International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 7 Issue 5, September-October 2023 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

An Assessment of Sentiment Analysis of Covid-19 Tweets

Ms. Tanzeela Qureshi[1], Dr. Mohit Singh Tomar[2], Dr. Ritu Shrivastava[3]

[1]Research Scholar, [2]Associate Professor, [3]Head and Professor,
[1, 2, 3]Department of CSE, SIRT, Bhopal, Madhya Pradesh, India

ABSTRACT

Various rumors and assumptions have circulated about the COVID-19 immunization, making it a heated subject of discussion in India. This prompted a reaction from the country's populace, who During the course of favorable, negative, and neutral evaluations, tweets and retweets on twitter. The number of these tweets are a jumble of unstructured data. The goal of this study is to have the statistics justify feeling implied by it. The purpose of this study is to take advantage of twitter's massive data pool and extract insights that have the implications that can be drawn from it. Comprehensive research on the people's feelings may help us arrive at a fair familiarity with the population at large's point of view toward preventing disease by vaccination. Dataset taken into consideration for vaccination-related tweets are collected for study. From 2020 to 2021, including a data mining of 16,05,152 tweets related to vaccination.

KEYWORDS: *Sentiment Analysis, Twitter, Polarity, Subjectivity, Natural Language Processing (NLP)*

How to cite this paper: Ms. Tanzeela Qureshi | Dr. Mohit Singh Tomar | Dr. Ritu Shrivastava "An Assessment of Sentiment Analysis of Covid-19 Tweets" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-7 | Issue-5, October 2023, pp.534-543, URL: www.ijtsrd.com/papers/ijtsrd59976.pdf

IJTSRD59976

I. INTRODUCTION

There was global anarchy as a result of the COVID-19 epidemic, which ravaged every country on Earth. All hope hinged on the vaccine because of the mutative and aggressive character of the virus. Pfizer, Moardina, Covi Shield, and many other international corporations worked hard to develop an effective vaccine. In any case, the notion that adverse effects for vaccinations are unavoidable was not effectively absorbed by the general population, despite there being a clear majority of approval. Many people's opinions are influenced by what they read or hear in the mainstream and social media. Consequently, social media played a crucial role in communication and expression of thoughts about vaccines, with Twitter in particular playing a pivotal role due to its unique features that allow users to tweet (i.e., express an opinion), retweet (i.e., support an opinion), and extend comments and like to a wider audience.

With over 500 million tweets sent every day, Twitter is a treasure trove of information that may be mined for insights if used correctly. Many academic investigations have used Twitter data. Twitter was used as a platform for individuals in India to openly discuss the topic of vaccination via tweets, retweets, etc. Many insightful conclusions may be derived by analysing people's moods according on the content of their tweets on Twitter utilizing sentiment analysis technologies. Opinion mining for sentiment analysis is a data-analysis method that may establish whether the data is good, negative, or neutral.

Therefore, the purpose of this research is to provide substantial insights by analysing the mood of all tweets on vaccines. The goal of this study is to use Sentiment Analysis to do an exploratory data analysis of all tweets and Twitter data. The results of this study will provide light on how the general public feels about COVID-19 vaccinations.

The study is organized as follows, with Section 2 focusing on prior studies that are pertinent to the topic at hand. Sentiment analysis is defined and briefly discussed in Section 3. The dataset that was utilized for this analysis is described in great depth in Section 4. In Section 5, we detail all the findings from our exploratory data analysis of the dataset. Experiment findings, key insights, and future plans for this model are presented in Sections 6 and 7, respectively.

II. LITERATURE WORK

By using the capabilities of Natural language processing (NLP) to analyze the sentiment that is being expressed in the specific data, the notion of opinion mining or sentiment analysis has been employed and modified for diverse studies throughout time. Previous studies that have shed light on this topic are discussed below.

In the paper[1], the BERT model is used to do Sentiment Analysis on Twitter data. Tweets were geotagged in order to classify the data utilized in this work. The BERT model for emotion categorization was used to train the data, and the SVM classifier was used to assess the model's effectiveness. On the whole, the acquired data was accurate to within 4%. Paper [2] presents an analytical framework for impact of COVID-19 on the stock market based on tweets during the outbreak. Supervised learning was used to train this model, which achieved an accuracy of 86.24 percent. The studies were conducted after the Coronavirus epidemic to aid businesses in forecasting stock prices, identifying new marketing opportunities, and monitoring their own growth. In paper [3], we analyze what people were tweeting about most during and after the first outbreak of the COVID-19 pandemic. For topic extraction, we used Latent Dirichlet Allocation (LDA), and for sentiment analysis, we relied on a Lexicon-based strategy. This report does a good job of summing up the concerns of different groups during the early stages of the epidemic. Using a dataset of 600,000 English-language tweets, the model was trained using 80% of the data and then tested using 20% of the data. The article used sentiment analysis to illustrate people's thoughts on the most discussed issues.

This paper [4] examines tweets from across all of India's states during the months of November 2019 and May 2022. In this article, we successfully used sentiment analysis to the gathered information and found that, on the whole, Indians had an optimistic outlook on life.

There was a correlation between the number of confirmed cases of COVID19 in a given state and the number of tweets sent from that state. The research [5] provides a comprehensive analysis of the tone of all tweets related to COVID-19. In this case, we evaluated the tone of the tweets using Logistic Regression, VADER sentiment analysis, and BERT sentiment analysis. In order to analyze public opinion on the issue of Coronavirus, the authors of paper [6] combine data from two sources: the textual tweets posted in April 2020 from six nations and the tweets of top 10 politicians. In the end, the report presents findings that shed light on the similarities and variances in public opinion among nations. The results showed that across all six nations, respondents felt the most "trust," "fear," and "anticipation." Sentiment analysis utilizing word weighting TF-IDF and Logistic Regression was performed on the Twitter data from 30th April 2020 in article [7]. This algorithm successfully classified the sentiment of the tweets with an accuracy of 94.71%.

