

Sentiment Analysis and Topic Modeling from Tweets about the Covid-19 Vaccine

Bhoomika Gupta ¹, Sunita Daniel ²

¹ Student, FORE School of Management, Qutub Institutional Area, New Delhi, India.

² Associate Professor, FORE School of Management, Qutub Institutional Area, New Delhi, India.

² sunita@fsm.ac.in

² ORCID ID: 0000-0003-3463-6464

Article Info

Page Number: 675 – 696

Publication Issue:

Vol. 71 No. 3s (2022)

Abstract

Coronavirus disease 19 (COVID-19) was discovered near the end of 2019 in Wuhan, China, and has since spread rapidly throughout the world. The most effective technique for combating the COVID-19 pandemic is to rapidly and successfully develop COVID-19 vaccines. Most people's primary source of information on health and vaccination is now the Internet. Twitter is one such platform that allowed users to gather as well as disseminate information. This paper uses data mining techniques to extract data from Twitter and evaluates the sentiments associated with COVID-19 vaccines and vaccination drives being conducted in India. The purpose of the research was to gain information about the perceptions of the general public towards COVID-19 vaccines in India using Exploratory Data Analysis, Sentiment Analysis and Topic Modelling. Moreover, the research aimed to analyse the underlying factors that contributed to the respective attitudes among people. VADER Sentiment Analyzer was used to identify Positive, Negative and Neutral sentiments among people and the dominant sentiment came out to be Positive. Then, Exploratory Data Analysis was used to study different types of users with regard to the three sentiments and a time-based analysis was done to identify the events that triggered a particular reaction. To further analyse the public emotions, text2emotion was used and 5 different emotions (happy, sad, angry, fear and surprised) were identified. Using Latent Dirichlet Allocation various topics were identified related to the vaccines.

Keywords: COVID-19 vaccine, VADER sentiment analysis, latent dirichlet allocation, text2emotion, text mining, exploratory data analysis.

Article History

Article Received: 22 April 2022

Revised: 10 May 2022

Accepted: 15 June 2022

Publication: 19 July 2022

1. Introduction

COVID-19 has had a significant impact on a number of aspects of our life, including the environment, economy, mental health, and public transport. The world economy decreased by 3% in 2020, resulting in a major loss of USD 9 trillion. The COVID-19 pandemic has forced a large segment of population to sit inside their homes. Fortunately, immunizations are now available to protect against the virus' harmful effects.

The immunization campaign for COVID-19 prevention in India began on January 16, 2021. Covishield, a brand of the Oxford–AstraZeneca vaccine manufactured by the Serum Institute of India, and Covaxin, developed by Bharat Bio-tech, were approved for emergency use in India at the start of the programme. The Indian government approved the Russian Sputnik V vaccine (sold locally by Dr. Reddy's Laboratories) as a third vaccination in April 2021, and it went into effect in May 2021. This drive has already surpassed 600,000 people in its first four days, and the government has declared that it will be increased in the following days to secure citizens' immunity.

Many influential people and scholars have shared their research on the effectiveness of COVID-19 vaccines and the general attitude of people towards these vaccines and the ongoing vaccination drive in India. The aim of this study is to understand the concerns and attitude of people towards three types of vaccines, i.e. Covishield, Covaxin and Sputnik V, currently available in India. Many studies have been carried out, but they have been confined to countries and region outside India and most studies do not make use of Natural Language Processing techniques. There is a lack of a formal research about how the consumer, specifically in urban cities like Delhi and Mumbai, feels towards the vaccine and the vaccination process.

According to [Hussain et al., (2021)] COVID-19 vaccine development and distribution initiatives are quickly progressing over the world. Herd immunity requires broad vaccination administration, which involves major public participation. As a result, it is critical for governments and public health organizations to understand public opinion on vaccines, as this information may be used to direct educational campaigns and other specific policy actions. There is still a segment of the population that is suspicious of the COVID-19 vaccine. The purpose of this study is to look at the sentiments expressed in tweets on accessible vaccines and vaccination programmes in India, more specifically in Delhi and Mumbai.

2. Literature Review

This section synthesizes the extant literature and identifies the gaps in knowledge. It also substantiates the presence of the research problem by addressing the gaps.

The article written by [Jain and Katkar, (2015)] outlines the way to analyse user sentiments using classification models for data mining. It also compares the performance of individual classifiers for the study of feelings using classifier ensemble. The experimental results show that the nearest neighbor classification yields very good predicted precision. The result shows also that individual classifiers outperform the classifier method ensemble.

According to [Xue, et al., 2020)], from 7 March to 21 April 2020, Twitter authors analysed 4 million COVID-19-associated posts that included a list of 20 hashtags (for example, coronavirus, COVID-19 and quarantine). We employed a Latent Dirichlet Allocation (LDA) machine teaching approach to identify common unigrams and bigrams, prominent themes and motifs, and feelings within the Tweets gathered. This study shows that Twitter data and

machine learning technologies were often used in an infodemiology study to investigate public debate and sentiment evolution during the COVID-19 epidemic.

According to [Singh, et al., (2020)], social media has become a vital part of everyone's life, serving as a virtual forum for people to discuss their topics of interest with like-minded people all over the world, thereby breaking down geographical barriers. People used Twitter extensively during the COVID-19 outbreak to express their feelings about COVID-19 and thereby related issues.

This study intends to map people's opinions during the COVID-19 epidemic, drawing inspiration from the widespread usage of Twitter by individuals all around the world. From March 29, 2020 to March 31, 2020, a fairly small sample of 10,403 tweets was collected using the keywords (CORONVIRUS and COVID-19) for testing purposes. After the tweets had been pre-processed, the sentiment analysis procedure was used. The method of extracting sentiment from a particular piece of text is known as sentiment analysis. The filtered tweets were then subjected to emotion analysis. The words are classified into eight e-motions in e-motion analysis: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, with the four emotions (anger, disgust, fear, and sadness) linked with negative sentiment and the remaining four emotions with positive sentiment.

It is mentioned in [Abbas and Hussein, (2020)], the wording used in most Twitter social media data is vague, making it difficult to distinguish between positive and negative attitudes. There are about one billion social media communications that must be recorded in a database and processed properly in order to be researched. An ensemble majority vote classifier is proposed in this research to improve sentiment classification performance and accuracy. To construct one ensemble classifier, the suggested classification model is integrated with four classifiers that use different techniques—naive Bayes, decision trees, multilayer perceptron, and logistic regression. In addition, a comparison is made between the four classifiers in order to assess the performance of the individual classifiers. The result demonstrates that, as a private classifier, the naive Bayes classifier outperforms the others. However, when the proposed ensemble majority vote classifier is compared to the four individual classifiers, the result shows that the suggested classifier outperforms the independent ones.

[Trajkova, et al., (2020)] mentions that people use social networking services like Twitter to express their thoughts, report real-life occurrences, and provide a viewpoint on what's going on in the world. During the COVID-19 pandemic, people used Twitter to distribute data visualizations from news outlets and government agencies, as well as to upload casual data visualizations that they had created themselves. To find out what others are saying, the authors did a Twitter crawl of 5409 visualizations. The research looks at what individuals are tweeting and retweeting, as well as the issues that will occur when understanding COVID-19 data visualization on Twitter. Findings reveal that a variety of factors influence the RT count of the first post, including the source of the data, who generated the chart (person vs. organization), the type of visualization, and hence the variables on the chart. They identify and discuss five challenges that arise when interpreting these casual data visualizations, as well as recommendations that Twitter users should keep in mind when designing COVID-19 data visualizations in order to simplify data interpretation and avoid the spread of misconceptions and confusion.

