Social Network Analysis and Mining (2022) 12:57
https://doi.org/10.1007/s13278-022-00885-w

ORIGINAL ARTICLE

Public reactions towards Covid-19 vaccination through twitter before and after second wave in India

Siddhi Mishra[1] · Abhigya Verma[1] · Kavita Meena[1] · Rishabh Kaushal[1]

Received: 7 December 2021 / Revised: 7 March 2022 / Accepted: 3 May 2022

Abstract

Social media have a significant impact on opinion building in public. Vaccination in India started in January 2021. We have seen many opinions towards vaccination of the people, as vaccination is one of the most crucial steps toward the fight against COVID-19. In this paper, we have compared the public's sentiments towards COVID vaccination in India before the second wave and after the second wave. We worked by extracting tweets regarding vaccination in India, building our datasets. We extracted 5977 tweets before the second wave and 42,936 tweets after the second wave. We annotated the collected tweets into four categories, namely Provaccine, Antivaccine, Hesitant and Cognizant. We built a baseline model for sentiment analysis and have used multiple classification techniques among which Random Forest using the TF-IDF vectorization technique gave the best accuracy of 69% using max-features and n-estimators as parameters.

Keywords Sentiment classification · Machine learning · Covid-19 vaccincation

1 Introduction

Covid-19 has immensely affected all aspects of our lives, the daily schedule, mental and physical health, the environment, and financial stability (Marois et al. 2020; McKibbin et al. 2020). The pandemic has led to a massive loss of human lives worldwide. The economic and social disturbance caused by it is unimaginable (Woolf et al. 2020). Due to such huge losses and gloomy news of the impact of Covid-19, the mental health (Chatterjee et al. 2020; Khan et al. 2020; Xiong et al. 2020) of people have suffered immensely. Consequently, there is an environment of disappointment, distress, and disenchantment (Chaix et al. 2020; Holingue et al. 2020). So, when the vaccines to protect oneself from Covid-19 were launched, it was only natural for people to lack trust towards them. People suspected whether sufficient safeguards (Kostoff et al. 2020; Shimabukuro 2021) were put in place in the manufacture of vaccines and raised doubts over the efficacy of vaccines (Heath et al. 2021; Lipsitch and Dean 2020). There is a hesitancy and uncertainty amongst the public due to fear of side effects, misinformation, difficulty in registration/slot bookings, and many more (Murphy et al. 2021; Machingaidze and Wiysonge 2021; Dror et al. 2020).

In this work, we build upon the prior work on response and opinion of people towards Covid-19 vaccines from the lens of social media. Nowadays, online social media platforms like Twitter are pretty popular among users, and in the context of Covid-19 induced social distancing and lockdowns, a lot of users had no other option but to express their opinions and sentiments on social media (Goel and Gupta 2020; Bridgman et al. 2020; Cuello-Garcia et al. 2020). We focus on the vaccination drive in India, the largest in the world, which started on 16 January 2021, where 165,741 people were vaccinated on the first day (Bagcchi 2021). The vaccination drive began in the backdrop of the survey

✉ Rishabh Kaushal
rishabhkaushal@igdtuw.ac.in

Siddhi Mishra
siddhimishra499@gmail.com

Abhigya Verma
abhigya006btit19@igdtuw.ac.in

Kavita Meena
kavita23meena.2002@gmail.com

[1] Department of Information Technology, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, Delhi 110006, India

Published online: 31 May 2022

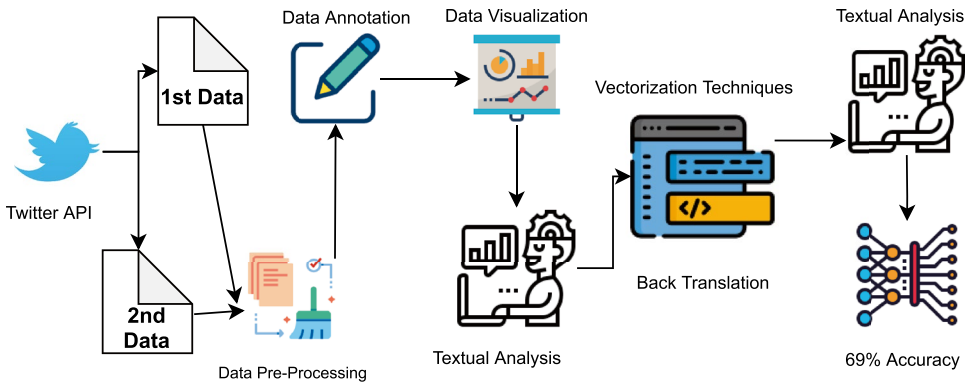


Fig. 1 Our workflow diagram involving data collection, annotation, visualization, text analysis, vectorization, data augmentation (back translation), and machine learning

