Sentiment Analysis of COVID-19 Vaccine Tweets

Apoorva Shete[1], Rohan Pradyuman[2], Sheetal Gondal[3]

[1]Department of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College, Mumbai, India
[2]Department of electrical and electrical engineering, Mahatma Gandhi institute of technology, Hyderabad, India
[3]Assist. Professor, Dept. of Information Technology Engineering, Thadomal Shahani Engineering College, Mumbai, India

Abstract - The COVID-19 vaccination has been a hot topic of debate in India as there were different speculations in the air. The people of the country responded to this in different ways (positive, negative and neutral) over Twitter in the form of tweets and retweets. Thousands of these tweets are scattered and unorganised data and this research aims at making the data speak about the sentiment behind it. This research intends to harness the potential of the huge amount of data on Twitter and derive meaningful conclusions from it. An elaborate study on the sentiments of people can provide us with an equitable understanding of the generalized perspective of the public towards vaccination. The dataset considered for the research is the collection of tweets related to vaccination from 12th Dec 2020 to 8th Aug 2021 consisting of a collection of 16,05,152 tweets concerned with the topic of vaccination.

Key Words: Sentiment Analysis, Twitter, Polarity, Subjectivity, Natural Language Processing (NLP)

1.INTRODUCTION

The COVID-19 pandemic mercilessly hit everyone across the planet and there was a lot of chaos among all the nations of the world. Due to the mutative and violent nature of the virus, all hope rested on the vaccination. A lot of multinational companies tried their best to come up with a vaccine and few such products were Pfizer, Moardina, Covi Shield etc. Well, the point that side effects for vaccines are inevitable was not received by the general public properly, though there was a majority of acceptance there was a visible amount of protest too. A lot of speculation in ethical and social media also has a weightage in the perspective of general audiences. That being said social media played a key role in communication and expression of thoughts about vaccines, Twitter in specific with its interesting features to tweet i.e express an opinion, retweet i.e to support an opinion and extend comments and like, helped people to express their thoughts on the internet to a larger set of people.

Over 500 million tweets made per day make Twitter an amazing pool of data that can result in meaningful conclusions if harvested and harnessed properly. There are a lot of studies that are based on data from Twitter. Even in India people reacted candidly over the issue of vaccination and expressed their views over Twitter as tweets, retweets etc. If the sentiment of people is analysed based on their tweets over Twitter using the technology of sentiment analysis then a lot of meaningful conclusions can be drawn out of it. Opinion mining for sentiment analysis is a technique that can analyse the data to determine the nature of the data (positive, negative or neutral).

This paper, therefore, focuses on providing major insights by performing sentiment analysis on all the tweets related to vaccines. The objective of this research is to give exploratory data analysis on all the tweets and Twitter data using Sentiment Analysis. This research will thus help us understand the overall public opinions and attitudes regarding the COVID-19 vaccines.

This paper is structured as follows, section 2 discusses the relevant research done previously in this field. Section 3 provides a brief explanation of sentiment analysis. Section 4 describes in detail the dataset used for this research. Section 5 gives an in-depth explanation of all the insights achieved through the exploratory data analysis of the dataset. Section 6 and 7 give the experimental results and important conclusions about this model along with the future scope of this project.

2. LITERATURE REVIEW

The concept of opinion mining or sentiment analysis has been used and adapted for different analysis over time by harnessing the power of Natural language processing (NLP) to analyse the sentiment that is being conveyed in the particular data. This section discusses the relevant research done previously in this field.

