ORIGINAL ARTICLE

COVID-19 vaccination perception and outcome: society sentiment analysis on twitter data in India

Anushtha Vishwakarma[1] · Mitali Chugh[1]

Abstract

This study examines the perceptions and results of COVID-19 immunization using sentiment analysis of Twitter data from India. The tweets were collected from January 2021 to March 2023 using relevant hashtags and keywords. The dataset was pre-processed and cleaned before conducting sentiment analysis using Natural Language Processing techniques. Our results show that the overall sentiment toward COVID-19 vaccination in India has been positive, with a majority of tweets expressing support for vaccination and encouraging others to get vaccinated. However, we also identified some negative sentiments related to vaccine hesitancy, side effects, and mistrust in the government and pharmaceutical companies. We further analyzed the sentiment based on demographic factors such as gender, age, and location. The analysis revealed that the sentiment varied across different demographics, with some groups expressing more positive or negative sentiments than others. This study provides insights into the perception and outcomes of COVID-19 vaccination in India and highlights the need for targeted communication strategies to address vaccine hesitancy and increase vaccine uptake in specific demographics.

Keywords Twitter · Sentiment analysis · Vader-lexicon · Random forest · SVM · LSTM

1 Introduction

The World Health Organization declared COVID-19 a pandemic in March 2020, and it influences the social, economic, and health sectors worldwide. Its rapid spread has led to the extensive use of online social media for sharing and broadcasting pandemic news and social communications specifically during lockdowns and quarantines. Hence, substantial data on COVID-19 exist and are even generated at present on social media. Its analysis can facilitate getting important cues about worries and perceptions about the pandemic.

Soon after the pandemic began, it became apparent that vaccination is the first viable means of preventing COVID-19 from spreading and making it easier to build up herd immunity. The overwhelming global efforts for vaccine development within a year made it possible to get WHO approval to immediately use different vaccines such as Moderna, Pfizer Biotech, and AstraZeneca. To conduct a successful worldwide campaign for vaccination public support appeared decisive and understanding the people's opinions and sentiments, and preparedness to be immunized was significant to take suitable measures for improving assurance about immunization tools. The social media platforms have vaccine-related conversations that support monitoring the factors influencing trust and assist strategies to improve it. The significance of social media in such shared conversations is increasing due to the growing presence of medical practitioners and healthcare professionals, with an active role in the discussions.

Because of the factors mentioned, the current study aims to evaluate the opinions and sentiments of Indian Twitter users against COVID-19 immunization by examining postings gathered from Twitter using Sentiment Analysis (SA) techniques (2021). India, the second-most populated nation on earth, has witnessed a significant surge in COVID-19 cases, with a high number of fatalities. As of March 2023, India has administered over 1.5 billion vaccine doses, covering a significant portion of the population. (2022) However, vaccine hesitancy and misinformation remain challenges in achieving full vaccination coverage in India. Understanding public sentiment can help health authorities and policymakers in developing

✉ Mitali Chugh
  mchugh@ddn.upes.ac.in

[1]   UPES, Bidholi, Dehradun, Uttarakhand 248001, India

targeted strategies for promoting vaccine uptake and addressing vaccine hesitancy (2022). To gain an understanding of public perception and behavior toward the COVID-19 vaccine and to answer the following research questions, the study would examine Twitter data about COVID-19 immunization in India.

(1) How did the sentiment toward COVID-19 vaccination evolve in India from 2020 to 2022?
(2) What factors contributed to the positive and negative sentiment toward COVID-19 vaccination during the vaccination drive in India?
(3) How did the expansion of the COVID-19 vaccination drive in India to cover all age groups affect the sentiment toward vaccination?
(4) What role did effective communication and efficient execution of the vaccination process play in maintaining positive sentiment toward COVID-19 vaccination in India?
(5) How can the sentiment toward COVID-19 vaccination in India be continuously monitored to identify and address any challenges or concerns that may arise during the vaccination drive?

The study outlines a methodology for analyzing a dataset of tweets related to vaccinations that were published in India between August 2020 and December 2022 to answer research questions for the evaluation of public opinions about vaccination and to improve insights into sentiments toward the COVID-19 vaccination.

To analyze social media data, the proposed methodology combines Natural Language Processing (NLP) and Sentiment Analysis methodologies. In this study, we have performed a qualitative analysis which is a needful step since the dataset used in work is quite noisy so have to perform data preprocessing to enhance the quality of text for further analysis, whereas the quantitative approach in the study was to classify tweets into three different classes that are positive, negative, and neutral. Categorization is performed by using predefined themes or codes to identify patterns and themes within the text.

The sentence describes how sentiment analysis is carried out utilizing the Lexicon Vader and sentiment intensity analyzers. The Lexicon Vader contains information about words, and the sentiment intensity analyzers classify text into different categories such as positive, negative, and neutral. The compound polarity score is then calculated by the sentiment intensity analyzer, which determines the polarity of each text. This process helps to analyze the sentiment of the text and understand the attitudes and feelings of people toward a particular topic or phenomenon.

The following goals are achieved by the current study using tweets from the social network Twitter in India.

(1) Analyze the public views to find sentiments toward the COVID-19 vaccination.
(2) Examine how these feelings have evolved throughout the years, starting at the beginning of the immunization campaign and ending after the booster dosage.

It will be helpful for policymakers, health authorities, and vaccine manufacturers to understand the challenges and opportunities in promoting COVID-19 vaccination in India by using the study's findings to gain valuable insights into how the general public in India feels and behaves toward COVID-19 vaccination.

The present study is structured as follows. In Sect. 2, an overview of studies related to COVID-19 and social media is presented. Section 3 details the proposed methodology and the tools used for analysis. In Sects. 4 and 5, the results obtained are shown and discussed. Section 6 concludes the study following the limitations of the study in Sect. 7.

2 Literature survey

Since COVID-19 infected people quickly, there has been an increase in research studies examining the connection between social media and vaccinations and how users' awareness of the latter was impacted by their use of the former as a source of information. As a result, the notable studies carried out from this perspective are detailed below, including both a global and an Indian perspective.