(The Way Forward Survey[1]) conducted in December 2020 by GOQii with about 11,000 respondents to understand the openness of Indians to take the vaccine. It was found that about 53% of the Indian population is unsure about taking COVID-19 vaccination. Henceforth, massive promotions were done to encourage people to vaccinate themselves, and meticulous planning was done to handle the large scale (Foy et al. 2021; Thangaraj et al. 2021; Jeyanathan et al. 2020). We focus on understanding and characterizing people's views and opinions towards Covid-19 vaccination in India, which is useful in the following ways: (1) It would help authorities to analyse the people's hesitancy (inferred from their sentiment) towards vaccination drive accordingly so they can undertake publicity campaigns accordingly; (2) Also, investors, shareholders, and companies can use this analysis to determine which vaccine is more popular among people and is drawing positive sentiments; (3) People at large can understand the opinion of other users, thereby helping them formulate their opinions accordingly.

While there are several papers analysing healthcare (Chatterjee et al. 2020), recovery, and economic impact of coronavirus (McKibbin et al. 2020), very few have worked towards the opinion mining of people towards vaccination from social media posts. Therefore, we curated a novel dataset by extracting tweets from Twitter API in this work. We employed keywords (popular hashtags) such as #Covaxine, #Covishield, #bharatbiotech, #LargestVaccineDrive, #IndiaVaccine, and #bjpVaccine to retrieve tweets. Like elsewhere in the world, the second Covid-19 wave started to have its impact in India from the second week of April 2021 and peaked subsequently (Asrani et al. 2021; Ranjan et al. 2021).

In Fig. 1, we explain our proposed workflow starting with data collection, annotation, visualization, text analysis, vectorization, data augmentation (back translation), and machine learning. More specifically, we make the following novel contributions.

- We collected 5977 tweets before the second wave and 42,936 tweets after the start of the second wave based on the relevant hashtags related to Covid-19 Vaccination in India. We make our dataset[2] available for public use.
- To understand people's opinion towards vaccination, we did not adopt the conventional sentiment classes of positive, negative, or neutral. Instead, we performed manual annotation of tweets into four unique classes: ' Provaccine,' 'Antivaccine,' 'Hesitant,' and 'Cognizant.'
- Given that our goal is to perform multi-class classification, we train models to classify a given tweet into one of these four categories. We employed different vectorization techniques and back translation for data augmentation. It turned out that a random forest classifier using the TF-IDF vectorization technique gave the highest accuracy of 69% for four-class classification.

2 Literature review

In the field of medicine, the vaccination process (Kaur and Gupta 2020; Malik et al. 2020) is a well-established practice to create immunity for a disease. However, not much work has happened to study sentiments of Indian towards Covid-19 vaccines, mainly to compare people's opinions before and after the second wave. Therefore, this section has also explored prior works related to vaccine research and people's sentiment towards vaccines.

Toll and Li (2021) inspected the users' attitude towards Measles, Mumps, and Rubella (MMR) vaccination. The different MMR vaccine doses and basic viewpoints to vaccination were modelled one by one with multi-nomial logistic regression. Perspective towards vaccine was highly in sync with MMR vaccine results. Although this research paper

[1] Way Forward Survey conducted by GOQii in Dec 2020. https://www.healthwire.co/53-indian-unsure-about-taking-covid-19-vaccine-survey/.

[2] https://github.com/rishabhkaushal/covid19vaccinesentiment.

is not covering our use case, we took it as a reference for analysing the methodology to extract the attitude towards the vaccination. Radzikowski et al. (2016) examined the sentiments towards vaccination through the tweets that were posted on Twitter after the 2015 measles outburst. A collection of 669,136 tweets was made during the time of February 1 to March 9 of 2015. Our work also involves the collection of tweets; however, we focus on Covid-19 vaccine-related tweets. Dubey (2021) examined the emotions in tweets posted on Twitter in India regarding the two vaccines, namely Oxford-AstraZeneca's Covishield and Bharat Biotech's Covaxin. Tweets were classified into a sentiment class by applying the pre-trained NRC Emotion lexicon (Mohammad and Turney (2013)). For Covaxin, 69% were tweeting with positive sentiments while 31% had negative sentiments. For Covishield, 71% tweets had positive sentiments while 29% tweets had negative sentiments associated with it. In contrast, our work curates an annotated dataset and trains a classifier from scratch for classifying tweets into one of the four categories: Provaccine, Antivaccine, Hesitant, and Cognizant.

