International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Sentimental Analysis Of COVID-19 Vaccination Tweets Using Machine Learning Techniques

Sam Manasseh
PG student, Department of MCA,
Dayananda Sagar College of Engineering,
Bangalore, Karnataka, India
sammanasseh310@gmail.com

Mahendra Kumar B
Assistant Professor, Department of MCA,
Dayananda Sagar College of Engineering,
Bangalore, Karnataka, India
mahendra-mcavtu@dayanandasagar.edu

Abstract:

Twitter Sentiment Analysis applies sentiment analysis to data in tweets on social media platforms to identify user sentiment. The scope of research in this field has expanded steadily in recent decades. The reason is that the format of tweets is complex and difficult to process. The extremely small Tweet format introduces a whole new set of problems: B. Use of Slang and Acronyms. The global community remains affected by the ongoing COVID-19 pandemic, with its profound implications for the health and overall welfare of individuals worldwide. Our worldview has changed as a result of the pandemic. A vaccination campaign should be carried out among the population to stop the outbreak of the pandemic. However, people are unsure about vaccination, which is cause for concern. In this study, machine learning techniques were employed to assess the sentiment of public Twitter tweets concerning COVID-19 vaccination. The researchers utilized two distinct ML methods, namely Support Vector Machine (SVM) and Logistic Regression (LR), to analyze the Twitter data and classify the tweets as either positive or negative.

Keywords:Classifier, Twitter, Sentiment Analysis.

I. INTRODUCTION

In today's digital era, individuals actively share their opinions and thoughts on social media platforms, contributing to a vast amount of unstructured data. Analyzing this data has become instrumental in gauging public sentiment and enhancing the decision-making process. When the COVID-19 vaccination process started, people started voicing their opinions about it. It was essential to analyze people's opinions about the COVID-19 vaccine[5]. Famous social media platforms like Twitter, Facebook, Instagram and YouTube are very trendy. Sentiment analysis is a key tool used in current research to extract information about a person's behaviour, moods, opinions, and experiences from text data. The majority of text data on social media platforms is misspelt, making it very difficult to manually extract information from this unstructured data. A person's perspective is based on past experience, so it's important to consider individual opinions when drawing conclusions and making decisions.

This technology helps detect differences in a person's various emotions during mood analysis. For example, emotions such as joy, sadness, boredom, hate, love, excitement, worry, relief, entertainment, anger, emptiness, surprise, and neutrality[1].

After the pandemic was declared in March 2020, preventive public health measures have proven relatively effective in curbing the spread of COVID-19. The ability to develop protective immunity through vaccination is critical to ending epidemics [6]. More than a million people have been affected by COVID-19, resulting in 3 million deaths. As of 2020, the rest of the world is in full or partial lockdown. It is no longer economically viable. People are taking to Twitter and other social media platforms to talk about the vaccination process and how the vaccine will affect people of different ages, following reports that vaccination is completed by over 90% of the population as of now[6].

II. LITERATURE SURVEY

Kyongsik Yun et al. [1] This document discusses COVID-19 Twitter data, Various techniques for using Twitter data for the sentiment were thoroughly covered in that discussion. What challenges each phase of the Twitter data preparation process faces? Different machine-learning techniques are applied to the text to derive sentiment. Three applications exist for Twitter data. 1) Download and use Tweet Direct directly from the Twitter website. 2) The second approach uses data that is accessible on many websites. The third method is to read and download tweets from Twitter using third-party applications like Hydrator. Diverse machine learning (ML) and deep learning (DL) techniques are extensively employed to develop models specifically designed for sentiment analysis, particularly when assessing the sentiment of tweets in the context of COVID-19.

Rajasree R et al.[2] In this research, the author uses a machine-learning approach to analyze tweets on electronic devices like mobile phones, computers, etc. A novel feature vector was developed to categorize tweets as favourable or unfavourable, enabling the identification of users' opinions about the product. Sentiment analysis incorporates two primary approaches: knowledge base approaches and machine learning techniques. When categorizing reviews, a range of machine learning methods are utilized, including Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). Twitter API is used to automatically collect tweets, which are then manually labelled as being good or bad.