Our understanding of the prior studies in this area was much enhanced by this literature study. Our project's trajectory is now clearer thanks to this.

III. SENTIMENT ANALYSIS

An application of Natural Language Processing (NLP), sentiment analysis classifies data and texts to reveal how people feel about a topic [8]. This helps in understanding the author's intentions and point of view. This technique uses a scoring system that shows the true meaning and viewpoint of the text. We can more quickly identify positive, bad, and neutral aspects of the material by using these evaluations. Businesses regularly use opinion mining (or "Emotion AI") and sentiment analysis (or "sentiment analysis") to get insight into how customers and the wider public feel about a brand or product.

To gauge public opinion about COVID vaccinations, we use Sentiment Analysis to data gathered from Twitter after the second wave of Coronavirus. The study's findings may provide light on the public's thoughts and feelings towards COVID-19 vaccines.

There are two main phases to any sentiment analysis:

1. Prioritizing, sanitizing, and selecting features from datasets
2. Applying Sentiment Analysis to the Data

IV. DATASET DESCRIPTION

The project began off with information collection and classification. For this study, we analyzed data from the 'Covid-19 All Vaccine Tweets' collection. The data, which covers the period from December 2020 to August 2021 and consists of 80,418 tweets, was acquired from kaggle.com [9]. Table 1 lists the characteristics and provides explanations for each.

Table 1: Attributes of the dataset and their description

ATTRIBUTES	DESCRIPTION
id	This gives the id of the tweet
user_name	User name of the person who has tweeted
user_location	The location of the person who has sent the tweet
user_description	The Twitter bio of the person writing the tweet
user_created	When the Twitter account of the user was created
user_followers	Number of followers of the person sending the tweet
user_friends	Number of friends of the person sending the tweet
user_verified	Binary value specifying whether the user is verified on Twitter or not
date	Date and time when the tweet was sent
text	The text in the tweet as it is
hashtags	Specifies all the hashtags that were used in the tweet
source	Gives information about the source(device or application) from which the tweet was sent
retweets	Number of times the tweet was retweeted
favourites	Number of people who have marked the tweet as a 'favourite'
is_retweet	Tells us if the tweet is a retweet or a new one

Following this, the tweets in the dataset were cleaned up by removing things like mentions, hashtags, retweet information, and links. Time stamps for tweets were also eliminated since they were deemed unnecessary. Some of the most salient characteristics from the aforementioned list are chosen for exploratory research.

V. METHODOLOGY

This section explains in depth how Sentiment Analysis was carried out on the selected dataset.

Gathering information that may be used in the analysis was the first stage. The same was discussed at length in the preceding paragraph. The dataset contains clean, pre-processed data. Eighty three hundred and six records survived after duplicate columns were removed from the dataset. The tweets were then cleaned up by removing any traces of mentions, hashtags, retweets, links, etc. We also scrubbed the data for tweet timestamps. Then, a subset of the aforementioned traits was chosen since it was more relevant to the data analysis being conducted. A few key graphs were displayed after a graphical study of the data was performed.

The number of tweets sent from each device type is displayed in Fig. 1. The majority of tweets were sent from Android devices, followed by the Twitter Web App, and then the Cowin Vaccination Availability platform, as seen in the provided scatter plot.

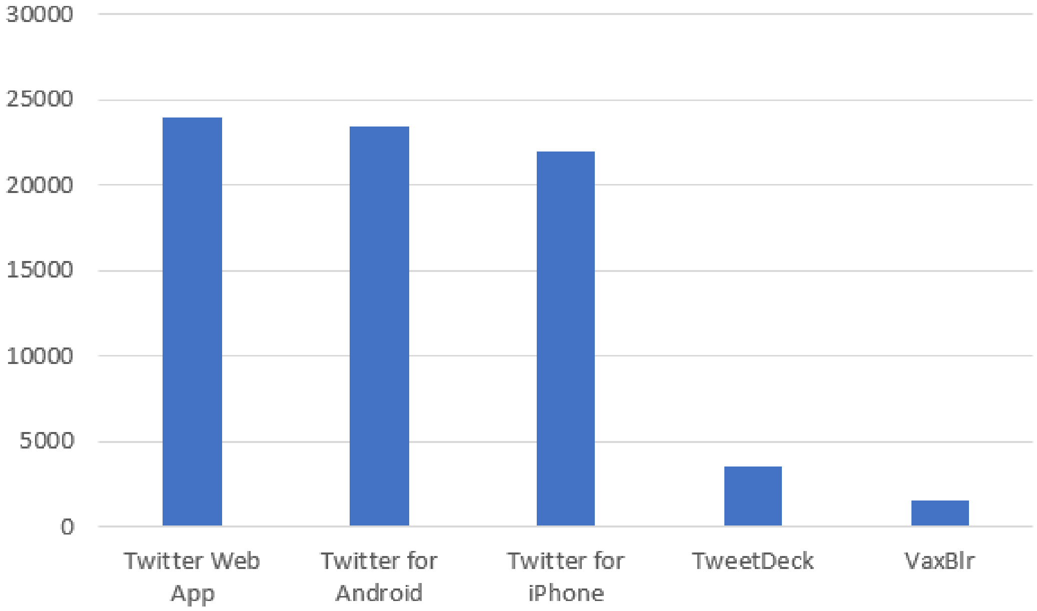


Figure 1: Plot showing the source of the tweets posted

Figure 2 displays the distribution of tweets between verified and unverified accounts. The narrative reveals that about 10% of tweets came from verified accounts, while the remaining 90% came from unverified ones.