According to [Sarracén (2021)], due to the rise of hate on the internet and its detrimental implications, automatic hate speech identification has become a critical issue. The focus of the research focuses on hate tweets on Twitter. The research hypothesis is that a combination of variables such as user activity and communities, as well as the photos that would be shared with tweets, can typically increase the prediction of hate speech when considering textual content. The authors have created ways for the automatic identification of hate using multimodal and multilingual approaches along the way. In addition, they have investigated the use of counter-narrative as a strategy for reducing the negative impacts of hate speech. To deal with the problem, they have used deep learning techniques, expanding on the research of graph-based approaches to representation.

The objective of [López-Chau et al., (2020)] was to explore popular discourse about the COVID-19 pandemic and policies implemented to manage it. Using tongue Processing, Text Mining, and Network Analysis to research corpus of tweets that relate to the COVID-19 pandemic, we identify common responses to the pandemic and therefore the way these responses differ across time. Moreover, insights on how information and misinformation were transmitted via Twitter, starting at the primary stages of this pandemic, are presented. Finally, this work introduces a dataset of tweets collected from everywhere the earth, in multiple languages, dating back to 22 January 2020, when the whole cases of reported COVID-19 were below 600 worldwide. The insights presented during this work could help inform decision makers within the face of future pandemics, and thus the dataset introduced are often used to acquire valuable knowledge to help mitigate the COVID-19 pandemic.

The paper [Negara et al., (2019)] discusses that Twitter could also be a well-liked social media for every user to issue thoughts and emotional forms which are tweets, tweets that only have 140 characters with limitations to write down in text. Twitter is one of the social media places to urge information that's always up thus far, tweets are categorized into big data because tweets are information which can be used as a source of data for research. Latent Dirichlet Allocation (LDA) as an algorithm which can process large text data (big data). during this study using the LDA method as an algorithm to provide topic modelling, each topic similarity, and visualization of topic clusters from the tweet data generated as many as 4 topics (Economic, Military, Sports, Technology) in Indonesian, where each topic features variety different tweets. The LDA method utilized within the processing of tweet data is successfully administered and works optimally, in each topic extraction, topic modelling, generating index words that are in each topic cluster and computer visualization within the subject. LDA output shows optimal performance within the method of word indexing in Sport topics with 1260 tweets with an accuracy of 98% better than the LSI method in Topic Modelling.

In the paper [Karlsson, et al., (2021)] it is mentioned that because people see COVID-19 as a dangerous disease, a vaccine against it is likely to be in great demand. When an individual selects whether or not to accept the vaccine, vaccine safety concerns may exceed the anticipated disease risks. The role of COVID-19's perceived risk and the perceived safety of a potential COVID-19 vaccination in determining intentions to simply accept a COVID-19 vaccine was studied. Eight hundred and twenty-five parents of young children, 205 residents of a neighborhood with inadequate vaccination coverage, and thirteen hundred and twenty-five Facebook users across Finland were polled.

It was observed from the literature studied that there are several studies conducted to analyze sentiments of people on COVID-19 disease as well as the lockdown period. Studies have also been conducted on perceptions related to vaccines but they are limited in number and are mostly conducted outside the Indian context. With these gaps the aim of this study is to develop and apply a Machine Learning and Natural Language Processing based approach to analyze public sentiments on social media in India COVID-19 vaccines to better understand the public attitude and concerns regarding COVID-19 vaccines and the vaccination process.

3. Materials and Methods

The following section discusses the methodology adopted during the course of the study as shown in Figure 1. The methodology has been explained in such a way that the study can be replicated in a similar way. The methodology consists of Conceptual Framework, Data Collection, Data Analysis and Description of the Data.

3.1 Data Collection

The data was collected by web scraping. Web scraping [Kaur, (2020)] is a data mining technique that allows us to mine data from various online platforms including social media websites like Twitter, Facebook, etc., online articles and journals, and so on. Some of the tools that are generally used to collect data from Twitter are Tweepy [Garcia, n.d.] and Twint [Project, n.d.] python libraries both of which are open source. Tweepy has a set of functions that have models and API endpoints that allow it to implement processes transparently. It encapsulates complex API functions and allows the user to work with other useful functionalities on top of encapsulated layer. Twint on the other hand allows extraction of data without API. Both Twint and Tweepy allow users to fetch data from Twitter however, there are certain features that make the extraction process easier.

- Tweepy puts a limit to the number of tweets that can be extract at a particular time frame. Twint has no such limit and scrapes almost all the tweets.
- No Sign-up or API required.
- Fast and convenient.