Saleh et al. (2021) examined the tweets over 10 months in the year 2020, all through the development phase of the vaccine, linking SA techniques and demographic information. The finding shows in general unstable positive trends dependent on news events. In its place, Lyu et al. (2021) hoped to use techniques like Latent Dirichlet Allocation for SA to aggregate important issues and viewpoints in a collection of tweets between 2020 and 2021 to achieve herd immunity.

Yousefinaghani et al. (2021) tweets were analyzed from January 2020 to January 2021 using VADER and ranked them as positive, negative, and neutral. They have discussed equally positive and negative reactions to the commencement of the vaccination campaign and the growing ambiguity around the vaccination's delivery.

Jabalameli et al.(2022) aimed to examine the online conversation on Twitter during the initial COVID-19 immunization deployment in Ohio and Michigan. The study looked at how people felt about the pandemic, the most common conversations, and how those conversations were

spread geographically. Based on a chronological examination of the tweets, the study also examined the effects of state government reactions and significant news on public opinion. Overall, the study's findings can be used to evaluate public demands and reactions, track the effects of county-level municipal policy, and prepare for future pandemic responses. The study emphasizes the value of social media in understanding public needs and opinions during a pandemic and provides informative data on the people's attitudes and opinions at the beginning of the COVID-19 vaccination deployment.

Qorib et al. (2023) The goal was to use social media data from open, live-streamed tweets found using an Application Programming Interface (API) search to analyze COVID-19 vaccine resistance. Five learning algorithms—Random Forest, Logistics Regression, Decision Tree, Linear SVC, and Nave Bayes—were utilized in the study to compute sentiment in various combinations with three vectorization techniques, including Azure Machine Learning, VADER, and Text Blob (Doc2Vec, Count Vectorizer, and TF-IDF). Lemmatization, potter stemming alone with lemmatization, and potter stemming alone with lemmatization were the other three vocabulary normalization methods employed in the study. The best outcomes were achieved by combining the Text Blob sentiment score, TF-IDF vectorization, and Linear SVC classification model. The study also discovered that the accuracy of the model is decreased when two vectorizations (Count Vectorizer and TF-IDF) are combined. The study also emphasizes how crucial it is to choose the right vectorization techniques, vocabulary normalization approaches, and sentiment computing techniques for reliable sentiment analysis of social media data.

Purwitasari et al. (2023) study provides a dataset of tweets from Indonesia about the COVID-19 vaccination that has been annotated with stance and aspect-based sentiment data. The dataset consists of nine thousand tweets that were thoroughly analyzed by three independent analysts and collected each month from January to October 2021 using certain keywords. Each tweet is annotated with three viewpoint labels and seven preset elements relating to tweets about the Indonesian COVID-19 vaccine. To find a long-term cure for the COVID-19 pandemic, the paper emphasizes the significance of examining and keeping track of public attitudes and positions on social media, particularly concerning vaccine-related themes. Among other forms of research, the dataset can be utilized for stance detection and aspect-based sentiment analysis.

Bokaee Nezhad & Deihimi (2022) study concentrated on examining Persian tweets from April through September 2021 that discussed the COVID-19 immunization program in Iran. The research aimed to evaluate Iranians' opinions on domestic and foreign vaccines and to learn how they felt about receiving the COVID-19 vaccine. According to the study, there was a slight variation in the percentage of people who had favorable feelings toward domestic and foreign vaccines, with the latter having the predominately positive polarity. Both types of immunizations seemed to be receiving more unfavorable press in recent months. The percentages of overall good and negative sentiments regarding vaccination among Iranians did not, however, differ much. The study concludes that to raise vaccination rates and put a stop to the epidemic, public health organizations should concentrate on spreading supportive messages and reducing negative ones on social media sites like Twitter.

Turón et al. (2023) study proposes a novel approach that combines multivariate statistical techniques with machine learning techniques, such as sentiment analysis using lexicons, to evaluate the evolution of social mood during the COVID-19 immunization protocol in Spain. The study examines 41,669 Spanish tweets written between February 2020 and December 2021 to identify the various attitudes represented in them using a list of Spanish phrases and their relationships with eight primary emotions and three valences. The results demonstrate the social atmosphere of the population, record the many opinion clusters that are formed, gage public sentiment via collective valence, and identify the predominance of various emotions during the several vaccination stages. The study concludes that combining subjective and objective data allows for a more realistic depiction of social reality and a more efficient approach to issue-solving. The results of this study may be useful to researchers who are interested in sentiment analysis and multivariate statistical methods used to social media data in the context of the COVID-19 immunization program in Spain.

Rahmanti et al. (2022) goal of the study is to determine whether vaccination rates, case growth, and case fatality rates in Indonesia are correlated with opinions regarding the COVID-19 vaccine expressed on Twitter. The Ministry of Health (MoH) and the KawalCOVID19 database's official websites were used by the researchers to gather data on the daily trends of COVID-19 vaccine coverage, case growth rates, and case fatality rates. Also, they obtained tweets from Indonesian Twitter users between October 15, 2020, and April 12, 2021, on the COVID-19 vaccine. To ascertain the thoughts, feelings, word choices, and tendencies of the general people, they examined all filtered tweets sent 90 days before and 90 days after Indonesia's national vaccination deployment. The study also discovered an upward trend in vaccination sentiment scores that was statistically significant and positively connected with growing vaccination coverage.

Catelli et al. (2023) analyze the attitudes and opinions of the Italian population concerning COVID-19 immunization, the paper introduces an innovative approach based on NLP and SA. After filtering 1,602,940 tweets with the term "vaccine," the study's analysis of a dataset of tweets

about vaccinations published in Italy between January 2021 and February 2022 yielded 353,217 analyzed tweets. The study's methodology and findings add to the corpus of information on leveraging social media data to assess public views and opinions on COVID-19 vaccination. Future research might employ these techniques to conduct more precise sentiment analyses of the public's attitudes about vaccination. Opinion holders could be divided into several user categories, and domain-specific lexicons could be used in sentiment analysis. Also, the report emphasizes the significance of addressing certain incidents that may have an impact on public views toward vaccination and the implementation of efficient communication tactics to combat unfavorable perceptions.