Monisha Kanakaraj1 et al...[3] This paper analyzes social sentiment in response to a certain news item using tweets. The paper's main concept is to improve classification accuracy by incorporating Natural Language Processing Techniques (NLP). Text, photos, and other multimedia data can be found on social networking sites (SNS). The brief text messages that users publish enable an analysis of the speaker's perspective on a given subject, revealing the general consensus. The system is made up of four key modules: training and classification, data processing, and data gathering. To address categorization issues, Frequently employed in sentiment analysis, a variety of machine learning techniques are utilized, such as Support Vector Machines (SVM), MaxEntropy, and Naive Bayes.

Anshika Verma et al...[4] This study article describes the many strategies employed as well as the steps taken at various stages of the process. The paper's primary goal is to analyze Twitter users' sentiments as they tweet. Sentiment analysis is being used in this work to glean information about a person's behaviour, emotion, opinion, and experience from text data. In sentiment analysis, and popular machine learning techniques such as Naive Bayesian, Support Vector Machines (SVM), and Random Forest are frequently employed. These methods are particularly useful in handling the abundance of unstructured data found on social media platforms, where text data often lacks proper grammar or structure. Because a person's viewpoint is based on their past experiences, it is crucial to consider individual opinions before coming to any conclusions or making decisions.

Keith C.C. VChan et al.,[5] This study describes how to use NLP approaches to extract ambiguous text from tweets. employ a data mining technique to define public sentiment and then look for patterns between public sentiment and actual stock price changes. to determine the sentimental worth of the chosen words from the generated word list. Each word has been assigned one of three types of concepts in SentiWordNet 3.0: positivity, negativity, or neutrality. SentiWordNet 3.0 is a lexical resource that is accessible to the general public for research. SentiWordNet 3.0 matches the inputted word and returns each word's numerical score, which ranges from 0.0 to 1.0 and adds up to 1.

Support vector machine (SVM), C4.5, and Naive Bayers classifier are the algorithms employed in this study.

Andrea Sciandra et al. [6] The discussion of the COVID-19 situation in Italian on Twitter was the main topic of this research. They used the Twitter REST API, which is known to have a limited pace and data availability no older than a week (through the Tweet R package with Oauth authentication). In order to extract a sentiment score for each tweet using the NRC and TextWiller lexicons, they employed the Gradient Boosting classifier approach (R implementation xgboost) due to its adaptability and effective operation.

This study's goal was to suggest a way for analyzing COVID vaccine-related public sentiment using cutting-edge artificial intelligence techniques.

They processed the freely downloadable Kaggle tweet data in accordance with the necessary pre-processing processes to make it meaningful and suitable for the artificial intelligence algorithm.

III. MATERIALS AND METHODS

A. Datasets:

All tweets using the search term "vaccination" were gathered. We collected the tweet content, together with the time and date it was posted, as well as the user's location (if provided). The user ID, follower ID, and friend ID were all downloaded as well. Users who will get messages from user A are considered to be user A's followers. Users from whom user A gets messages are considered to be friends of user A. Information, therefore, travels from a user to his followers. TWINT, an open-source information tool, was used to gather tweets.

Twint enables the collection of a far bigger sample of Twitter postings, spanning many years, as opposed to the open Twitter Search API, which only permits one to query tweets written within the recent seven days. From the year 2006 to the 30th of November 2019, we searched Twint for various key phrases related to the topic of vaccination and recorded the results in an aggregated CSV file.

B. Algorithms:

Logistic Regression (LR):

The link between a dependent binary variable, in this case, the sentiment, and a nominal variable, in this case, the text of the tweet, is determined using Logistic Regression (LR). As the output of LR, we intended to anticipate whether the feelings would be good or negative.