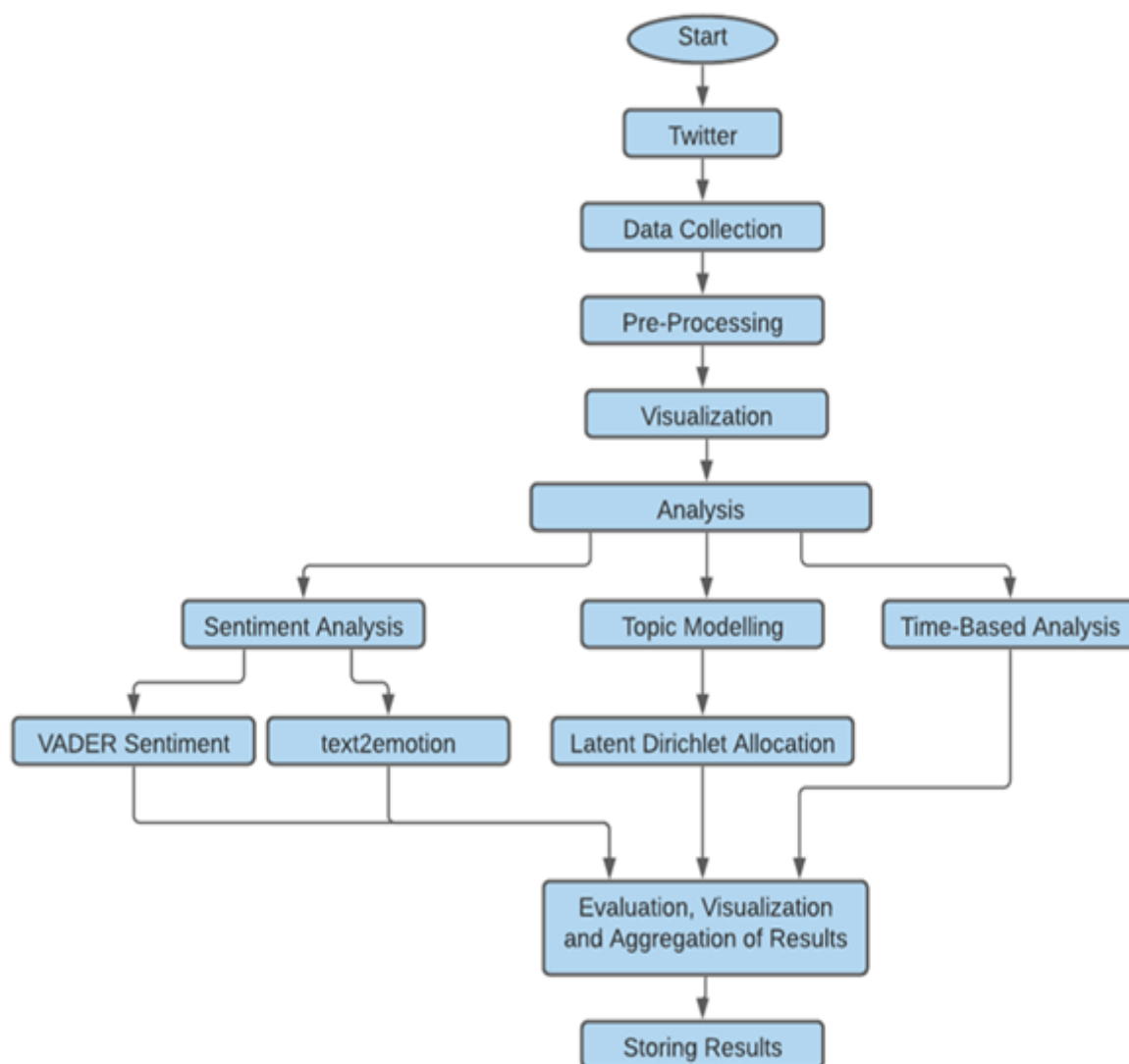


Figure 1; Flow of Study

For the purpose of the study, data has been collected for Mumbai and Delhi for 32 different keywords with 11960 data points from December 12, 2020 to May 31, 2021 using Twint library as in Figure 2. The keywords used are shown in Table 1.

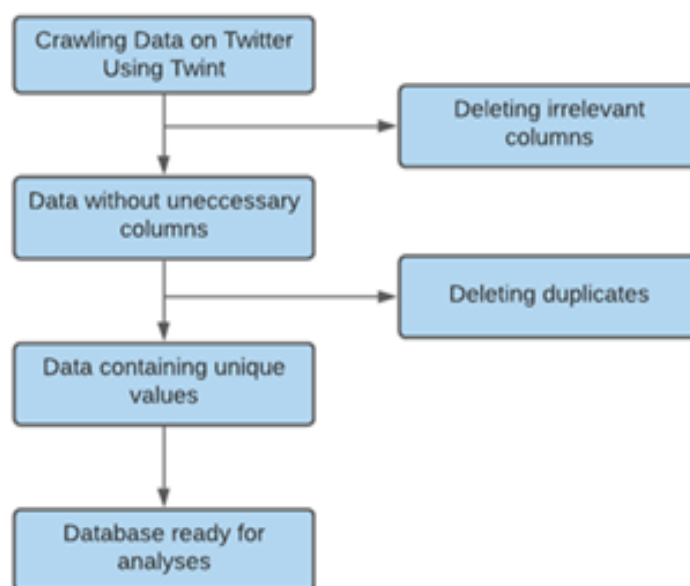


Figure 2; Method for Data Collection Using Twint

Table 1. Keywords Used

Keywords Used	Covishield, vaccineswork, getvaccinated, covidcovishield, covid19vaccines, covishieldworks, covishieldvaccine, covishieldindia, deltavariant, seruminstitute, SII, oxford-covishield, astrazeneca, covishield works, effective vaccine, vaccine fear, covaxin, sputnik, sputnikV, coronavaccine, bharatvaccine, largest vaccine drive, bharat biotech, bharatbiotechvaccine, covaxinworks, covaxineffects, getcovaxinated, sputnikworks, russianvaccine, indiafightscorona, vaccinateindia, cowin
----------------------	---

3.2 Description of Data

The data collected for the study contains 11960 rows and 12 columns. The rows represent different tweets posted by the users on Twitter using the hashtags taken into account while columns represent different attributes for each tweet. Each column is explained in Table 2.

Table 2. Description of Data

Unnamed: 0	Represents row index
date	Date on which the tweet was posted
time	Time at which the tweet was posted
username	Username of the user posting the tweet
name	Name (as on Twitter) of the user posting the tweet
language	Language in which the tweet was posted
replies_count	Number of replies on a particular tweet
retweets_count	Number of retweets on a particular tweet
likes_count	Number of likes on a particular tweet

hashtags	Hashtags associated with a particular tweet
near	Place at which the tweet was posted

The datatypes for each column were noted. It was observed that index, replies_count, retweets_count and likes_count are integers. Whereas, Date, time, username, language, name, hashtags and near are string type. Moreover, near and language are categorical in nature.

3.3 Data Cleaning and Pre-Processing

Usually, if the data is discarded or data is provided for analysis, it is always in its natural human sentence or sub-paragraph style, etc. Prior to analyzing this, we must change and filter this language so it can be understood by the computer in the proper manner. Data pre-processing for a machine learning model is a critical step. The outcome would be reliable if the data are fairly pre-processed. In Natural Language Processing (NLP) technique, pre-processing is the initial step before creating the learning model for machines as shown in Figure 3.

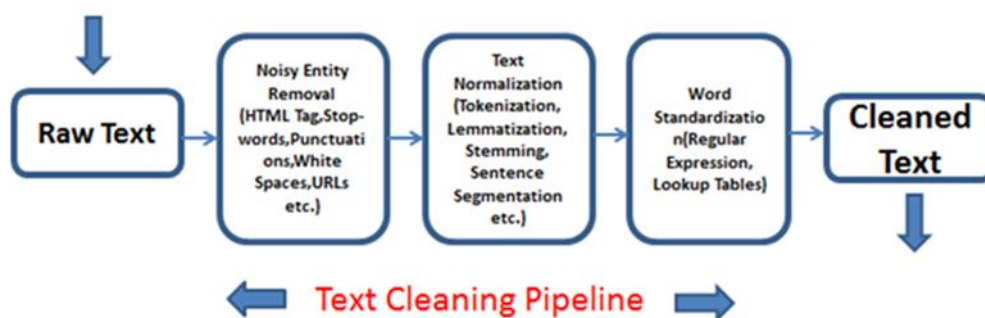


Figure 3; Text Cleaning Pipelining Process

Therefore, in order to clean and pre-process the data, following steps have been followed.

1. Checking for missing values in each row.
2. Check for languages other than English: The dataset consisted of languages such as Hindi, Marathi, etc.
Hence in order to utilize NLP models, values containing language other than English have been Dropped.
3. Clean the corpus: The entire text (tweet) is converted in lower case. Then hashtags, emojis, punctuations, multiple spaces, URL's, numbers and special characters have been removed from the tweets.
4. Remove stopwords to discard words that have no contribution or significance towards the emotion of the text (tweet).
5. Tokenize text (tweets): Tweets are converted in list of strings so that they can be analysed by the model.
6. Lemmatization and Stemming: Words are converted into their base form (stemming) and lowered into their present language (lemmatization).