S. Praveen et al. (2021a) according to surveys, 47% of social media posts on vaccinations were neutral, while 17% of those about the COVID-19 vaccine were critical. Fear of health problems and vaccine allergies are the two main issues that Indian citizens have with the COVID-19 vaccination. For the study, the data is collected from social media posts made by Indian residents about the COVID-19 vaccine using Python. To ascertain how the general opinion of the Indian population toward the COVID-19 vaccination evolves over many months of the COVID-19 crisis, we conducted a sentimental analysis in Study 1. To further understand the main concerns that the general public has about the COVID-19 vaccination, we used topic modeling in Study 2.

SV et al. (2022) researched over, sentiment analysis and topic modeling from natural language processing used to examine how Indians view the COVID-19 booster dosage vaccine. For this study, we examined tweets created by Indian citizens. The Indian government accelerated the COVID-19 booster dose immunization process as of late July 2022. Out of the 76,979 tweets, the sentiment analysis study found that more than half ($n = 40,719$ tweets, or 52.8%), 24,242 tweets, or 31.5%, had neutral attitudes, and 12,018 tweets, or 15.6%, had positive sentiments.

Dumre et al. (2021) COVID-19 vaccines are made available in India, and opinions concerning them are starting to take shape. The study's major goal was to use data analysis tools to analyze survey responses and come to certain conclusions. To determine what prevents people from getting immunized, we conducted a Sentimental Analysis of the participants' responses for this research.

Paliwal et al. (2022) research examines opinions posted on Twitter about the ongoing COVID-19 immunization campaign in India. The study acknowledges that people used social media to voice their worries and feelings during the pandemic and lockdown. The study focuses on issues like the necessity of immunization, how it is administered, and vaccine confidence. Both a vocabulary-based approach and a machine learning method were used to analyze the sentiment

of the collected tweets after they had been pre-processed to eliminate noise and unnecessary data. According to the study's findings, the majority of tweets either indicated positive or neutral views, and this conclusion holds for both techniques of analysis.

Sv et al. (2021) used posts from social media, the study investigates how Indian citizens feel about the COVID-19 vaccine's negative effects. Two investigations were carried out after social media posts focusing on the COVID-19 vaccine's negative effects were gathered and processed using Python. In the first study, emotional analysis was used to determine how Indian citizens felt generally about the COVID-19 vaccine's negative effects. In the second trial, topic modeling was used to examine the main side effects that people who received the COVID-19 vaccination reported experiencing. According to the findings, 78.5% of tweets made by Indian residents about the COVID-19 vaccine's negative effects were either neutral or favorable. The topic modeling research discovered that the two main factors influencing Indian individuals' negative attitudes regarding the COVID-19 vaccine's side effects were fear of death and fear of efficiency at work. The study's findings support the need for the Indian government to vigorously promote vaccination among its people while simultaneously addressing their worries and fears through programs of education and awareness.

S. V. Praveen et al.(2021b) Machine learning is being used in the study to look at how the general people perceived stress, trauma, and worry during the COVID-19 outbreak in India. Python was used in the study's analysis of 840,000 tweets gathered between March 2020 and June 2020. The tweets were cleaned up and pre-processed before machine learning methods were used to determine the sentiment and topics covered. The sentiment analysis revealed that the majority of the tweets had negative attitudes, and the results indicated that worry, stress, and trauma were the main issues mentioned. The article indicates that the COVID-19 pandemic has significantly impacted Indian society's mental health, causing elevated levels of stress, trauma, and anxiety.

Mudassir et al.(2021) the study describes the perspective of Indian residents toward the COVID-19 vaccine is examined, and it is noted that the COVID-19 outbreak coincided with a spike in traffic to social media sites. People's alternatives and platforms for expressing their opinions are limited, which is the cause of this. Using this, the author performs sentiment analysis on English tweets sent by users in India using three different models and then selects a deep learning model after evaluating the findings.

Melton et al.(2021) purpose of the paper was to look into the attitudes and topics being raised by users in COVID-19 vaccine-related Reddit discussions. The study analyzes textual data gathered from 13 Reddit forums with

a focus on the COVID-19 vaccination between December 1, 2020, and May 15, 2021, using sentiment classification and latent Dirichlet-based topic modeling. According to the data, these communities have consistently exhibited a more positive than a negative attitude about discussions about vaccinations. Instead of discussing conspiracies, the group members mostly concentrated on addressing adverse effects. To ease the adoption of suitable messages, digital interventions, and new legislation to boost vaccine trust, the study emphasizes the necessity of analyzing public mood and topic modeling around vaccines.

Sakthi Kumaresh (2021) study focuses on the sentiment analysis of tweets about the India-delivered COVID-19 immunization. It is simpler to gage public opinion when utilizing machine learning algorithms to analyze Twitter data. The tweets are classified as favorable or negative using machine learning techniques like logistic regression and the Naive Bayes algorithm. When utilizing logistic regression, classification accuracy was 84.8%, compared to 83.7% when using the naive Bayes algorithm.

Rani & Jain. (2023) research study shows that the coronavirus pandemic has raised numerous healthcare issues. It is extremely amazing how hard healthcare specialists have worked over the last two years to create a variety of vaccines to fight this illness. This paper proposes a novel architecture dubbed a deep fusion model (DFM) with a meta-learning ensemble technique based on sentiment analysis of public opinions on the COVID-19 vaccinations and the omicron version on Twitter. The proposed method combined natural language processing with deep learning models like LSTM, GRU, CNN, and their many combinations. The study aims to determine how the general public feels about COVID-19 vaccinations and the omicron version. In addition, the trial showed effectiveness with accuracy that might reach 88% when compared to cutting-edge models.