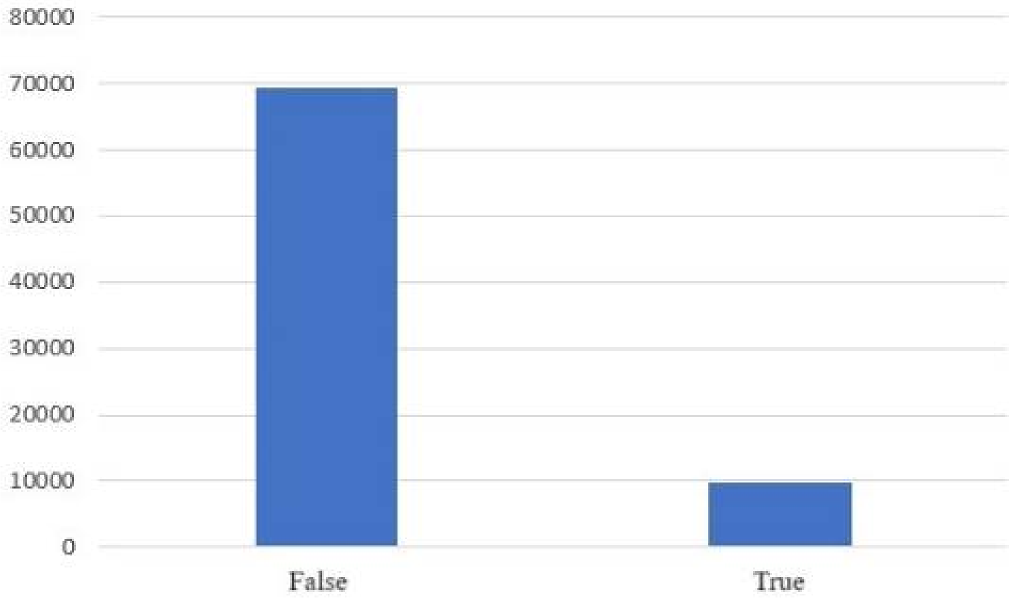
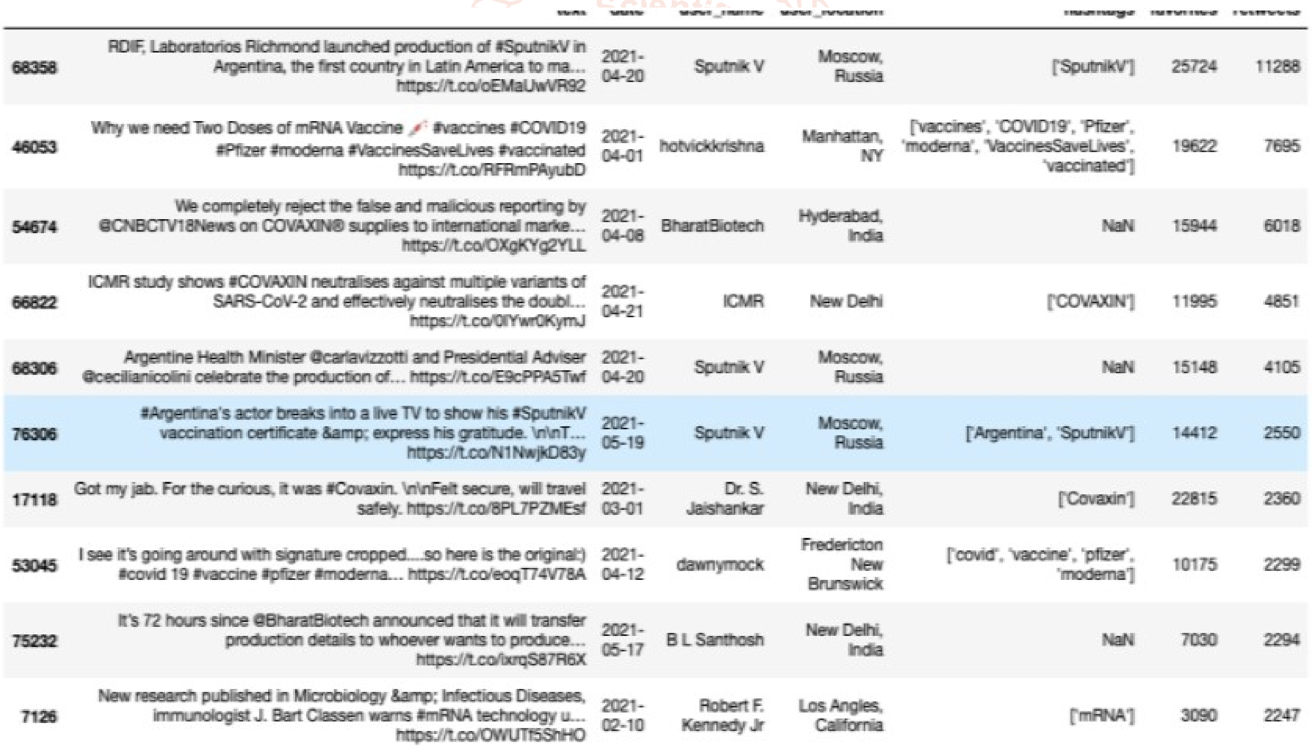


Figure 2: Plot showing the number of tweets sent from verified or unverified accounts

The most popular tweets about COVID-19 vaccinations were identified by taking the top 10 most retweeted tweets from the dataset. They look like Fig. 3 down below.