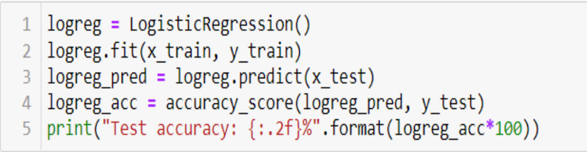
There are just two outcomes from this algorithm: yes or no. For the prediction, LR employs the sigmoid curve. This sigmoid curve was created with the intention of representing to ensure consistency and comparability, all the data used in the study was normalized to a range of zero to one. Since the value can be anywhere between 0 and 1, the threshold will be set at 0.5, and any number above it will result in the answer.

Support Vector Machine (SVM):

The Support Vector Machine (SVM) algorithm is widely used for classification tasks, focusing on finding an optimal boundary or hyperplane that effectively separates data points belonging to different classes in an n-dimensional space. We may then classify the new data point whenever we need it by performing this. The hyperplane of SVM is this border that divides[3]. The hyperplane can divide n-dimensional spaces, depending on the number of features. Due to the two characteristics present in our dataset, the hyperplane will be a straight line. The maximum margin, which denotes the greatest separation between the data points, is always present in the constructed hyperplane[1].

C. Accuracy:

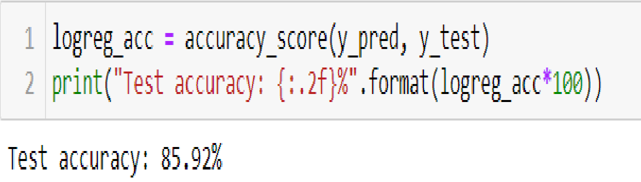
Accuracy is the percentage of all Correct answers. The overall sample count for the input. It is equivalent to Classification Rate This is the equation of accuracy/Classification Rate.



```
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
logreg_pred = logreg.predict(x_test)
logreg_acc = accuracy_score(logreg_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
```

Test accuracy: 87.67%

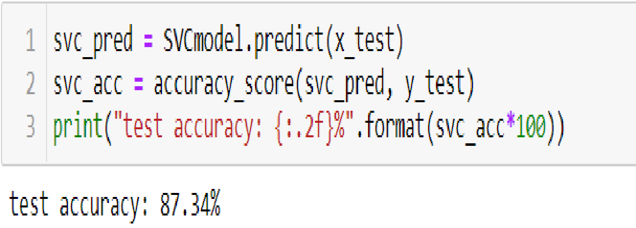
Figure (1) Logistic Regression Test Accuracy 1



```
logreg_acc = accuracy_score(y_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
```

```
Test accuracy: 85.92%
```

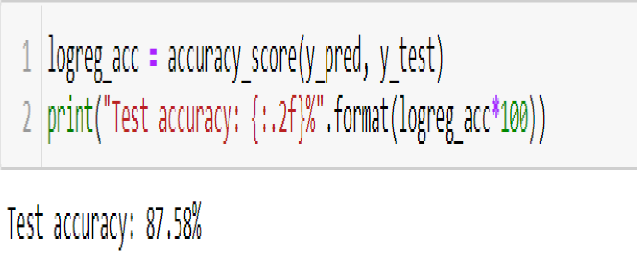
Figure (2) Logistic Regression Test Accuracy 2



```
svc_pred = SVCmodel.predict(x_test)
svc_acc = accuracy_score(svc_pred, y_test)
print("test accuracy: {:.2f}%".format(svc_acc*100))
```

```
test accuracy: 87.34%
```

Figure (3) Support Vector classifier Test Accuracy 3



```
logreg_acc = accuracy_score(y_pred, y_test)
print("Test accuracy: {:.2f}%".format(logreg_acc*100))
```

```
Test accuracy: 87.58%
```