3 Methodology

People express their opinions in the form of text, and their sentiments and feelings are been expressed through such text. Sentimental analysis is used for the classification and identification of sentiments. Social media platforms such as Twitter in which people tweet their opinions regarding different genres to express their thoughts and views to describe the situation according to their feelings. From user-generated data using sentimental analysis, it can classify it into positive, negative, and neutral. It is a technique that analyzes people's thoughts, and feelings. Text mining here it's been used for the extraction of useful information
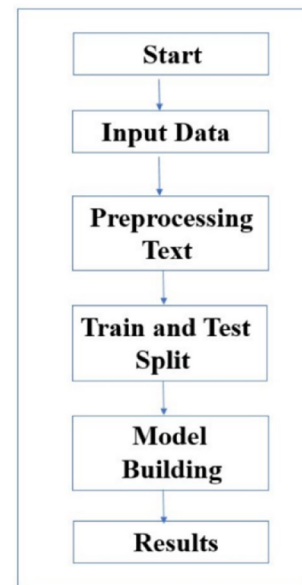

Fig. 1 Block diagram of work-flow

from text. In the field of natural language processing sentimental analysis which is also known for opinion mining, sentimental analysis is a subcategory that is used for the categorization of opinions under positive, negative, and neutral.

We are following this flow work in Fig. 1 here to carry on our methodology.

3.1 Data analysis and data collection

The Twitter dataset contains different columns which describe user names, user descriptions, user locations, and tweets by the user regarding vaccinations. Here the shape of the data is (399,647, 13). By looking into the source column which describes various sources from where tweets have been published and visualize it according to tweet count. For getting information about which location has the highest number of tweets it visualizes according to tweet counts.

Here we have used a large dataset from Kaggle named vaccination_all_tweets. It contains 399,647 tweets which are further used for data preprocessing to get some good insights out of it. The dataset consists of tweets collected based on the #CovidVaacine hashtag. The dataset consists of the reviews that are being written by humans over COVID-19 vaccination analysis. The Twitter dataset contains a large volume of unstructured data. To get some insights and information from the dataset, first, perform Exploratory data analysis [EDA] which describes the shape of the overall data frame, and the type of data such as an object and integer.
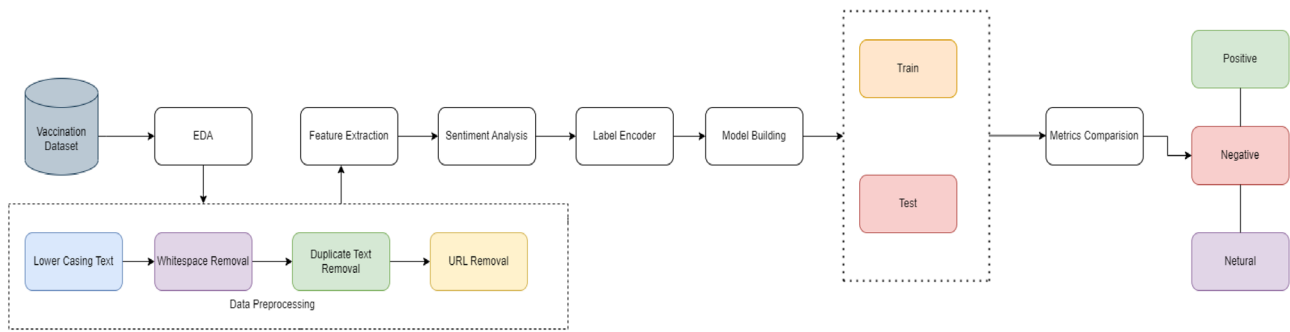
Fig. 2 Proposed methodology

Checking for duplicate values and unique values and then performing visualization concerning tweets made by people (Fig. 2).

3.2 Data pre-processing

Hereafter exploratory data analysis data preprocessing step includes NLTK which natural language toolkit. NLTK classifies and categorizes data. Here we are using Vader Lexicon for sentiment analysis which is being used as a rule-based tool. Vader model which is available for NLTK packages, which directly applied to labeled text. Lexical features map to intensities of emotions called sentiment score. The NLTK package categorizes data and classifies it and Vader Rule-based rule is used for calculating the sentiments of tweets.

Lexicon Vader in which it contains information about words. Then sentiment intensity analyzers are used which classify the text under different categories such as positive, negative, and neutral. The compound polarity score is calculated in sentiment intensity analyzers which classify each text polarity.

Here then by importing a sentiment intensity analyzer which analyzes the emotions of a particular text, and classifies the text into positive, negative, and neutral. The sentiment intensity analyzer classifies text and also calculates compound scores of texts which are the sum of positive, negative, and neutral scores. Which is categorized into $+1$ and -1 the compound score near $+1$ is classified as the most extreme positive, and the compound score near -1 is classified the as most extreme negative. Here by sentiment compound score when equal to zero categorizes the sentiment type as neutral, when less than equal to zero the sentiment type is, negative, and greater than equal to zero then the sentiment type is positive.

Then in data preprocessing cleaning of text is there in which it first converts all text in the lower text by using the lower () method in Python. Then by removing most repeated patterns by the replace method. In which removal of special character URL, non-ASCII characters white space, etc. Then the clean text is retrieved after preprocessing and the text is classified row-wise as positive, negative, and neutral which has been calculated by sentiment intensity analyzers. After that data frame is clean and in a structured format.

The most important step while implementing text analysis is to break the text in form of tokens. Tokens in which individual list is created which stores each text. Then again removal of special characters from clean text if exist any in the data frame. Then stop word removal is where most supporting words are removed. Stop word removal also removes words that do not have any appropriate meaning such as (a, the, if, any). Then using the final text while creating a word cloud for the most frequent word used while writing a positive review or negative review about the vaccines.

3.3 Feature extractions

Then applying label encoder which works on categorical data and is used for the normalization of text. It formats labels into machine readable form where it converts labels into integer format. It is the most important step for applying supervised learning to structured data. By importing the SK learn library which is a tool for label encoding. It classifies text into zero, one, and two as per the text is classified as example 0 as negative, 1 as neutral, and 2 as positive.

3.4 Model building

NLP uses a variety of language models, including deep learning and machine learning models. The many algorithms employed include LSTM, Random Forest, and Support Vector Machine. After data preprocessing the data received are cleaned of all the noise and unwanted characters making it use full to work for model building.

3.4.1 Long short-term memory

RNN easily works on short-term memories. In RNN it temporarily stores the first word and after a new word is added the previous one gets replaced. Then In LSTM, it introduces a memory unit where it stores the words memory unit also called cells. LSTMs are created to prevent the vanishing gradient problem.