	text	date	user_name	user_location	hashtags	favorites	retweets
68358	RDIF, Laboratorios Richmond launched production of #SputnikV in Argentina, the first country in Latin America to ma... https://t.co/oEMaUwVR92	2021-04-20	Sputnik V	Moscow, Russia	['SputnikV']	25724	11288
46053	Why we need Two Doses of mRNA Vaccine 💉 #vaccines #COVID19 #Pfizer #moderna #VaccinesSaveLives #vaccinated https://t.co/RFRmPAyubD	2021-04-01	hotvickkrishna	Manhattan, NY	['vaccines', 'COVID19', 'Pfizer', 'moderna', 'VaccinesSaveLives', 'vaccinated']	19622	7695
54674	We completely reject the false and malicious reporting by @CNBCTV18News on COVAXIN® supplies to international marke... https://t.co/OXgKYg2YLL	2021-04-08	BharatBiotech	Hyderabad, India	NaN	15944	6018
66822	ICMR study shows #COVAXIN neutralises against multiple variants of SARS-CoV-2 and effectively neutralises the doubl... https://t.co/0IYwr0KymJ	2021-04-21	ICMR	New Delhi	['COVAXIN']	11995	4851
68306	Argentine Health Minister @carlavizzotti and Presidential Adviser @cecilianicolini celebrate the production of... https://t.co/E9cPPA5Twf	2021-04-20	Sputnik V	Moscow, Russia	NaN	15148	4105
76306	#Argentina's actor breaks into a live TV to show his #SputnikV vaccination certificate &amp; express his gratitude. \n\nT... https://t.co/N1NwjkD83y	2021-05-19	Sputnik V	Moscow, Russia	['Argentina', 'SputnikV']	14412	2550
17118	Got my jab. For the curious, it was #Covaxin. \n\nFelt secure, will travel safely. https://t.co/8PL7PZMEsf	2021-03-01	Dr. S. Jaishankar	New Delhi, India	['Covaxin']	22815	2360
53045	I see it's going around with signature cropped....so here is the original:) #covid 19 #vaccine #pfizer #moderna... https://t.co/eoqT74V78A	2021-04-12	dawnymock	Fredericton New Brunswick	['covid', 'vaccine', 'pfizer', 'moderna']	10175	2299
75232	It's 72 hours since @BharatBiotech announced that it will transfer production details to whoever wants to produce... https://t.co/lxrqS87R6X	2021-05-17	B L Santhosh	New Delhi, India	NaN	7030	2294
7126	New research published in Microbiology &amp; Infectious Diseases, immunologist J. Bart Classen warns #mRNA technology u... https://t.co/OWUTf5ShHO	2021-02-10	Robert F. Kennedy Jr	Los Angles, California	['mRNA']	3090	2247

Figure 3: Top 10 most retweeted tweets related to COVID-19 vaccine.

The top 20 accounts based on the frequency of the tweets were found out. They are as shown in Fig. 4 below.

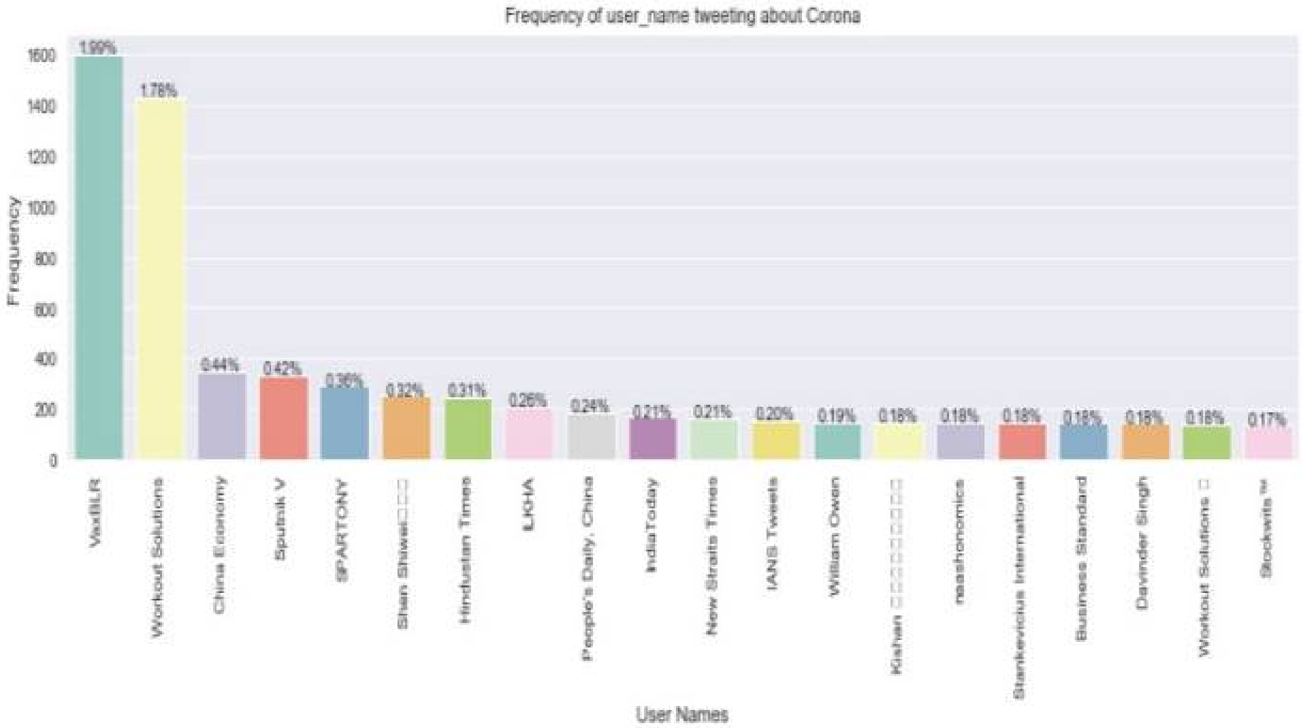


Figure 4: Top 20 accounts based on the frequency of the tweets

The information was divided into three groups, one for each polarity value. Tweets with polarity values between -1 and -0.01 were classified as negative. There are three types of tweets: positive (1), negative (-1) and neutral (0). Tweets with polarity values between -0.01 and 0.01 were labeled as "Neutral," while those with polarity values between 0.01 and 1 were labeled as "Positive." The number of tweets that fall into each of these three categories is shown in Fig. 5.