3.4 Conceptual Framework

3.4.1. Sentiment Analysis

Sentimental analysis [Beri, (2020)] is a textual analysis method which finds polarity in a text, regardless of whether the document, paragraph, sentence, or clause is entirely polar (i.e. a positive or negative opinion). It attempts at measuring a speaker/attitude, writer's feelings, assessments, positions and emotions based on the computer handling of subjectivity in a text.

- **VADER Sentiment Analysis**

VADER [Beri, (2020)] (Valence Aware Sentiment Reasoning Dictionary) is a model used for analysis of the textual feelings, sensitive to both polarity (positive/negative) and emotional intensity (strongness). It can be used with the NLTK package directly for text data that is not labelled. VADER sentimental analysis relies on a lexicon that links lexicological characteristics with so-called emotional values. The sentiment score of a text can be achieved by summarising the intensity of each word.

- **Text2emotion Sentiment Analysis**

text2emotion [Bilakhiya, (2020)] is a Python library that was created with the goal of revealing hidden human emotions in text. While sentimental analysis just reads and categorises text as positive, negative, or neutral, text2emotion assists in categorising the tone of text into five basic human emotions: happy, angry, surprised, fear, and sad. To summarize, text2emotion is a Python library that may be used to extract emotions from text.

- Recognizes the emotions of any textual content.
- Happy, Angry, Sad, Surprise, and Fear are among the five emotion categories that are compatible.

3.4.2. Topic Modelling

Topic Modelling [Bansal, (2016)] is a method for automatically identifying subjects in a text object and deducing hidden patterns from a text corpus, as the name suggests. As a result, it will help you make better decisions. It's an unsupervised method for locating and observing groups of words (referred to as "topics") in big groups of texts.

- **LDA**

The Latent Dirichlet Allocation (LDA) [Bansal, (2016)] technique is a common topic modelling algorithm that is well-implemented in Python's Gensim package. The problem is determining how to extract high-quality, clearly separated, and significant subjects. The quality of text pre-processing and the approach for determining the appropriate number of subjects play a big role in this.

3.4.3. Time-Based Analysis

A collection of quantities assembled across even time periods and ordered chronologically is called time series (Erica, 2019) data. The frequency of data gathering is referred to as the time series frequency. As the name implies, it entails dealing with time-based data (years, days, hours, minutes) in order to uncover hidden insights and make better decisions.

Python makes use of the Pandas software library, which was created primarily for financial sector analysis and forecasting. Time stamps (individual points in time), time deltas (total

duration), and time periods are all used in this language (intervals). The in-built capability includes these fundamental objects that include dates and timings.

4. Results

In this section, analysis is carried out based on the conceptual framework mentioned in Section-3. An in-depth study of the data collected has been conducted using Python. The outputs obtained have been analyzed and the results drawn as discussed in this section in accordance Figure 4.

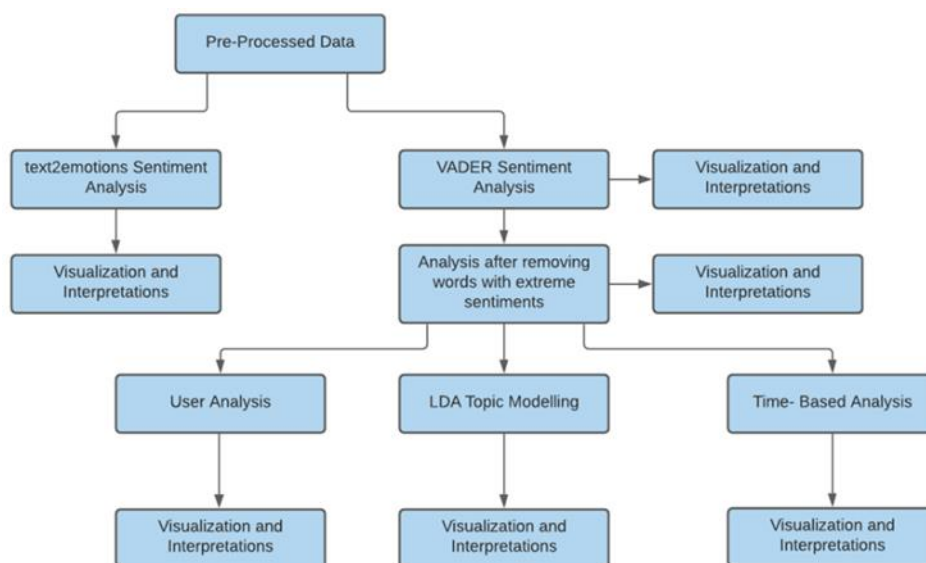


Figure 4; Flow of Analysis

4.1.VADER Sentiment Analysis

After the data was pre-processed, sentiment analysis using VADER model was performed. Sentiment Intensity Analyser was used and compound score for each sentiment using the utility function polarity scores was calculated. Based on that if the score was greater than 0 then the sentiment was termed as 'Positive' whereas, if the score the was less than 0 the sentiment was termed as 'Negative'. Scores with compound score equal to zero were termed as 'Neutral'.

Figure 5 shows the distribution of Positive, Negative and Neutral sentiments in terms of count and percentage. There are 4189 tweets that are Positive (43.8%), 3884 Neutral (40.6%) tweets and 1478 Negative (15.4%) tweets. Based on the distribution it can be inferred that the dominant sentiment towards COVID-19 vaccine and vaccination drive is positive or neutral while negative emotion is significantly less as compared to positive tweets reflecting a feeling of optimism among consumers.

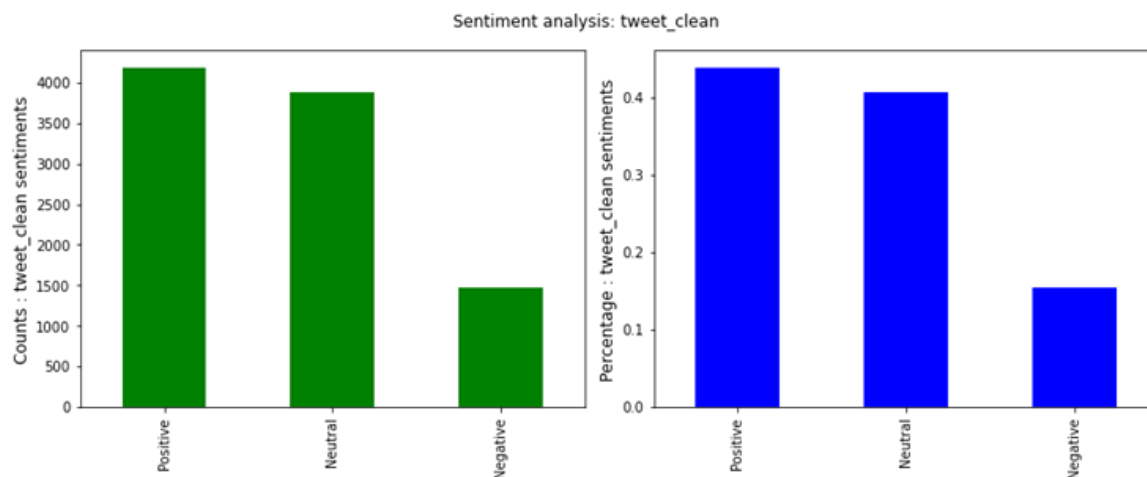


Figure 5; Distribution of Positive, Negative and Neutral Sentiments

The barplot in Figure 6 gives the distribution of positive, negative and neutral sentiments in Delhi and Mumbai. It is observed that while positive sentiment is dominant, Mumbai is more positive as compared to Delhi. Neutral sentiments in both the cities are nearly equal while negative sentiments in Mumbai are more as compared to Delhi.