The paper[1] performs Sentiment Analysis on twitter data with the help of the BERT model. The data used in this paper was categorized based on the locations of the individual tweets. The data was trained using the BERT model for emotion classification and the performance of the model was evaluated using the SVM classifier. An accuracy of nearly 94% was achieved on the collected dataset. In paper [2], a model was built for analyzing the effect of COVID-19 on stocks using the COVID-19 twitter data. This model was trained using supervised learning with an accuracy of 86.24%. The aim of this research was to help companies predict stock prices, discover new marketing strategies, and track the development of the company post the Coronavirus pandemic. Paper [3] focuses on the most discussed topics on Twitter during and after the first wave of the COVID-19 pandemic. Topic extraction was carried out using Latent Dirichlet Allocation (LDA) and a Lexicon based approach aws taken for performing sentiment analysis. In this paper, the interests of everyone regarding various topics during the first wave of the pandemic were well presented. A dataset of 600,000 tweets in English language was used, the model was then trained with 80% of the dataset while 20% of the data was used for testing of the model. People's feelings related to the most traded topics were presented in the paper using sentiment analysis.

Paper[4] focuses its research on twitter data from all the Indian states from November 2019 to May 2022. In this paper, sentiment analysis was successfully implemented on the collected dataset and it was concluded that the overall sentiment of the Indian people was positive. Certain states had a higher number of tweets as compared to others because of the higher number of positive COVID-19 cases. The paper [5] presents a comparative study of twitter sentiment of all the COVID-19 tweets. Here, VADER sentiment analysis ,BERT sentiment analysis and Logistic Regression was used to determine the sentiment of the tweets. Paper [6] uses two datasets, the textual posts of twitter in the month of April 2020 from six different countries and the tweets of top 10 politicians based on the topic of Coronavirus. The paper concludes by showing results that help understand sentiment and the differences between the sentiments of each country. "Trust", "Fear" and "Anticipation" were proven to be the top emotions among people from the six different countries. In paper [7], the twitter data from 30th April 2020 was used for sentiment analysis which was carried out using term weighting TF-IDF and Logistic Regression. An accuracy of 94.71% for sentiment classification of the tweets was achieved by this model.

This literature review helped us gain essential insights about the research done previously in this field. This helped us understand how to go about with our own project.

3. SENTIMENT ANALYSIS

Sentiment Analysis is a text and data classification tool that leverages Natural language Processing (NLP) for data analysis and gives us insights into the overall sentiment (positive, negative or neutral) about the data[8]. This helps us in determining the author's opinions and attitude towards a written piece of text. This technique uses a scaling system that tells the emotions and attitude of a piece of text. This scoring makes it easier to bifurcate the text into three categories: positive, negative and neutral. Sentiment analysis, also known as Opinion Mining or Emotion AI is widely used in the business sector to understand the customer feedbacks and public's opinions towards a brand or product, insights from which can be used for marketing strategies, advertising campaigns, market research, service or product improvement, etc.

This paper uses Sentiment Analysis for analysing the public's opinion about the COVID vaccines through Twitter Data after the second wave of Coronavirus. The trends and patterns observed in this paper can be useful for gaining insightful knowledge about the public's opinion and attitude towards the COVID-19 vaccines.

The two main steps followed for Sentiment Analysis are:

1. Dataset preprocessing, cleaning and feature selection
2. Performing Sentiment Analysis on the data for exploratory data analysis

4. DATASET DESCRIPTION

The first and foremost step in this project was data collection and preparation. The dataset used for this research is the 'Covid-19 All Vaccine Tweets' dataset. It has been taken from kaggle.com[9] and provides all the tweets related to vaccines from 12th Dec 2020 to 8th Aug 2021 with 80,418 tweets and 15 attributes. Table 1 shown below, gives the information about each attribute and its description.

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

After this, all the unnecessary information from the tweets in the dataset, like the mentions, hashtags, retweet information and links, was removed from the tweets. The exact time of each tweet aws also removed from the dataset as it is not necessary. From all the aforementioned attributes, a few important ones are selected for exploratory data analysis.

6. METHODOLOGY

The section gives a detailed explanation about the methodology followed for performing Sentiment Analysis on the chosen dataset.