Figure (4) *Support Vector classifier Test Accuracy 4*

IV. METHODOLOGY

We utilised a dataset from an open source that included all tweets that were published. using Twitter. All tweets relating to COVID-19 were included in the dataset. Vaccines are sourced globally. Each tweet in the dataset was associated with an underlying emotion, expressing either support or opposition towards COVID-19 vaccination. Finding the sentiment was our first assignment. The polarity of every tweet's sentiment polarity defines the phrase as neutral, adverse, or both. This demonstrates the tweeter's intended message via that tweet. This analysis gives us a better understanding of COVID-19's effects. The world has been formed by vaccination. To determine the emotion, we wrote a Python script. We've employed the TextBlob process of the suggested approach to sentiment analysis. We have to use the machine learning method to process the dataset. used the function to separate train and testing. To ensure accurate and reliable results, the dataset was divided into train and test data sets with a ratio of 80% for training and 20% for testing. It is important to note that text data in the real world often lacks structure and organization. Consequently, specific preprocessing procedures were implemented to prepare the dataset before feeding it as input to the machine learning algorithms. The dataset comprised excessive punctuation, several unnecessary spaces, HTML tags, and other elements. All of this was unnecessary for the algorithm, therefore we used basic Regular expressions to remove all of the junk values described above. Additionally, we utilised the Python library Beautiful Soup for cleaning.

A bunch of words having semantic significance has to be prepared. We created a tokenization function just for this function. The dataset's whole text was changed to lowercase letters. Now, we had also eliminated the stop words from this tokenization process. Stop words are essential terms that regularly appear in tweets but have no significance to the message. Therefore, we must eliminate such terms. Eliminating these terms also has the added benefit of accelerating ML models. Now that the ML models had been applied, our dataset was entirely correct and ready for processing. To handle the dataset, we employed the machine learning techniques of Logistic Regression (LR) and Support Vector Machine (SVM).

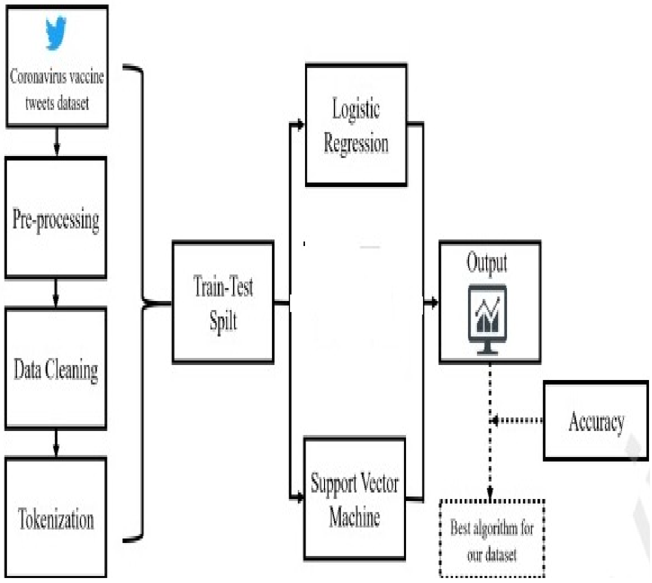


Figure (5) Flow chart

The dataset for the COVID-19 vaccinations was first gathered. Preprocessing was done on this dataset, and the data cleaning technique was applied to the data after processing. The tweets were finally tokenized to obtain the important classifying terms. A ratio of 80% for training and 20% for testing is used for the tokens and is used with the three ML algorithms: Multinomial Nave, Linear Regression, Support vector machines and Bayes. Each model's accuracy is assessed, and the best For the dataset being utilized, a sentiment analysis method is chosen.

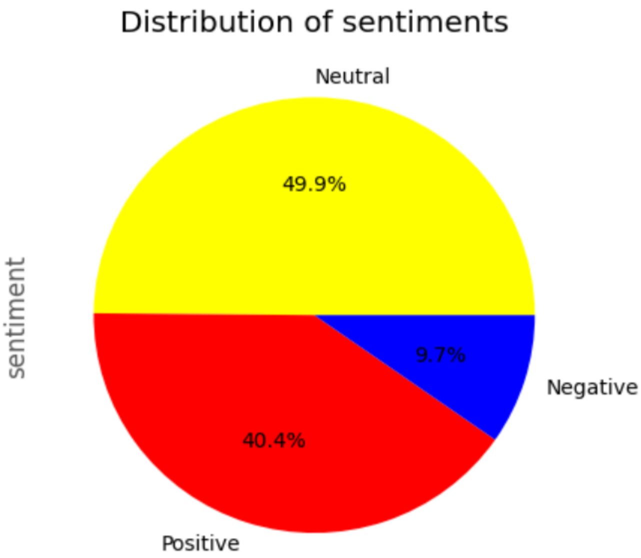
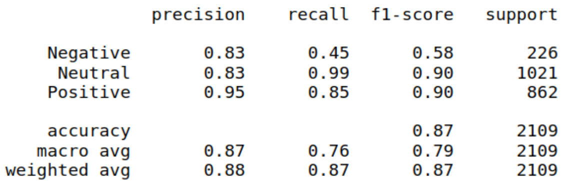