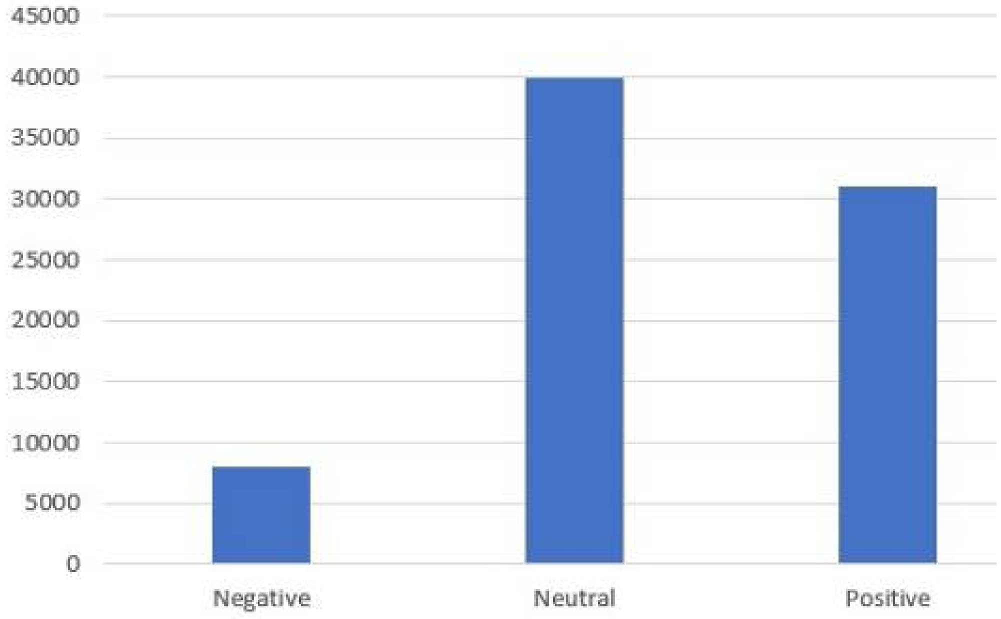
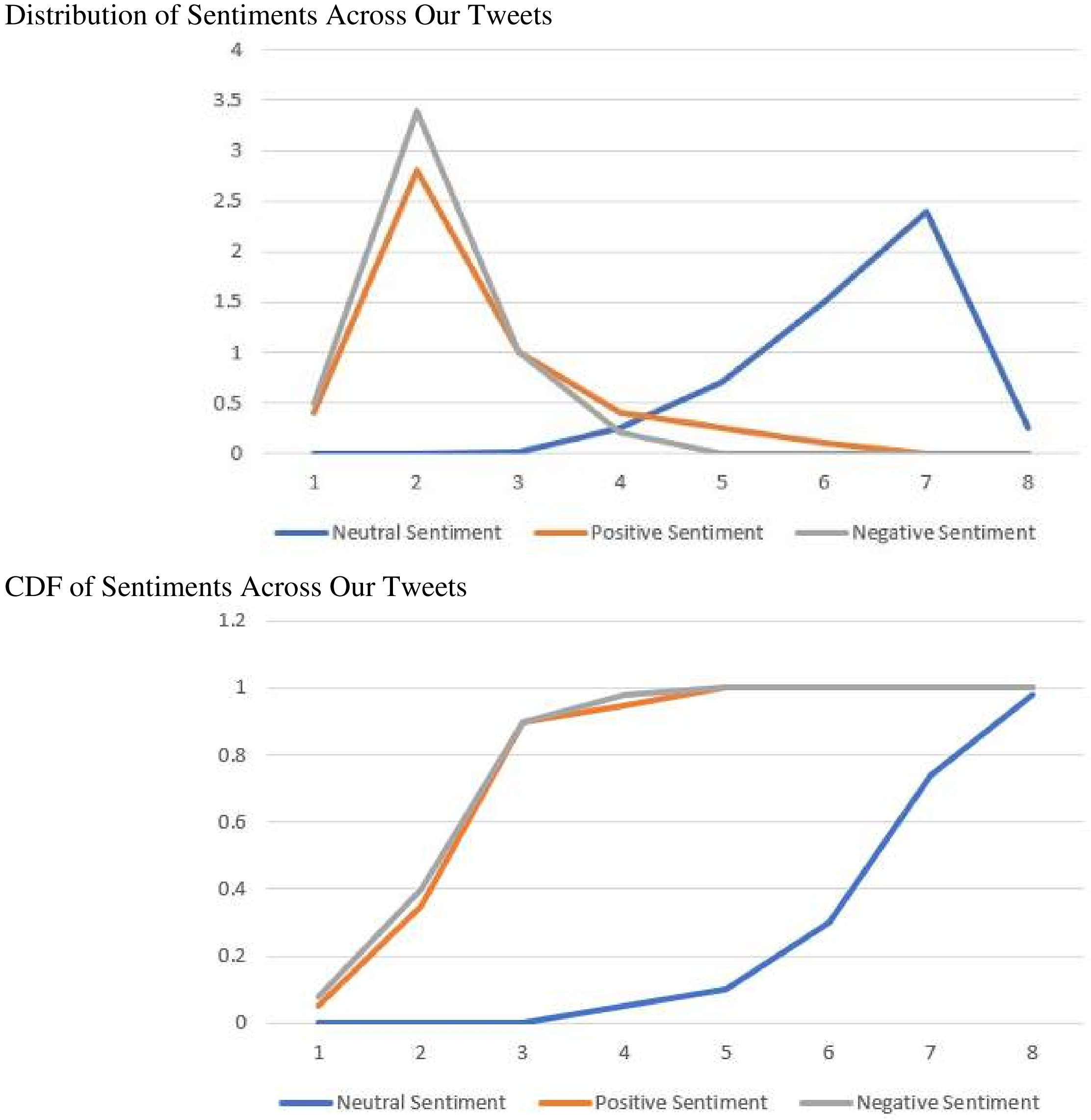


Figure 5:Tweet count given as positive, negative or neutral class

Figure 6 shows the CDF of tweet sentiments and the distribution of tweet sentiments throughout the sample.


