning. Advances in Intelligent Systems and , vol 1432. Springer, Singapore. <https://doi.org/10.1007/978-981-19-5443-6_27>

[8] Improving sentiment reviews classification performance using support vector machine-fuzzy matching algorithm Vivine Nur Cahyawati, Zuriani Mustaffa <https://doi.org/10.11591/eei.v12i3.4830>

[9] Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, Paul Cotae, Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset,Expert Systems with Applications,Volume 212,2023, <https://doi.org/10.1016/j.eswa.2022.118715>.