The first step was a collection of data suitable for the analysis. A detailed description of the same was given in the previous section. The data in the dataset was cleaned and pre-processed. The duplicate columns from the dataset were dropped and 80306 entries remained after this step. The tweets were then stripped of mentions, hashtags, retweet information, any links, etc. The exact time of each tweet was also removed from the data. Next, few of the important features were selected from all the previously mentioned attributes for the performing data analysis. To understand the data better, a graphical exploration of the data was done and a few important graphs were plotted.

Fig. 1 shown below tells us how many tweets were sent using which kind of device. From the given plot it is clear that most of the tweets were sent using an Android device followed by the Twitter Web App, while the least tweets were posted using the Cowin Vaccination Availability platform.

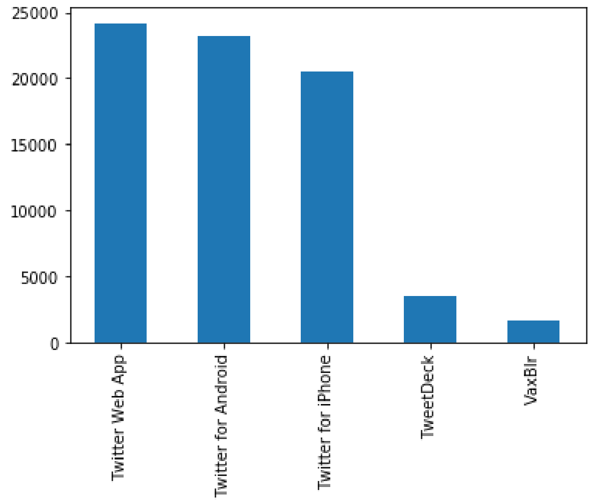


Fig -1: Plot showing the source of the tweets posted

The next plot, Fig. 2, shows the number of tweets that were sent using verified and unverified accounts. From the plot, it is clear that nearly 70,000 tweets were posted from unverified accounts and approximately 10,000 tweets were posted using a verified account.

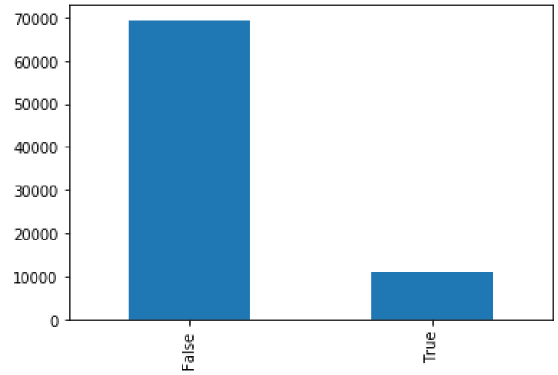
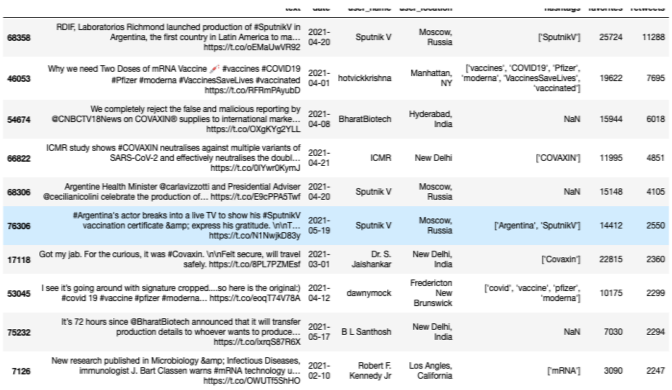


Fig -2: Plot showing the number of tweets sent from verified or unverified accounts

To see the most famous tweets related to the COVID-19 vaccines, the top 10 most retweeted tweets from the dataset were extracted, irrespective of the location. They are as shown in Fig. 3 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/ixrqS87R6X	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

Fig -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.