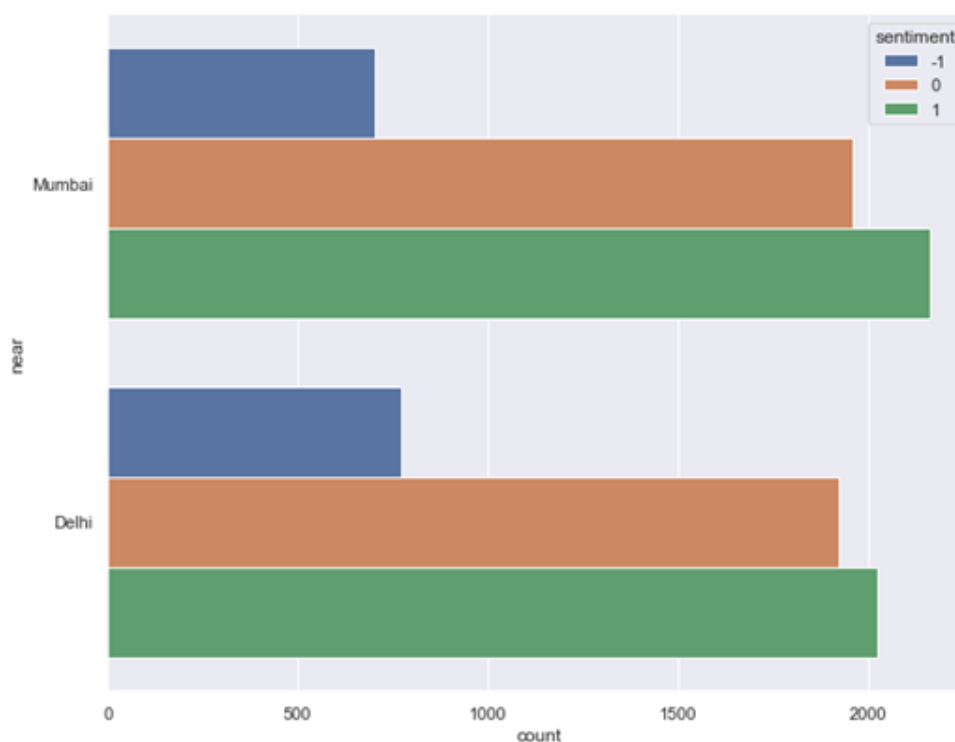


Figure 6; Distribution of Sentiments in Delhi and Mumbai

After sentiment for each data point was collected, Wordclouds were plotted for positive negative and neutral sentiment based on the location and polarity.



Figure 7; Positive Sentiments in Delhi Figure 8; Positive Sentiments in Mumbai

The wordclouds in Figures 7-8 show positive tweets in Delhi and Mumbai respectively. Words like ‘dose’, ‘approve’, ‘time’, ‘avail’ etc are prevalent. These words relate to the availability of vaccines for consumption. Other words such as ‘thank’, ‘efficacy’, ‘safe’, ‘good’, ‘kind’ etc also relate to a positive emotion. It was analysed in the tweets that people thanked the government for their efforts to roll out vaccines and make it reach to the masses. There was also a positive response towards the working of the vaccines which can be associated by words such as ‘best’, ‘work’, ‘hope’ etc.

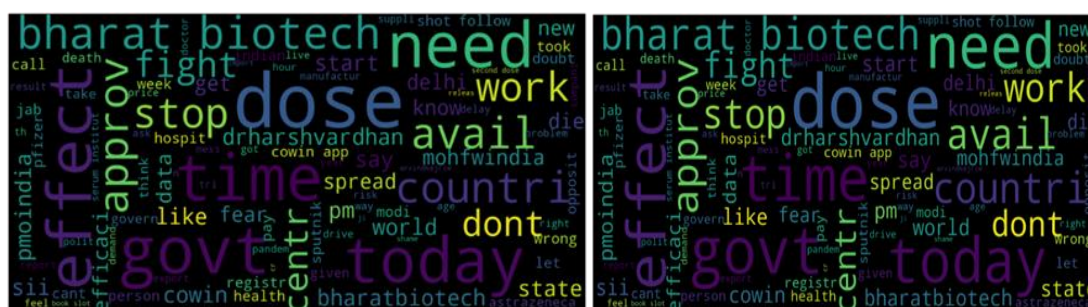


Figure 9; Negative Sentiments in Delhi Figure 10; Negative Sentiments in Mumbai

The wordclouds in Figures 9-10 show negative sentiments associated with the vaccines. Words like ‘stop’, ‘need’, ‘fear’, ‘new variant’, ‘problem’ etc. are prevalent. These words could be associated with fear inherent among the people in regards to the efficacy and working of the vaccines. After analysing the tweets with negative sentiment, it was observed that people had a fear of side effects and even death that could be a result of the vaccine. People also fear whether or not the vaccine is effective against the new variants that were a major contributor to the second wave. Another reason why consumers of the vaccine displayed a negative emotion was due the limited availability of slots for the vaccines.



For negative sentiments words greater than intensity 0.25 have been removed and for positive sentiments words greater than intensity 0.4 have been removed as shown in Figure 12. After the removal of these extreme words, the word clouds in Figure 13 were obtained.



Figure 13 gives a clearer distribution of words in positive and negative sentiments. After this Wordclouds for top 10 words accounting for positive and negative sentiments are plotted.

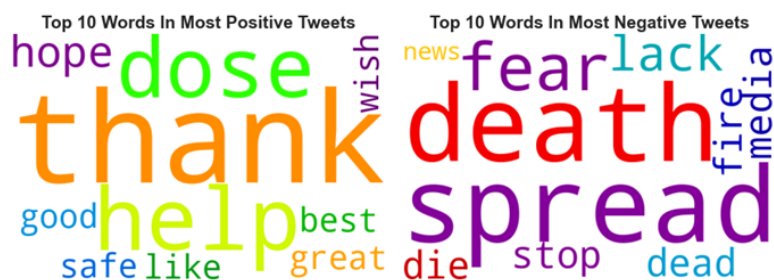


Figure 14; Top 10 Words

As it can be seen in Figure 14 most prominent words for positive sentiments are ‘thank’, ‘help’, ‘good’, ‘hope’, ‘great’, ‘safe’, ‘best’. Words like ‘thank’ and ‘hope’ are related to government’s efforts towards rolling out vaccines and conducting vaccination drives. Words like ‘great’, ‘best’ and ‘safe’ are associated with efficiency and working of the vaccines. Negative sentiments are represented by words such as ‘death’, ‘fear’, ‘lack’, ‘panic’, ‘new’, ‘slot’ etc. Words like ‘panic’ and ‘fear’ are a result of consumers’ apprehensions towards the vaccines while ‘lack’ and ‘slot’ are associated with shortages in appointment booking slots and vaccines. To identify the strengths of the sentiments from December, 2020 to May, 2021, a time-series chart has been plotted. It can be seen from the plot in Figure 15 that the dominant sentiment throughout the entire timeline is positive or neutral. It can also be observed that there have been sudden spikes and drops in the number of tweets on some days, analysis for which has been carried out further in the study.