LSTM is a chain-like structure in which it has a repeating module that has different structures. LSTM is divided into three stages Input, Forget gate and output. The first stage is the input stage where useful information is stored and that information is regulated using the sigmoid function. Then in forget gate, the information which is no longer needed is removed from the cell.
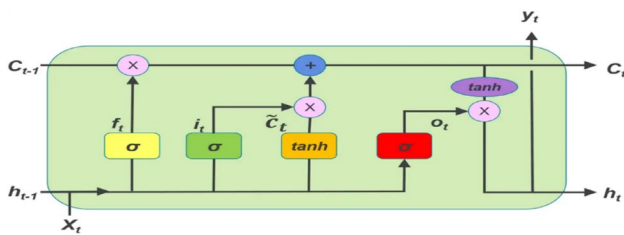
Functioning of LSTM (Divyanshu Thakur 2018).

There is two input first input at a particular time x-1 and the second input which is the output of the previous cell that is h_t-1 multiplied with weight matrices. Then the results are passed into the activation function which returns either 0 or 1. Then if 0 returns it is multiplied with the previous cell information lost and if it returns 1 it will be used in the future. In the output stage, the task is extracting information from the current state.

3.4.2 Random forest

Supervised learning algorithms such as random forest work for text classification. As known, this classification algorithm is made up of trees. It creates a tree of data sets in which it predicts the results by voting. Random Forest works very well for large datasets rather than decision trees. It first randomly selects samples from the data frame, then constructs a decision tree to get prediction results. Then voting is performed for predicted results and lastly selects final predicted results. Random Forest works for multi-class labels for text classification.

3.4.3 Support vector machine

SVM also called the Support vector machine works for both regression and classification problems. SVM usually segregates classes such as hyperplanes. In SVM it separates data concerning labels. Basic work of SVM in which it separates data with hyperplanes. SVM works on the following

Fig. 3 Bar chart of tweets according to Location

Fig. 4 Comparison of most common words between sentiments

concepts such as support vectors, hyperplane, and margins. In SVC classification implementation can be done by importing modules such as sklearn.svm.svc.

3.4.4 Model results

The results of model training on a large dataset for text classification into different classes, to predict sentiment. Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory were three different algorithms that were used (LSTM). The accuracy of the SVM algorithm, which served as the basic model, was 71.2%. The Random Forest algorithm was then used, which created a tree-like structure of individual texts and predicted sentiment based on probabilities. This algorithm achieved an accuracy of 72.2%. Finally, the LSTM algorithm was used, which is an extension of the Recurren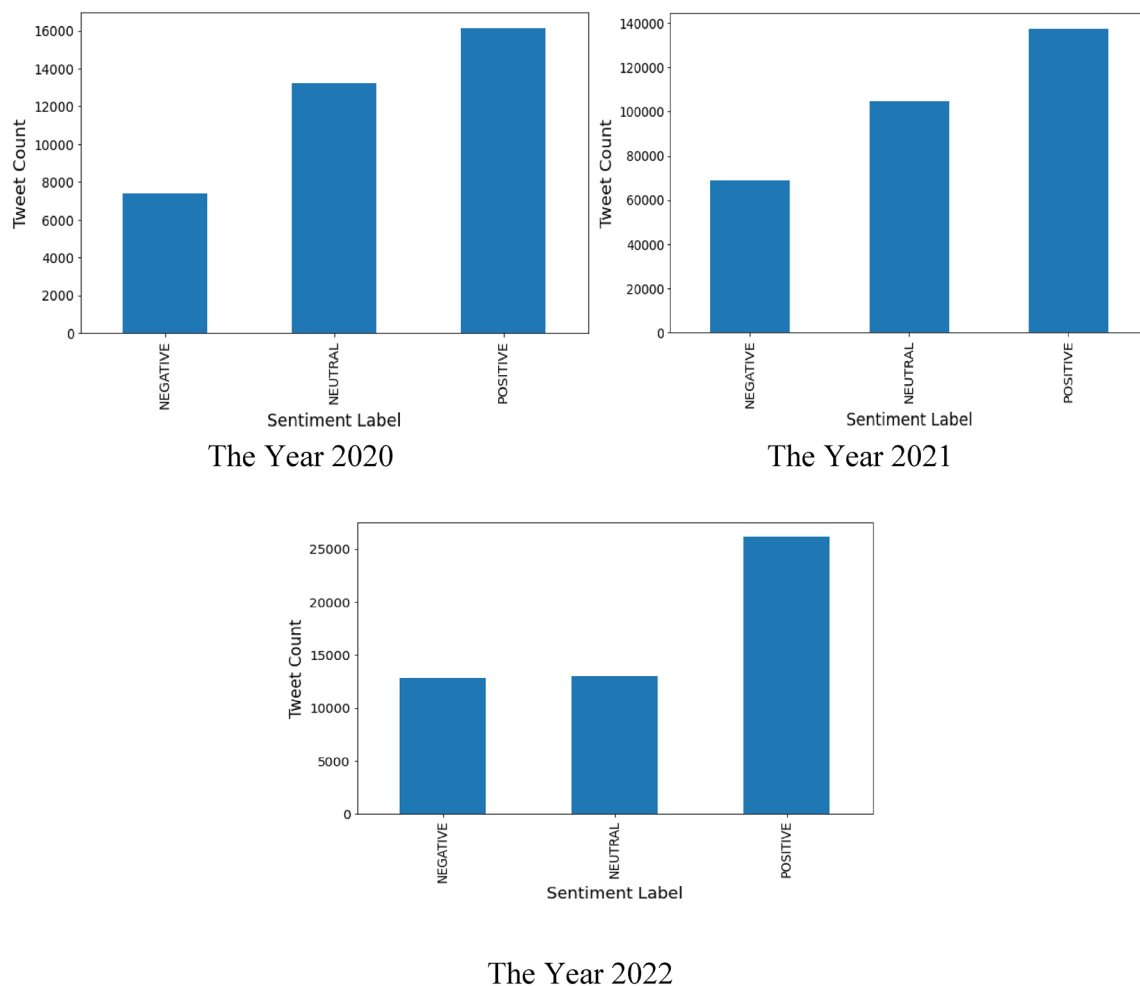t Neural Network (RNN) model that solves the vanishing gradient problem in text classification. The LSTM algorithm achieved an accuracy of 88.6%, which was the highest among the three algorithms. Based on these results, the paragraph concludes that deep learning models like LSTM provide better accuracy in identifying the sentiment of text, compared to traditional Machine Learning models like SVM and Random Forest. These results highlight the importance of using advanced deep learning models for text classification tasks, especially when dealing with large datasets.