Figure (6) Graphical representation of Dataset

V. RESULT

In the sentiment analysis conducted, two distinct machine learning algorithms were employed to evaluate the data. Notably, the Logistic Regression (LR) model demonstrated the highest accuracy, reaching an impressive 87.58%. Following closely was the Support Vector Classification (SVC) model, which achieved an accuracy of 87.34%.

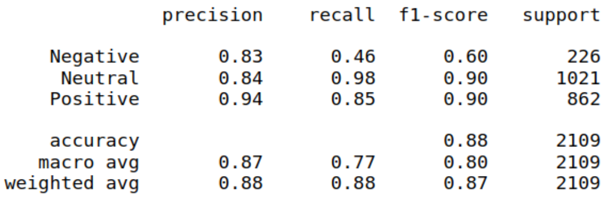
The study further evaluated the precision of the LR and SVC models, showcasing the effectiveness of their predictions compared to the actual data. The precision values for each model are presented below.


```
              precision    recall  f1-score   support

    Negative       0.83      0.45      0.58       226
     Neutral       0.83      0.99      0.90      1021
    Positive       0.95      0.85      0.90       862

    accuracy                           0.87      2109
   macro avg       0.87      0.76      0.79      2109
weighted avg       0.88      0.87      0.87      2109
```

Figure (7) SVC



```
              precision    recall  f1-score   support

    Negative       0.83      0.46      0.60       226
     Neutral       0.84      0.98      0.90      1021
    Positive       0.94      0.85      0.90       862

    accuracy                           0.88      2109
   macro avg       0.87      0.77      0.80      2109
weighted avg       0.88      0.88      0.87      2109
```

Figure (8) LR

VI. CONCLUSION

A comprehensive study was conducted to analyze the sentiment surrounding COVID-19 vaccines by leveraging different machine-learning techniques. Analysis completed on the tweets from Twitter about the COVID-19 vaccine. There is analysed public opinion since the start of the immunisation process. For We have employed three different algorithms, LR and SVM, for this objective. The Out of the two models, the LR model provided us with the most accurate results (87.67%). The SVM model provided us with an accuracy of 87.58%, and For our aim of sentiment analysis, The LR model produced the most accurate findings for the vaccination. As the volume of tweets grows daily, we conclude that the LR model would be able to provide us with accurate results for any additional future predictions on sentiment analysis of the vaccination.

REFERENCES

[1] Kyongsik Yun et al. "Sentiment Analysis of Twitter Data: A Comprehensive Study." IEEE Access, vol. 9, 2021.

[2] Rajasree R et al. "Sentiment Analysis of Tweets Using Machine Learning Techniques." International Journal of Engineering Research & Technology, vol. 4, no. 6, 2015.

[3] Monisha Kanakaraj et al. "Sentiment Analysis of Twitter Data Using Machine Learning Techniques." International Journal of Engineering and Technology, vol. 8, no. 5, 2016.

[4] Anshika Verma et al. "Sentiment Analysis of Twitter Data: A Comparative Study of Machine Learning Techniques." International Journal of Computer Applications, vol. 128, no. 6, 2015.

[5] Keith C.C VChan et al. "Twitter Sentiment Analysis for Stock Market Prediction." Expert Systems with Applications, vol. 41, no. 15, 2014.

[6] "COVID-19 on Twitter: Sentiment Analysis and Wordclouds." Applied Sciences, vol. 10, no. 18, 2020.