Figure 16 shows the standard deviation over time and one interesting observation was that on December 27, 2020 there was a sudden drop and then incline in the mean and standard deviations of positive and negative sentiments. Moreover, it has been observed that standard deviation of positive and negative sentiments becomes less stable. The reason behind this instability has been explored in Figure 17. The plot shows distribution of positive sentiment. It is observed the spikes and declines are a result of government’s policies and decisions related to the vaccination usage and rolling out process.

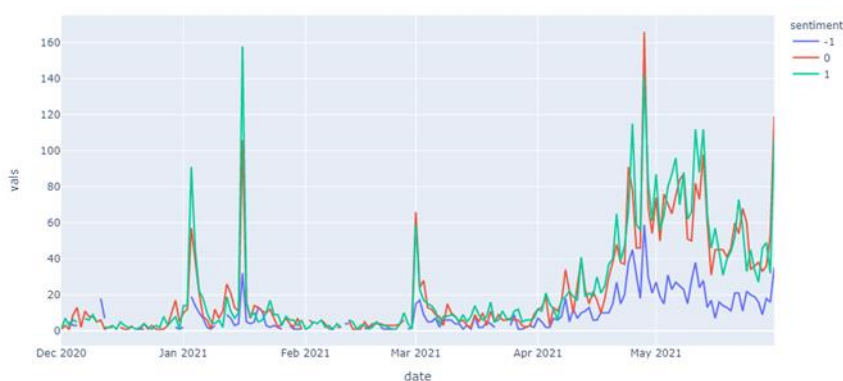


Figure 15; Distribution of Sentiments



Figure 16; Standard Deviation Change with Time

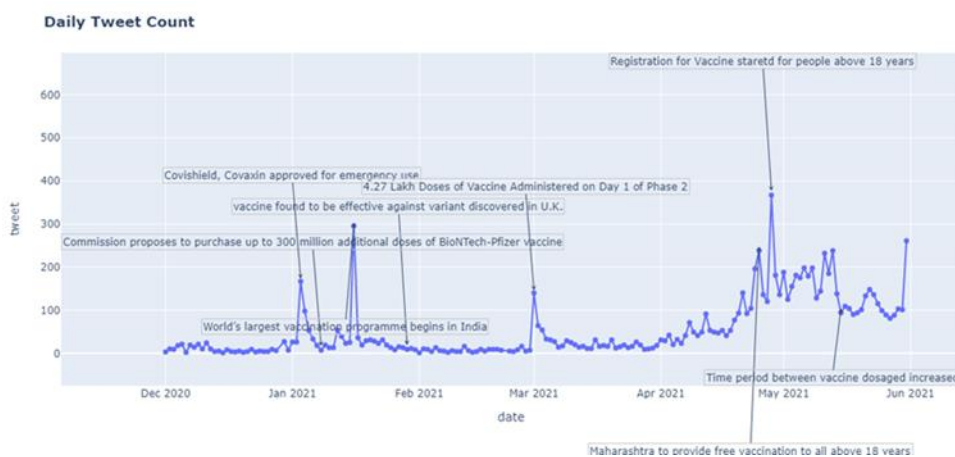


Figure 17; Events that Triggered Tweets

4.2.Latent Dirichlet Allocation (LDA) Analysis

Though sentiment analysis provided insight into the public's perception of the COVID-19 vaccine and immunisation campaign, it did not assist us comprehend the key factors that determine the public's perception. For tweets concerning the vaccine with unfavourable sentiments, Latent Dirichlet Allocation topic modelling was used to assess the general public's concerns about the vaccine and vaccination process.

To carry out topic modelling, first a dictionary of the pre-processed data is created using `corpora.Dictionary()` function. Using this dictionary, a Document-term matrix is formed using `doc2bow()` function. A document term matrix is basically a table or a matrix that contains frequency of each word in a document. This matrix is then fed to LDA model along with number of topics required, dictionary, number of passes, random state, alpha, and beta values. For this study, the optimal number of topics obtained from LDA tuning was 3 with alpha value 0.01 and beta value 0.9.

In order to find the topic that is most prevalent, dominant topic for each row is assigned with percentage contribution, topic keywords and representative text. With the help of this

dataframe a Dominant topic across the entire dataset is identified. It is observed in Figure 18 that topic 0 is the most dominant topic while topic 1 is the least dominant. Each topic is then associated with a sentiment to derive meaning in Table 3. It is observed that all 3 topics are positive while in nature.

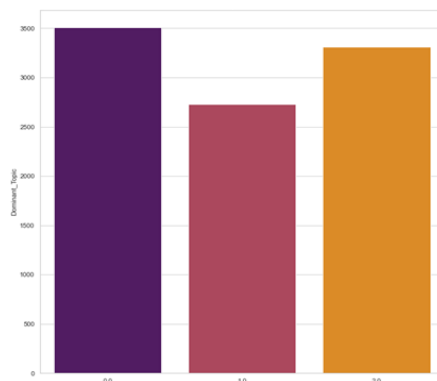


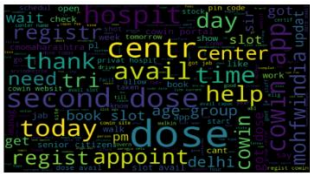


Figure 18; Dominant Topics

Table 3. Topic Modelling Results

Topic	Wordcloud	Top Words	Sentiment	Label	Description
Topic 1	 <p>Topic 1</p>	Dose, COWIN, Slot, Avail, Book, App, Center, Dose, Register, Need	Positive	Availability and Accessibility	Information related to availability of vaccines in different vaccine centers, hospitals and COWIN App.
Topic 2	 <p>Topic 2</p>	Hope, Fight, Thank, Got Dose, Like, Great, Support, Works, Take	Positive	General Sentiment Towards Vaccine	Information related to attitude and general perception of people towards the vaccine and vaccination drive.

Topic 3	 <p>Topic 3</p>	Indian, Sputnik, Pfizer, Effect, Need, Approve, Government, Astrazenica, Efficacy, Data	Positive	Efficiency and Other Vaccine Related Information	Information about working of the vaccines, their efficacy, approval, usage etc.
---------	--	---	----------	--	---

4.3. User Analysis

After studying the topics that the general public is talking about, user analysis was done wherein, the type of user was identified using K-Means clustering. To form clusters, first, user popularity was defined as 'user_popularity' based on number of likes, retweets and replies a particular user got on their tweets. A weightage of 65% was given to number of likes (likes_count) a person received, weightage of 25% was given to number of retweets (retweets_count) and 10% to number of replies (replies_count). The weightage was assigned on the basis of maximum usage of the said twitter utilities (likes, replies and retweets) by the users. As a general notion, users interact majorly with likes on twitter followed by retweets and replies. Popularity metric was calculated using (1).