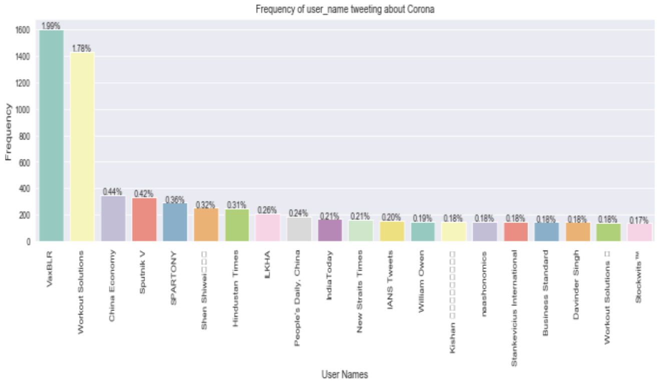


Fig -4: Top 20 accounts based on the frequency of the tweets

The data was segregated into three different classes as per the polarity values. The tweets that had polarity values ranging from -1 to -0.01 were labelled as 'Negative'. The tweets having polarity values between -0.01 to 0.01 were marked as 'Neutral' and the tweets having polarity 0.01 to 1 were considered to be 'Positive' i.e. the three classes: positive(1), negative(-1) and neutral(0). The plot in Fig. 6 shown below gives the tweet count of the tweets as these three classes.

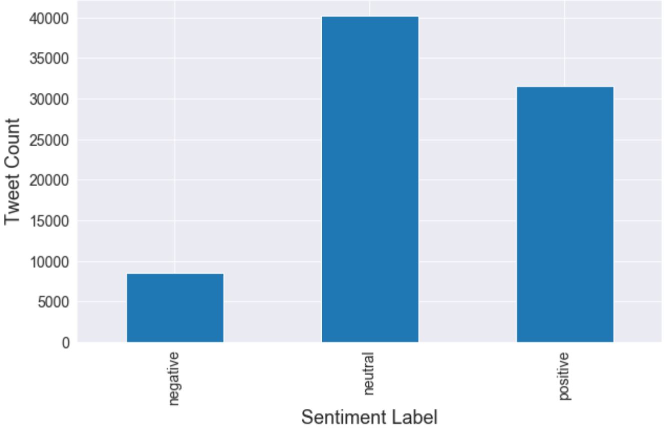


Fig -6:Tweet count given as positive, negative or neutral class

In Fig. 7, the distribution of sentiments across the tweets in the dataset was plotted along with CDF of sentiments across the tweets.

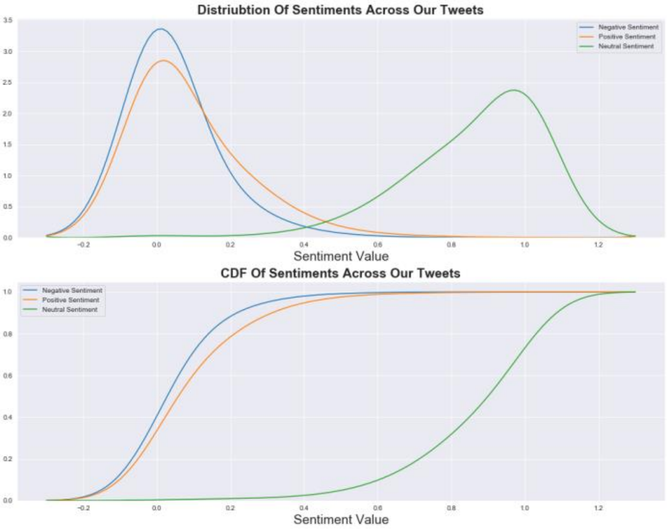


Fig -7: Distribution and CDF of Sentiments across tweets in the dataset

In the next step, the word clouds for the common words among the most positive and most negative tweets in the entire dataset were found. This is shown below in Fig. 8. As it can be seen from the figure below, the most common words in the positive tweets were: good, thank, effective, vaccinated, great, happy, safe, etc., highlighting the positive response of the public towards the COVID-19 vaccines. Similarly, the most common words for the negative tweets were: emergency, forced, alone, Canda, stop, second, death, India, Ontario, etc.