Distribution of Sentiments Across Our Tweets

CDF of Sentiments Across Our Tweets

Figure 6: Distribution and CDF of Sentiments across tweets in the dataset

Next, we identified the most frequently used terms in both the most positive and negative tweets in the total dataset and created word clouds for them. Fig. 8 below illustrates this point.

The chart below shows the public's overwhelmingly favorable reaction to the COVID-19 vaccinations, with the most prevalent phrases being "good," "thank," "effective," "vaccinated," "great," "happy," "safe," etc. Emergency, forced, alone, Canada, halt, second, death, India, Ontario, etc. were other frequent terms in the unfavorable tweets.

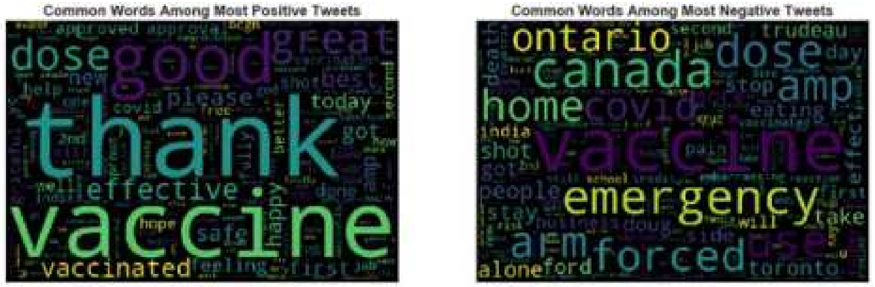


Figure 7: Word Clouds for the common words among the most positive and most negative tweets

To go further, we plotted word clouds from tweets about a select number of nations and places. Figures 8 and 9 and Figure 10 depict them below.

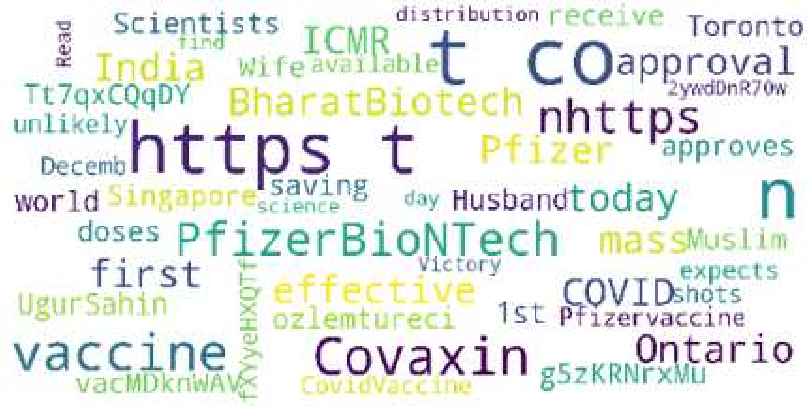


Figure 8: Most common words in tweets related to India

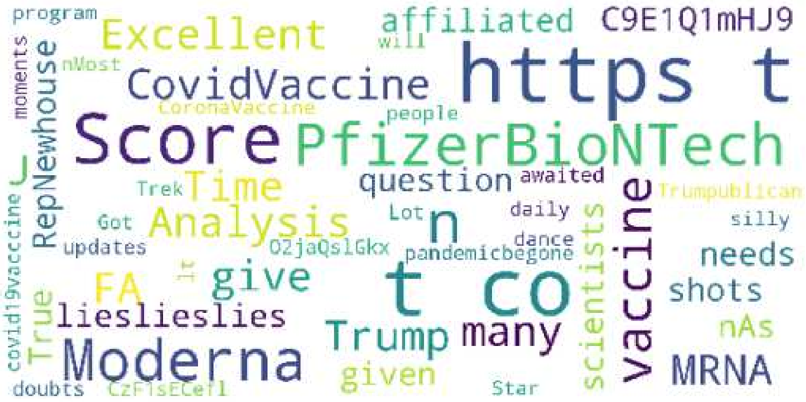


Figure 9: Most common words in tweets related to USA

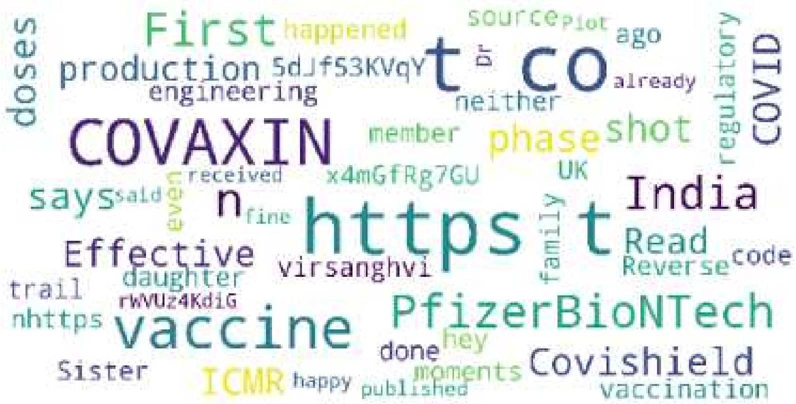


Figure 10: Most common words in tweets related to Mumbai

Due to the limited vaccination options in India, we created a word cloud from tweets about Covaxin and Covishield. Fig. 11 displays this.