User Popularity(ρ) =

$$\sqrt[2]{(0.65 * (\text{No. of Likes})) + (0.25 * \text{No. of Retweets})) + (0.10(\text{No. of Replies}))} \quad (1)$$

After finding out the popularity of a particular user, new features Number_Of_Words (number words in a tweet) and Mean_Word_Length (average length of a tweet) were generated.

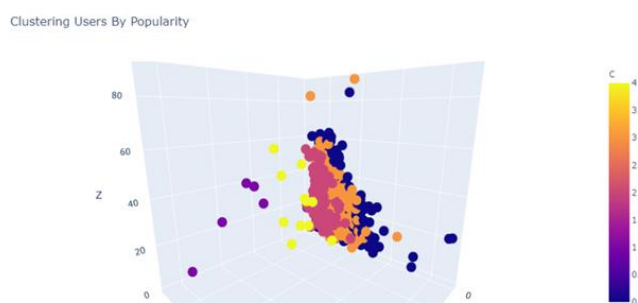


Figure 19; K-Means Clustering

Clusters were formed using K-Means clustering as shown in Figure 19. Optimal number of clusters was found out to be 5 with the help of Elbow Method which uses Within-Cluster-Sum of Squared Errors. These clusters were identified as type of user. Based on metrics, clusters were labelled as:

- Less than Average User
- Average User
- Popular User
- Very Popular User

- **Superstar User**

Positive and negative sentiments associated with the above-mentioned types of users along with the distribution of these users were plotted. It can be seen in Figure 20 that majority number of users are constituted by Less than Average users followed by average users. Popular users displayed the maximum number of positive sentiments while very popular users displayed maximum number of negative sentiments.

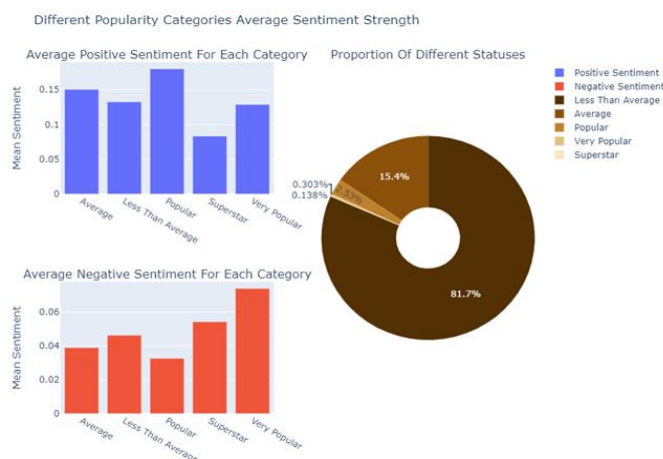


Figure 20; Average Sentiment of Each Popularity Division

4.4.Text2Emotions Sentiment Analysis

It is evident from Figure 21 that the happy is the emotion with maximum score followed by fear and surprise. Happy can be associated with people talking about vaccines being efficient. Moreover, people are gaining immunity from COVID-19, this itself contributes to a positive emotion. Fear is attributed towards shortage of vaccines and side-effects of the vaccines. Further, the time period for which the data was collected was during the second wave.

Therefore, fear can also be associated with people being afraid of what might happen to them and their loved ones during the unprecedented time. Further, shortage of vaccines and unavailability of vaccine slots also contributes towards fear among people. Sad and surprise are emotions that could be a result of the side effects of the vaccines whereas, angry could be an outcome of a situation wherein people were unable to obtain vaccines for themselves. The diagrams in Figure 22 show Wordclouds for the derived emotions.

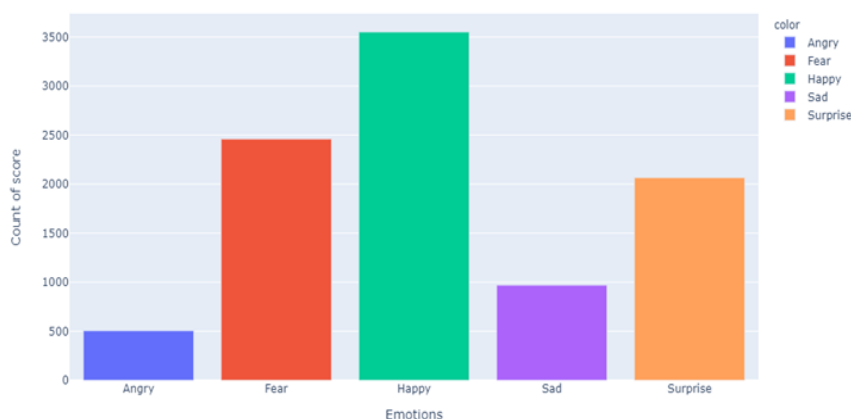


Figure 21; Distribution of Emotions



Figure 22; Wordclouds for Different Emotions

As analysed from the tweets, due to unavailability of vaccines in several regions people were afraid of either missing their second dose or missing their first dose entirely. Further, several myths and rumours being circulated around vaccines also result in fear among the people. Some of them include:

- Vaccines cause deaths
- Vaccine will be replaced by a fake medicine
- Side effects caused by vaccines, etc.

Surprise emotion could be a direct result of people getting infected even after getting vaccinated specially after the first dose. On a lighter side, people getting rare and severe side effects from these doses also contribute to this emotion. Government's decision of starting phase III of the vaccination drive suddenly also came as a surprise to a lot of people. As stated earlier, due to shortage of slots and vaccines people were unable to get vaccinated. The inability to find slots on COWIN app even after searching for a very long time made people angry and frustrated with the app. Moreover, government's decision of rolling out vaccines for people of age group 18-44 when there was already a shortage of slots and vaccines further made people furious. As the vaccines started rolling out, Delhi and Mumbai announced that the states would be providing free vaccinations to people in the age group 18-44. This is resulted in optimism among people residing in these cities. Moreover, the general public expressed contentment and gratitude towards the governments of these cities for their efforts to provide immunization towards the virus.

5. Discussion

The sentiment strength of a tweet provides us with a domain from which we can learn how the population is reacting to the vaccine. This insight allows different governments to direct their advertisements towards more negative groups that typically refuse to believe in the vaccine's integrity, and to observe the change over time, as we saw between December 13th and 27th.

Furthermore, the overall or dominating positive emotion in response to vaccinations is 'happy,' implying that the general emotion circling around the vaccine is contentment, relief, support, and thanks toward authorities' attempts to carry out the vaccination process. 'Fear' is the most prevalent negative emotion, owing to misconceptions and rumours as well as popular concern about vaccines. According to the findings, several Twitter users in Mumbai and Delhi

supported COVID-19 infection control measures and contradicted misinformation. Governments should study and implement an effective vaccine promotion plan in addition to encouraging the development and clinical administration of COVID-19 vaccines. In addition, an observation was made that users could not register for vaccination using the COWIN app. This is because of the app's failure and, above all, the reality that most people in the country don't use the internet. As a result, in order to see a higher vaccination rate, the government should start offering walk-in registrations so that those who are unfamiliar with computers and smartphones can get vaccinated.

The analysis further identified three topics using Latent Dirichlet Allocation Analysis. These topics were identified as follows.