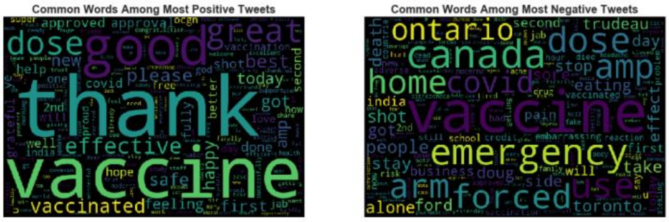


Fig -8: Word Clouds for the common words among the most positive and most negative tweets

For further analysis, word clouds from tweets related to a few important countries and cities were plotted. They are shown below in Fig. 9, Fig. 10 and Fig. 11.

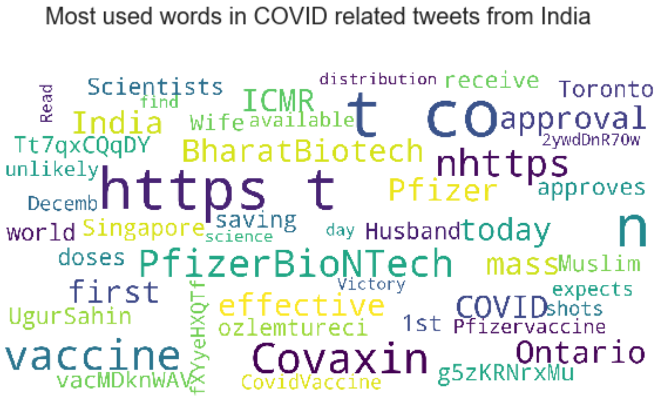


Fig -9: Most common words in tweets related to India

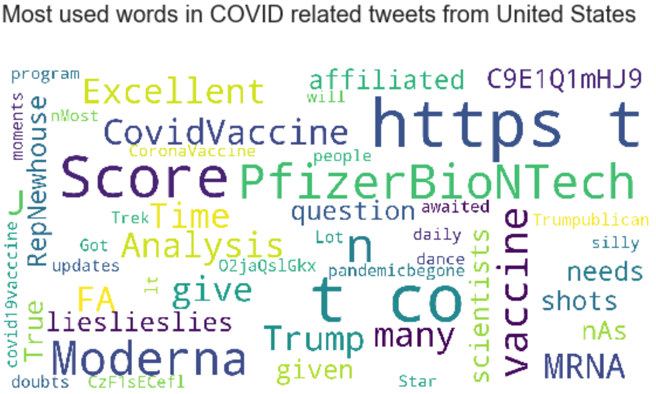


Fig -10: Fig. 10 Most common words in tweets related to USA

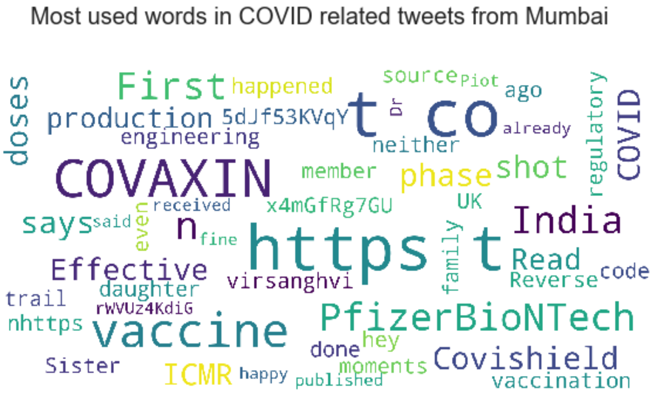


Fig -11: Fig. 10 Most common words in tweets related to Mumbai

Since in India, mostly only Covaxin and Covishield vaccines are available, word cloud for the tweets of these vaccines (combined) was plotted. It is shown below in Fig. 12.