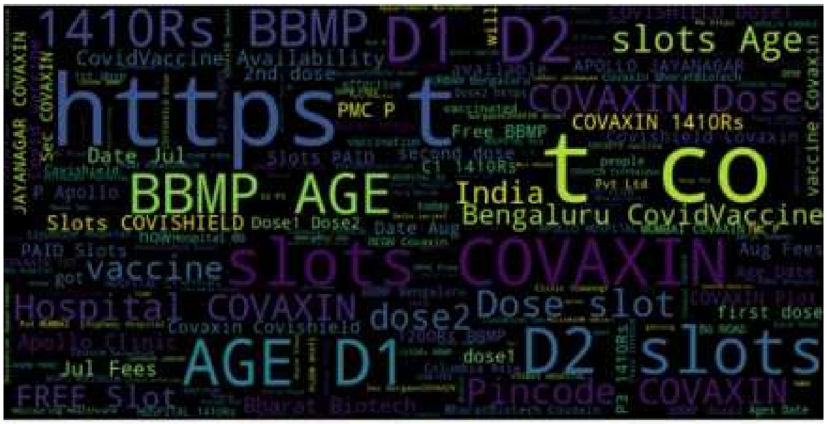


Figure 11: Word cloud for the tweets Covishield and Covaxin

The data was then processed using some sophisticated methods to generate the color-coded word clouds for the divided tweets. The information was once again scrubbed. The tweets were filtered to remove any grammatical or spelling errors or nonsense. After that, the whole Twitter data was classified into positive, negative, and neutral categories, and corresponding cloud words were formed for each. That's what Fig. 12 shows.

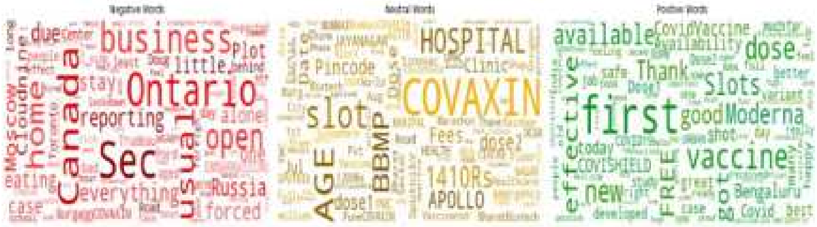


Figure 12: Colour-coded word clouds for all the tweets

Similarly, the colour-coded word clouds for the Covishield and Covaxin were also plotted which is shown in Fig. 13 below.

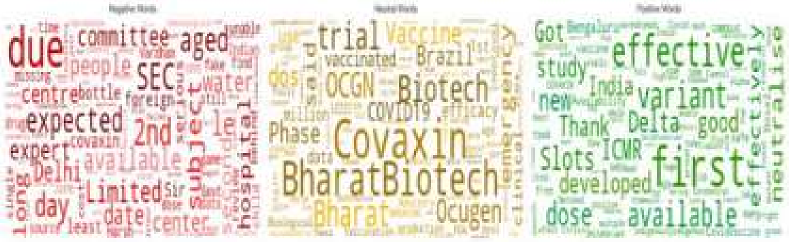


Figure 13: Colour-coded word clouds for the Covishield and Covaxin

A. Polarity

A word's polarity is the degree to which it expresses a negative, neutral, or positive emotion. Words with a positive polarity have a value of 1, whereas neutral words have no polarity and negative words have a value of -1. The polarity of a tweet is calculated by taking the mean of all the words in it, which is a float value between -1 and +1. Polarity of a tweet is a matrix that breaks down a tweet's emotional tone into positive, negative, and neutral categories.

B. Subjectivity

The ratio of subjective to objective details in a tweet or paragraph depends on the

speaker. The degree to which a writing is subjective rises as more private details are included and falls as more objective data is presented. It's a proxy for the author's degree of involvement in the tweet or other source content.

A small subset of the dataset's generalizations about vaccination was then chosen for testing of polarity and subjectivity in terms of sentiment. In Figures 4 and 5, we see the polarity and subjectivity ratings.