4 Results

With the advancement of technology, people are using social media like Twitter which is one of the most enhanced technologies to spread news and articles with almost no cost to spread true or false information to misguide people. Twitter is one major online platform to connect a large population at once which helps people to know about the current situation and the latest news. Hence this can also be harmful if some misinformation being published will reach a large population which can create chaos.

In the above-proposed framework, we have worked on the sentiments of people toward vaccines.

The first step for text mining to get some useful insight is EDA which analyzes information and useful data to get conclusions about people's opinions.

Society sentiment analysis on Twitter data in India, Fig. 3 is a bar chart that shows the distribution of tweets related to COVID-19 vaccination across different cities in India. The chart shows that Bengaluru has the highest number of tweets related to vaccination, followed by Delhi. This information can be useful in understanding the level of public awareness and engagement with COVID-19 vaccination across different cities in India.

In Fig. 4, it is a comparative word cloud where people have expressed their likings and dislikings about the different vaccines for this they have used some frequent words to justify their emotions regarding the vaccination. The commonly taken vaccines are Pfizer, covaxine, and Covidshield. When people were counting on the positive side of vaccines to provide more information to other people they used words like "Approved"," effective"," Availability" which express that the vaccination was a success for a group of people who thinks that this vaccine was a huge savior for life in the such difficult phase of life. Whereas few have just opposite thinking as they were cases where people died even after taking a second dose of vaccine for that social media as a medium to aware more number of people that vaccination is not a complete solution to deal with the deadly virus for

The Year 2020

The Year 2021

The Year 2022

**Fig. 5** Sentiment classification on basis of year

that they used some common words like "health"," effects"," sore"," problem". Neutral comments consist of words that justify how many people as individuals have taken the vaccine which includes words like "Free"," fully"," today" etc. which express that some have taken their while some have expressed on which date they got vaccines.

In Fig. 5, the data are classified into different years from the collected dataset in which the year 2020 the sentiments of people are positive regarding vaccines because the motive behind the vaccine is to decrease mortality of virus its very effective for some people in 2020 the positive tweets were 16,159 although it is starting of waves and vaccination trials are going on. In 2021 positive sentiments of people increased by 137,632 in this year's processes of vaccination for all age groups started in which there are negative reviews increases compared to 2020. As in 2022, the count off of positive opinions was more in comparison to the negative ones since the vaccination process is still in a continuous phase for both people who are taking the vaccine also

for those who are developing it. As the recent drive is for booster dose which is been provided to those people who have completed two doses of vaccination. Therefore, the opinion of the patient varies periodically depending on the effectiveness of the vaccine on the mass population. Fig. 6

Here we can see a description of the whole dataset where user location, user_description, etc. are there respective to their tweets. From here we can take text and the text in the input framework and by prediction, it predicts where the text is positive, negative, or neutral.