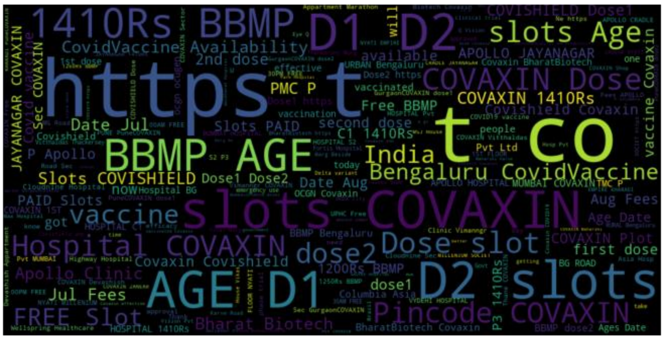


Fig -12: Word cloud for the tweets Covishield and Covaxin

After this, a few advanced operations were performed on the data to obtain the colour-coded word clouds for the segregated tweets. The data was cleaned once again. All the misspelt and nonsensical words from the tweets were cleaned and rejected. After this, the colour coded cloud words were generated for the positive, negative and neutral classes for the overall Twitter data. This is shown in Fig. 12 below.

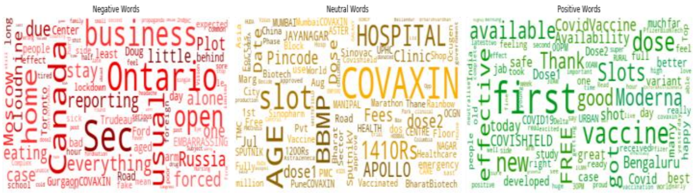


Fig -12: Colour-coded word clouds for all the tweets

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

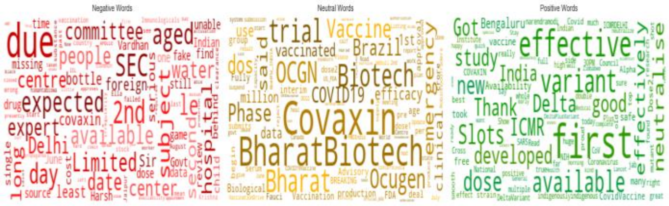


Fig -13: Colour-coded word clouds for the Covishield and Covaxin

6.1 Polarity

Polarity is a concept of assigning a value of -1, 0 or +1 to a word based on the sentiment it conveys. For a positive word polarity is +1, for a neutral word its 0 and for a negative word its -1. The average of all the words in a tweet gives us the polarity of the tweet between a float of (-1 to +1). Polarity is a tweet is a matrix to analyse the sentiment of a tweet into three categories of positive, negative and neutral.

6.2 Subjectivity

Subjectivity defines the amount of personal information and factual information in a paragraph or a tweet in this case. With more amount of personal information in a text the subjectivity rate increases and with more amount of factual information the subjectivity rate decreases. It's basically a measure of how personal or subjective the tweet or the given text in general.

In the next step, few of the generic statements related to vaccination were selected from the dataset and their sentiment was tested using polarity and subjectivity. The polarity and subjectivity scores are displayed in Fig. 4 and Fig. 5 respectively.