- Accessibility and Availability of the Vaccines
- General Perception of the Public Towards Vaccines
- Vaccine Related Information and Government Decisions

Findings reveal that many Indians still feel aggravated by the whole outbreak. The citizens will reject the vaccine with such an attitude. In addition, skepticism over vaccine nationality, vaccine trial skepticism, side effects from the vaccine, the fear of death that the vaccine could lead, allergic reactions was seen among the general public. Also, distrust of pharmaceutical companies, doubts regarding vaccine company data, the prevalence of many vaccines and concerns about selecting the vaccine which provided greater safety, and hastily providing vaccines were other concerns shared by Indians. While the Indian general population is expressing legitimate worries about the COVID-19 vaccine, myths such as exaggeration of the spread of COVID-19 and a hate or disbelief for specific vaccines due to nationality were also echoed. The study also revealed that many individuals do not trust vaccines, and that governments are instilling fear in people who receive vaccinations.

With recent study it is crucial for politicians, governments, pharmaceutical firms and NGOs to invest in teaching the public at large about the importance of vaccination to return to normal life. There should be special attention on ensuring that any hoaxes and insecurities among the general public concerning vaccinations are dealt with.

6. Conclusions

The study used Natural Language Processing Techniques such as Sentiment Analysis and Topic Modelling to identify the perception of the general public towards COVID-19 vaccine and vaccine drive in India. With the help of sentiment analysis, the general sentiment of the people was understood. Topic Modelling further explored the narrative that shape these opinions. It was concluded that the dominant sentiment with regards to tweets is Positive (43%). However, the negative sentiment associated with vaccines is around 15% which is also a significant proportion when it comes to a population as large as India. These negative sentiments can attribute to the shortcomings of the entire vaccination drive and therefore need to be dealt efficiently by the policy makers and manufacturers. The three topics dealt with the general information about the vaccines, general perception of the people towards the vaccines and availability of the vaccines. These further tell the policy makers about the apprehensions as well as confidence levels of the public and would help them device campaigns and policies

accordingly. User analysis also helped in understanding the kind of users' policy makers need to target as any opinion on a social media platform can impact hundreds of people.

This research also has its own limitations. Rather than first-hand encounters with vaccines, the majority of Twitter users submitted tweets based on indirect experiences gained via the Internet and the press. This is because vaccination of a small number of people, including healthcare personnel, began during the study period. Examining the change in the public's emotional score following immunization could yield another interesting conclusion. Social media has a high level of engagement, spread, and change and hence the content on Twitter is unpredictable. Depending on the outbreak and vaccine development, the COVID-19 vaccine's hashtag and keyword may change at any time. As a result, study findings on this subject are subject to change throughout time. Only the tweets made during the research period are significant.

In the future, this will be subjected to a more thorough assessment. The functions of the VADER tool and the LDA model can be tweaked and modified for sentiment analysis in the future. With the addition of new features, the accuracy of sentiment identification will improve even further. Meanwhile, additional twitter posts for model training can be collected across a greater region and population, ensuring that the data is representative of thoughts and opinions and, in turn, can aid in the formation of a generalized opinion in the future. This examination would help to improve the results' reliability.

Acknowledgments

The infrastructural support provided by FORE School of Management, New Delhi in completing this paper is gratefully acknowledged.

References

1. Abbas, A., & Hussein, Q. M. (2020). Twitter Sentiment Analysis Using an Ensemble Majority Vote Classifier. *Journal of Southwest Jiaotong University*. doi:10.35741/issn.0258-2724.55.1.9
2. Bansal, S. (2016). *Beginners Guide to Topic Modeling in Python*. Retrieved from Analytics Vidhya.
3. <https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>
4. Beri, A. (2020). *Sentimental Analysis Using Vader*. Retrieved from Towards Data Science.
5. <https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664>
6. Bilakhiya, K. (2020). *How I Created A Python Package – text2emotion*. Retrieved from Analytics India Mag.
7. <https://analyticsindiamag.com/how-i-created-a-python-package-text2emotion/>
8. Erica. (2019). *Introduction to the Fundamentals of Time Series Data and Analysis*. Retrieved from Aptech.
9. <https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/>
10. Garcia, M. (n.d.). *How to Make a Twitter Bot in Python with Tweepy*. Retrieved from Real Python.
11. <https://realpython.com/twitter-bot-python-tweepy/>
12. Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., & Ali, A. (2021). Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. *Observational Study J Med Internet Res* 2021. doi:<https://doi.org/10.2196/26627>

13. India, P. T. (2020). Recovered Covid-19 patients last immunity for 8 months, raise hopes for vaccine: Study. Retrieved from India Today: <https://www.indiatoday.in/coronavirus-outbreak/story/covid-19-antibody-immunity-lasts-8-months-study-1752290-2020-12-23>
14. Jain, A. P., & Katkar, V. D. (2015). Sentiments analysis of Twitter data using data mining. 2015 International Conference on Information Processing (ICIP). doi:10.1109/INFOP.2015.7489492
15. Karlsson, L. C., Soveri, A., Lewandowsky, S., Karlsson, L., Karlsson, H., Nolvib, S., Antfolk, J. (2021). Fearing the disease or the vaccine: The case of COVID-19. *Personality and Individual Differences*.
16. doi:<https://doi.org/10.1016/j.paid.2020.110590>
17. Kaur, H. (2020). What is Web Scraping and How to Use It? Retrieved from Geeks for Geeks: <https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/>
18. López-Chau, A., Valle-Cruz, D., & Sandoval-Almazán, R. (2020). Sentiment Analysis of Twitter Data Through Machine Learning Techniques. Ramachandran M., Mahmood Z. (eds) *Software Engineering in the Era of Cloud Computing. Computer Communications and Networks*. doi:https://doi.org/10.1007/978-3-030-33624-0_8
19. Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS).
20. doi:10.1109/ICECOS47637.2019.8984523
21. Project, T. (n.d.). TWINT - Twitter Intelligence Tool. Retrieved from Github.
22. <https://github.com/twintproject/twint>
23. Sameh N Saleh, M., Christoph Lehmann, M., Samuel McDonald, M., Mujeeb Basit, M., & Richard J Medford, M. (2020). Understanding Public Perception of COVID-19 Social Distancing on Twitter. *Open Forum Infectious Diseases*, Volume 7, Issue Supplement_1. doi:<https://doi.org/10.1093/ofid/ofaa439.679>
24. Sarracén, G. L. (n.d.). Multilingual and Multimodal Hate Speech Analysis in Twitter. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*.
25. doi:<https://doi.org/10.1145/3437963.3441668>
26. Singh, P., Singh, S., Sohal, M., Dwivedi, Y. K., Kahlon, K. S., & Sawhney, R. S. (2020). Psychological fear and anxiety caused by COVID-19: Insights from Twitter analytics. *Asian J Psychiatr*.
27. doi:10.1016/j.ajp.2020.102280
28. Trajkova, M., Alhakamy, A., Cafaro, F., Vedak, S., Mallappa, R., & Kankara, S. R. (2020). Exploring Casual COVID-19 Data Visualizations on Twitter: Topics and Challenges. *Informatics*.
29. doi:<https://doi.org/10.3390/informatics7030035>
30. Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*. doi:10.2196/20550