Society sentiment analysis on Twitter data in India is a screenshot that shows the Sentiment Prediction tool. The tool allows users to enter any text related to COVID-19 vaccination, such as tweets or news articles, and receive a prediction of the sentiment of the text as either positive, negative, or neutral. The screenshot in Fig. 7 shows an example of how the Sentiment Prediction tool can be used to analyze the sentiment of a tweet related to COVID-19 vaccination. In this example, the tool was used to analyze a tweet that
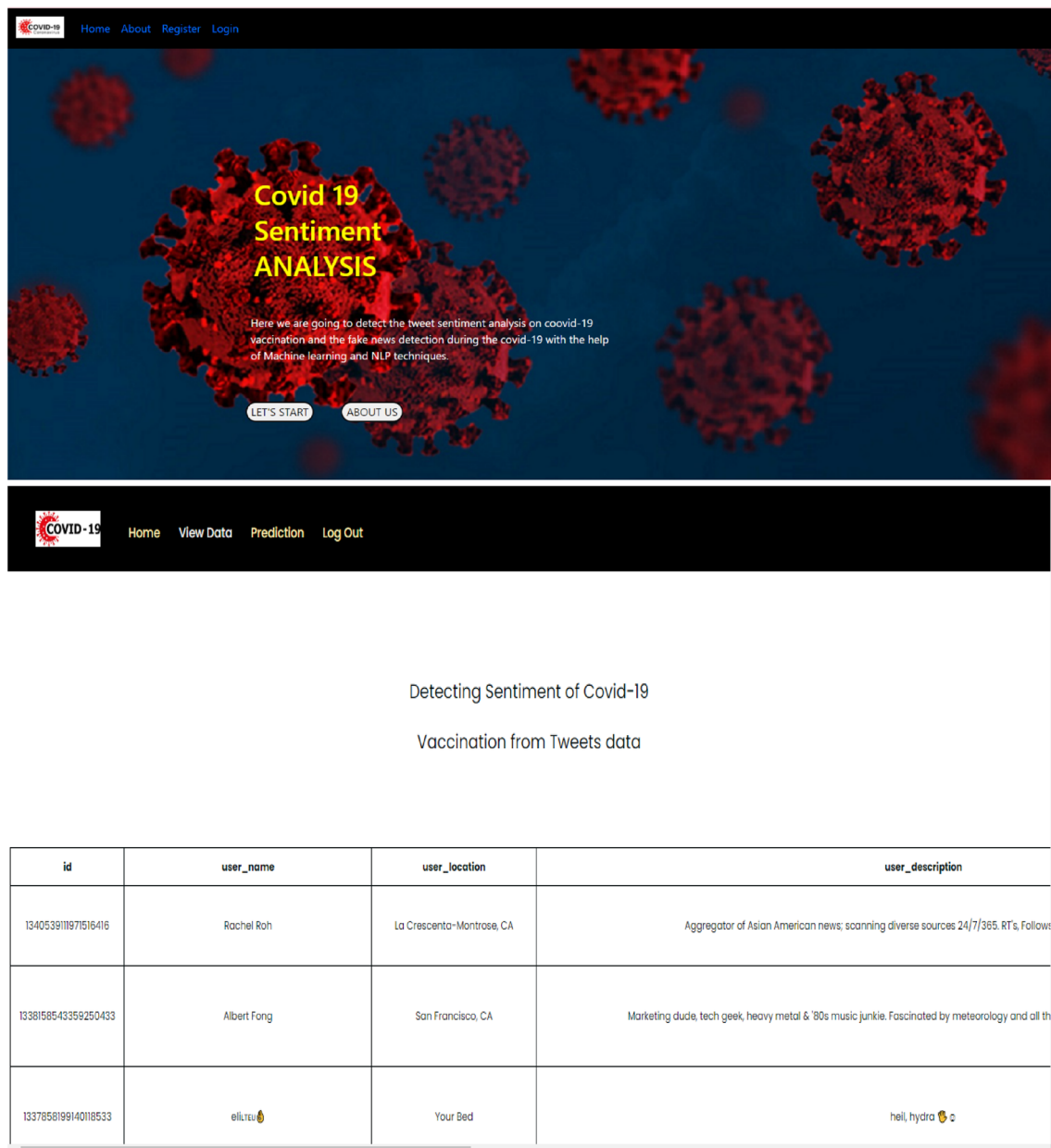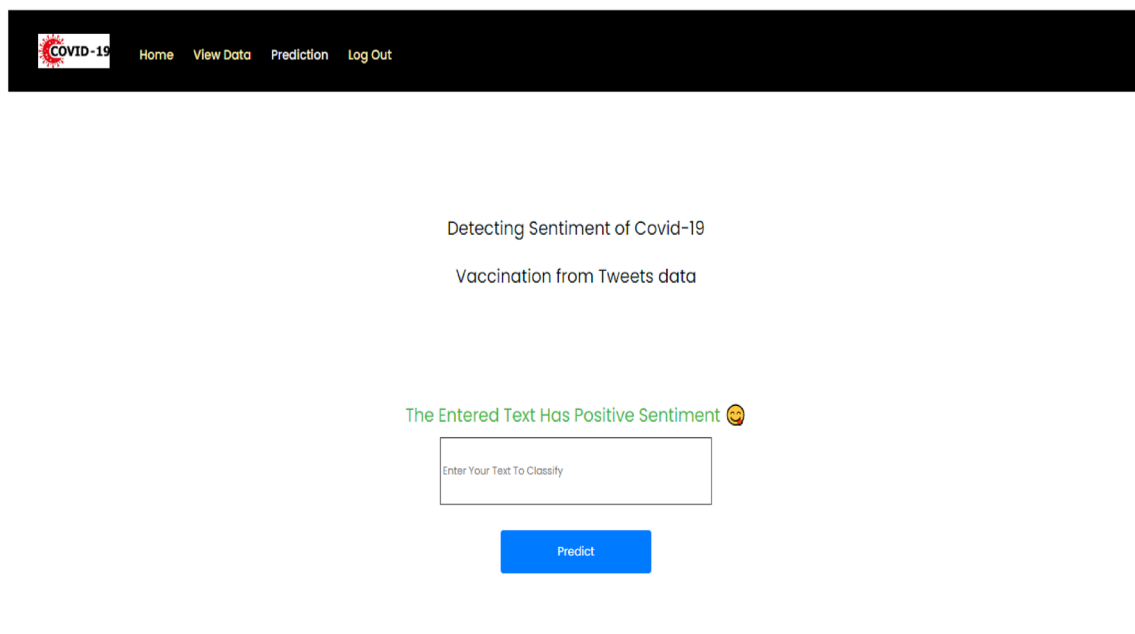
Fig. 6 Depiction of the screenshot showing the view page

reads "Just got my COVID vaccine shot! Feeling relieved and grateful to science and healthcare workers." The tool correctly predicted the sentiment of the tweet as positive.

5 Discussion

The study found the sentiment toward COVID-19 vaccination was overwhelmingly positive in 2020 when the vaccination drive had just begun. This positive sentiment was likely driven by the urgency of the situation and the hope that the vaccine would help to reduce the mortality rate of

the virus. However, as the vaccination drive progressed and more people were vaccinated, the sentiment toward vaccination became more nuanced. In 2021, the positive sentiment toward vaccination increased significantly as the vaccination process was expanded to cover all age groups. However, this expansion also led to an increase in negative reviews, likely due to challenges in the vaccination process such as long waiting times, insufficient vaccine supply, and difficulties in accessing vaccination centers. Interestingly, the data also suggest that in 2022, the sentiment toward vaccination remained positive even though the vaccination process was still ongoing. This could be because people were becoming

Fig. 7 Depiction of the screenshot showing sentiment prediction

more familiar with the vaccination process and were experiencing its benefits. Overall, the analysis presented in the paper highlights the importance of effective communication and efficient execution of the vaccination process in maintaining positive sentiment toward COVID-19 vaccination. It also underscores the need for continuous monitoring of the sentiment toward vaccination to identify and address any challenges or concerns that may arise during the vaccination drive. Overall, the study provides valuable insights into the sentiment toward COVID-19 vaccination in India and its outcomes. The study highlights the importance of effective communication about the efficacy and safety of the vaccine and the availability of the vaccine in promoting positive sentiment toward vaccination. The study also highlights the need for a smooth and efficient vaccination process to avoid negative sentiment toward vaccination.

6 Conclusion

In the era of fast-growing technology, social media can connect people making them rely on it to make life easy and people use it for spreading information and the latest news. As a huge volume of people is connected this helps in spreading the information fast. Twitter tweets on the COVID-19 vaccine were subjected to sentiment analysis in this study. Although there are some unfavorable opinions about vaccinations, the study overwhelmingly reveals support for it. The results comparison uses a variety of machine learning techniques for categorization. LSTM displays the highest level of accuracy for the dataset. But the proposed strategy's primary drawback is the prevalence of tweets posted in regional dialects like Hindi. We did not include any of the Hindi terms in the positive and negative corpus because they are written in English. Hence, any tweets written in Hindi but translated into English are regarded as neutral. These terms might be added in the future, improving the accuracy of the sentiment analysis. Similar methods can be utilized in future work to lower the neutral count. This classification can be further used for categorizing the text as fake or real. Whether the text published over different social media is done by real authors or is being published for fake agendas. Classifiers are the two data mining classifiers we have chosen to employ in this research. There are further classifiers, including the C4.5 classifier, the Bayesian network classifier, and the Neural Network classifier. Such classifiers might be used going forward to provide more data to compare with because they were not included in this work.