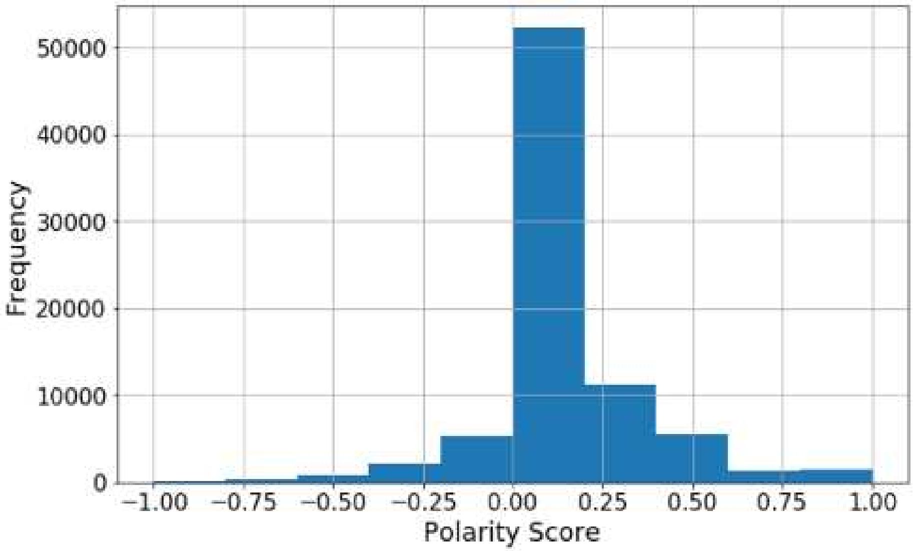


Figure 14: Polarity score of the tweets

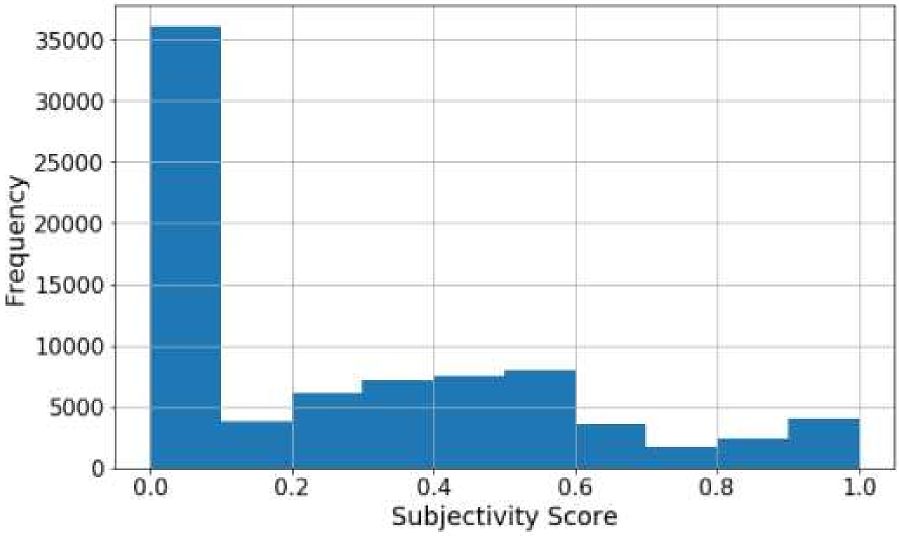


Figure 15: Subjectivity score of the tweets

The following is a flowchart depicting the research procedure that was used.

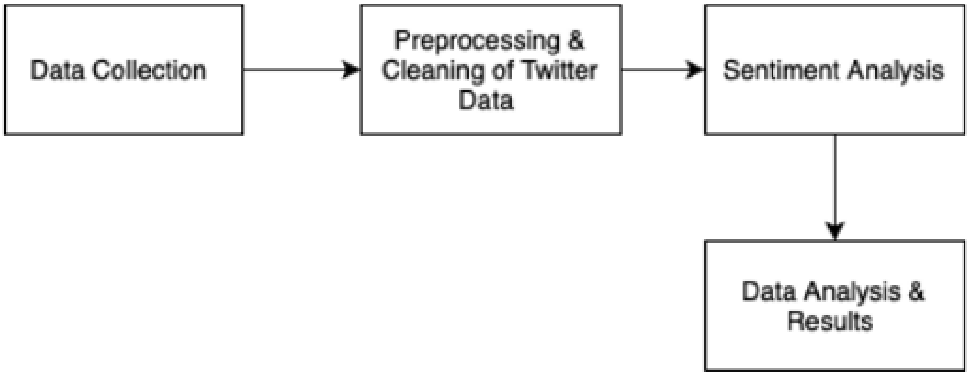


Figure 16: A flowchart of the methodology followed for implementation of this project

VI. RESULTS AND DISCUSSIONS

In this study, 16,05,152 tweets and retweets were analysed for attitude about immunization in India. After being cleaned, pre-processed, and having duplicates removed, the data utilized in the research was ready for analysis. The tweets had their timestamps, mentions, hashtags, retweets, and links deleted. Word clouds of varying permutations were extracted while also plotting crucial graphs like emotion labels graphs and distribution graphs, among others, to get the gist of the data. Additionally, a few general comments about vaccination were chosen and examined for polarity and subjectivity to determine how they were received. Data analysis shows that the majority of Indians have a favourable opinion about vaccination, but that there is still a strong negative attitude towards the practice in the country.

VII. CONCLUSION AND FUTURE SCOPE

The public has been affected in a variety of ways by the coronavirus, and this sort of study may aid government and other research organizations in comprehending public sentiment and filling in knowledge gaps. Twitter data analysis is crucial since it is a place where many individuals express their honest, sometimes controversial, opinions. With the help of Natural Language Processing (NLP) methods like subjectivity and polarity, the study analyses millions of public views expressed via tweets on the Twitter network and provides us with the necessary analysis as outputs in the form of graphs and tables. This study's findings highlight the need for increased vaccination awareness and provide new insight into the factors that make some individuals feel uneasy about being vaccinated.

Since it is crucial to grasp the public's mood in a variety of scenarios, this research has significant future potential. Sentiment analysis may be conducted again to examine public opinion on the third wave, public opinion on immunization delay in India, and similar subjects. There is a wealth of relevant data at our disposal, which we may decipher and analyse for use as a springboard for future action. Any piece of private information may be utilized as a dataset to help analyse the tone of a tweet or other piece of data. In today's lightning-fast, data-powered world, having a matrix to better comprehend the reasoning behind opinions on how to use data is critical.

REFERENCES

[1] Singh, M., Jakhar, A.K. & Pandey, S. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc. Netw. Anal. Min. 11, 33 (2021). https://doi.org/10.1007/s13278-021-00737-z

[2] International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 16, Number 2 (2020), pp. 87-104 © Research India Publications https://dx.doi.org/10.37622/IJCIR/16.2.2020.87-104

[3] Manal Abdulaziz, Alanoud Alotaibi, Mashail Alsolamy and Abeer Alabbas, "Topic based Sentiment Analysis for COVID-19 Tweets" International Journal of Advanced Computer Science and Applications (IJACSA), 12(1), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120172

[4] T. Vijay, A. Chawla, B. Dhanka and P. Karmakar, "Sentiment Analysis on COVID-19 Twitter Data," 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1-7, doi:10.1109/ICRAIE51050.2020.9358301.

[5] A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment on COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778, doi:10.1109/ICCMC51019.2021.9418320.

[6] G. Matošević and V. Bevanda, "Sentiment analysis of tweets about COVID-19 disease during pandemic," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 1290-1295, doi:10.23919/MIPRO48935.2020.9245176.

[7] Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regression," 2020 6th Information Technology International Seminar (ITIS), 2020, pp. 238-242, doi:10.1109/ITIS50118.2020.9320958.

[8] "Sentiment Analysis" https://brand24.com/blog/sentiment-analysis/

[9] "Dataset"https://www.kaggle.com/gpreda/allcovid19-vaccines-tweets