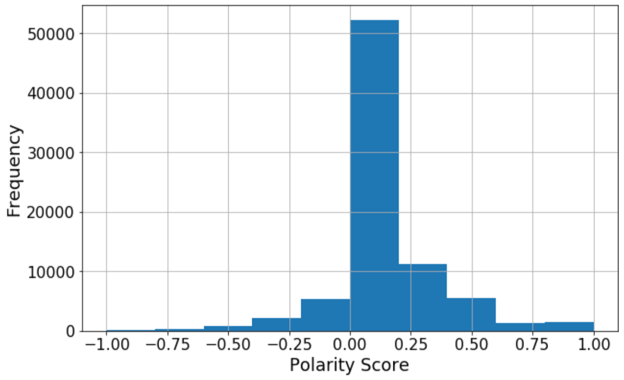


Fig -14: Polarity score of the tweets

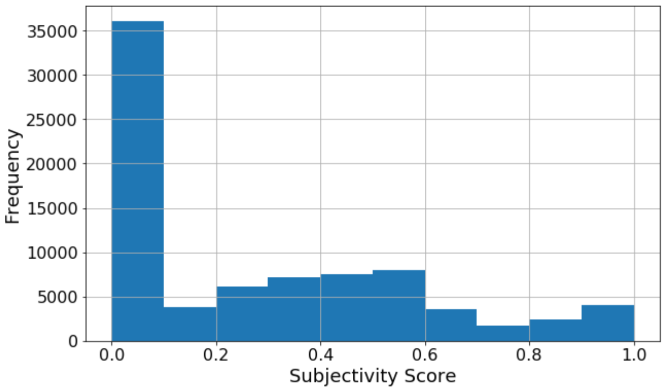


Fig -15: Subjectivity score of the tweets

A workflow of the methodology followed for this research has been shown below.

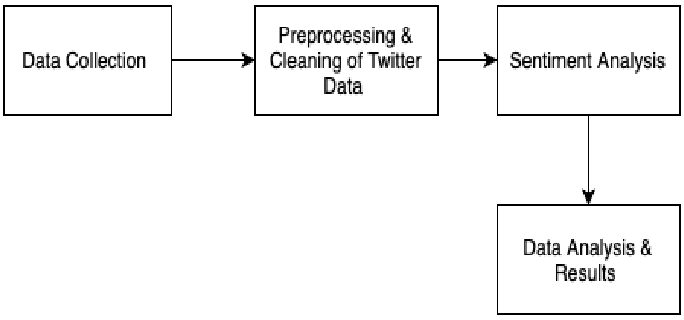


Fig -16: A flowchart of the methodology followed for implementation of this project

7. RESULTS AND DISCUSSIONS

In this research, sentiment analysis was performed on a twitter data set consisting of 16,05,152 tweets and retweets about vaccination in India. The data used for the study was cleaned and pre-processed and duplicates were eliminated. Mentions, hashtags, retweets and links were removed from the tweets and exact time of tweet was also removed. All the required attributes were selected and all the important graphs such as sentiment labels graph, distribution graphs etc were plotted and simultaneously word clouds of different combinations were extracted to understand the sentiment of the data. Furthermore few generic statements related to vaccination were selected and their sentiment was tested using polarity and subjectivity. After a thorough analysis of the data it is evident that the sentiment of the people in India about vaccination turns out to be **majoritively positive** but yet there is a **visible negation** towards vaccination in India.

8. CONCLUSION AND FUTURE SCOPE

The coronavirus has impacted the people in multiple ways and this kind of analysis can help government and other research organisations to understand the public opinions and take necessary steps to fill the gaps of awareness that's required. The analysis of twitter data is very important as that's where a lot of personal subjective viewpoints are poured down candidly by a lot of people. The paper thoroughly analyzes thousands of public opinions in the form of tweets on the Twitter platform by using techniques of Natural language processing (NLP) such as subjectivity and polarity and provides us the necessary analysis as outputs in the form of required graphs and tables. The analysis of this study opens up a scope for understanding the reason for discomfort of people towards vaccination and a need for more awareness about vaccination.

This study in particular has a tremendous scope in the future as it's really important to understand the sentiment of people in various cases. It can again be performed to analyse the people's opinion on third wave, public's opinion on vaccination delay in India and many more topics like these by using sentiment analysis. We can understand and analyse a lot of meaningful information and can be used as a basis for any further action. The datasets can go beyond the Twitter platform and any personal information can be used as a dataset to understand the sentiment of the given information. In this fast running world which is running with data as its essential fuel it's important to have a matrix to understand the opinion behind it for the better utilisation of data.

ACKNOWLEDGEMENT

We would like to thank Assist. Prof. Sheetal Gondal for her support and guidance throughout this project. We would also like to express our gratitude to our Head of Department, Dr. Ashwini Kunte, and our Principal, Dr. G.T. Thampi.

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/all-covid19-vaccines-tweets