7 Limitation of study

Twitter users may not be representative of the broader population in India. Because Twitter users are typically younger, better educated, and wealthier than the overall population, the results may not apply to other demographics. The sentiment analysis algorithms used in the study may not capture the nuances of language and cultural context in India, which

may lead to bias in the classification of tweets into positive, negative, or neutral categories. Additionally, the accuracy of sentiment analysis tools can be affected by the quality of training data and the context in which the data was collected.

Author contributions   Ms. Anushtha Vishwakarma has worked on the guideline provided by Dr. Mitali Chugh and has performed the analysis for data. Manuscript has been reviewed by both authors.

Declarations

Conflict of interest   The authors declare no competing interests.

References

Bokaee Nezhad Z, Deihimi MA (2022) Twitter sentiment analysis from Iran about COVID 19 vaccine. Diabetes Metab Syndr: Clin Res Rev. https://doi.org/10.1016/j.dsx.2021.102367

Catelli R, Pelosi S, Comito C, Pizzuti C, Esposito M (2023) Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on twitter in Italy. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2023.106876

Divyanshu Thakur (2018) https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af

Djuric O, Ottone M, Vicentini M, Venturelli F, Pezzarossi A, Manicardi V, Greci M, Giorgi Rossi P (2022) Diabetes and COVID-19 testing, positivity, and mortality: a population-wide study in Northern Italy. Diabetes Res Clin Pract 191:110051. https://doi.org/10.1016/j.diabres.2022.110051

Dubey AD (2021) Public sentiment analysis of COVID-19 vaccination drive in India. SSRN Electron J. https://doi.org/10.2139/ssrn.3772401

Dumre R, Sharma K, Konar K (2021) Statistical and sentimental analysis on vaccination against COVID-19 in India. In: 2021 international conference on communication information and computing technology (ICCICT). IEEE, pp 1–6

Jabalameli S, Xu Y, Shetty S (2022) Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: at the early stage of vaccination. Int J Disaster Risk Reduct. https://doi.org/10.1016/j.ijdrr.2022.103204

Lyu JC, Le HE, Luli GK (2021) Covid-19 vaccine-related discussion on twitter: topic modeling and sentiment analysis. J Med Internet Res. https://doi.org/10.2196/24435

J MK (2021) Syndicate. J Manag. Vol 21. ISSN 2278-8247

Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A (2021) Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: a call to action for strengthening vaccine confidence. J Infect Publ Health 14:1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010

Mudassir MA, Mor Y, Munot R, Shankarmani R (2021) Sentiment analysis of COVID-19 vaccine perception using NLP. In: 2021 third international conference on inventive research in computing applications (ICIRCA). IEEE, pp 516–521

Paliwal S, Parveen S, Afshar Alam M, Ahmed J (2022) Sentiment analysis of COVID-19 vaccine rollout in India. pp 21–33

Praveen S, Ittamalla R, Deepak G (2021a) Analyzing the attitude of Indian citizens towards COVID-19 vaccine–a text analytics study. Diabetes Metab Syndr 15:595–599. https://doi.org/10.1016/j.dsx.2021.02.031

Praveen SV, Ittamalla R, Deepak G (2021b) Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19–a machine learning study of 840,000 tweets. Diabetes Metab Syndr 15:667–671. https://doi.org/10.1016/j.dsx.2021.03.016

Purwitasari D, Putra CBP, Raharjo AB (2023) A stance dataset with aspect-based sentiment information from Indonesian COVID-19 vaccination-related tweets. Data Brief 47:108951. https://doi.org/10.1016/j.dib.2023.108951

Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P (2023) Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination twitter dataset. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.118715

Rahmanti AR, Chien CH, Nursetyo AA, Husnayain A, Wiratama BS, Fuad A, Yang HC, Li YCJ (2022) Social media sentiment analysis to monitor the performance of vaccination coverage during the early phase of the national COVID-19 vaccine rollout. Comput Methods Programs Biomed. https://doi.org/10.1016/j.cmpb.2022.106838

Rani S, Jain A (2023) DFM: deep fusion model for COVID-19 vaccine sentiment analysis. pp 227–235

Saleh SN, McDonald SA, Basit MA, Kumar S, Arasaratnam RJ, Perl TM, Medford RJ (2021) Public perception of COVID-19 vaccines through analysis of twitter content and users. medRxiv. https://doi.org/10.1101/2021.04.19.21255701

Sv P, Tandon J, Vikas HH (2021) Indian citizen's perspective about side effects of COVID-19 vaccine–a machine learning study. Diabetes Metab Syndr: Clin Res Rev 15:102172. https://doi.org/10.1016/j.dsx.2021.06.009

Sv P, Lorenz JM, Ittamalla R, Dhama K, Chakraborty C, Kumar DVS, Mohan T (2022) Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding COVID-19 booster vaccine shots in India: crucial to expanding vaccination coverage. Vaccines 10:1929. https://doi.org/10.3390/vaccines10111929

Thorpe A, Fagerlin A, Drews FA, Shoemaker H, Scherer LD (2022) Self-reported health behaviors and risk perceptions following the COVID-19 vaccination rollout in the USA: an online survey study. Publ Health 208:68–71. https://doi.org/10.1016/j.puhe.2022.05.007

Turón A, Altuzarra A, Moreno-Jiménez JM, Navarro J (2023) Evolution of social mood in Spain throughout the COVID-19 vaccination process: a machine learning approach to tweets analysis. Publ Health 215:83–90. https://doi.org/10.1016/j.puhe.2022.12.003

Yousefinaghani S, Dara R, Mubareka S, Papadopoulos A, Sharif S (2021) An analysis of COVID-19 vaccine sentiments and opinions on twitter. Int J Infect Dis 108:256–262. https://doi.org/10.1016/j.ijid.2021.05.059