Samuel et al. (2020) studied the general outlook of the public towards the pandemic using tweets related to Covid-19 by using sentiment analysis packages of R statistical software.[3] Their methodology can be divided into two parts: data visualizations and text analysis. The second part deals with the classification of texts retrieved, i.e., tweets, into negative and positive classes. They found that the naïve bayes classifier had the best accuracy of 57%, whereas logistic regression has a precision of 52%.

Shaban-Nejad et al. (2015) provided a method, referred to as the Vaccine Attitude Surveillance using Semantic Analysis (VASSA) Framework, for layout and build-out of an integrated semantic platform to grasp the understanding of vaccine sentiments, reliance, and attitude using ontology. VASSA allows structuring the huge unstructured data around various blogs to ease the analysis of content and discover patterns of individual words or sentences in blogs. NLP module can be used to analyse the idea extracted by VASSA ontology. Monitoring and analysis of vaccines during real-time could help vaccination programs further lead to better and well-timed plans for any particular health concern. In recent times, various trials have been made to design a more efficient system to monitor the concerns regarding vaccines where refusal might lead to potential pandemics. Nemes and Kiss (2021) designed an emotion prediction model by co-relating words and further labelling the words to certain entries instead of the usual negative and positive classification. RNN is used along with NLP processing and Sentiment analysis for classic as well as Deep Learning styles. Most comparisons were made against TextBlob[4] which came up with decent results. There were also times when the neutral results were above 30% compared to their RNN model, which we were not useful to facilitate additional evaluations for the RNN model. Lyu et al. (2021) studied a human guided machine learning framework to establish state-of-the-art transformer language models to collect opinions towards Covid-19 vaccines, further classifying them into Pro, Anti, and Hesitant classes. Almost 40,000 meticulously picked tweets by 20,000 different users were selected. The subgroups were based on opinions as well as individual factors such as geographical location, financial conditions, religious beliefs, and political opinion. Researchers manually read and labelled the tweets into a particular class if it was relevant. A lower acceptance rate was found in the Southeast part of the US This research paper manually labels the data into pre-self-define categories vaccine-hesitant, pro-vaccine and anti-vaccine. XL-Net model is used for the training purpose (with labelled data) for three different types of hypotheses based on societal status and country.

In Table 1, we compare the prominent prior works related to our work. We briefly explain each work and provide our remarks. Recall that we move beyond the typical sentiment classification by classifying a tweet into four classes: Provaccine, Antivaccine, Hesitant, and Cognizant. Furthermore, we perform a comparative study of people's opinions and sentiments towards the Covid-19 vaccine before and after the second wave in India.

Next, we discuss some works related to understanding people's interests in general topics, opinions, and viewpoints. Comito (2021) studied a combination of peak detection and clustering method on a real-world tweet dataset related to Covid-19 in the US. They demonstrated that social media data could be used as a key predictor of epidemics. Consequently, data produced by consumer presence on digital platforms has become an asset for seizing and comprehension of pandemic outbreaks. Another significant study emphasizing the utilization of social media data Al-Dhuhli and Ismael (2013) used two primary methodologies, interviews and questionnaires, to assess the influence of social media on influencing the behaviour of people that desire to buy on the internet. Results of this study exhibit that buyers are primarily influenced by explanatory and layout variables, which ultimately lead people to decide what's best in selecting the most appropriate networking site to purchase one's desired items on the internet. Young (2018) discussed the risk and benefits of using social media data for health monitoring in order to improve public health and medical care. Schoen et al. (2013) introduced a prediction model using

[3] https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html.

[4] https://textblob.readthedocs.io/en/dev/.

Table 1 Comparative analysis between prior research works

Paper	Description	Our remarks
Toll and Li (2021)	The aim was to inspect the reasons for attitude towards Measles, Mumps, and Rubella (MMR) vaccination. The data used for the analysis were from The Longitudinal Study of Australian Children (LSAC). The impact of different MMR vaccines and attitudes towards vaccination were modelled using multinomial logistic regression. This longitudinal study began with a nationally representative sample of over 10,000 children and their families in 2004	Although this research paper is not covering our use case, we are taking it as a reference for analysing the methodology to extract the attitude towards the vaccination
Piedrahita-Valdés et al. (2021)	They studied hesitancy among users towards vaccines in general during the period 2011 to 2019 through the lens of social media. They collected 1,499,227 tweets regarding vaccines and performed sentiment analysis. Polarity analysis was performed using an association model based on a combination of lexical-based approaches and supervised machine learning methods. The results showed that 69.36% of the classified tweets are neutral, 21.78% were positive, and 8.86% were negative	We use this research paper to find the data preparation and extraction methodology
Dubey (2021)	They investigated the public's sentiment towards the vaccination drive of Covid-19 in India with respect to two vaccines, namely, Oxford-AstraZeneca's Covishield and Bharat Biotech's Covaxin. Tweets were classified into a sentiment class by applying the pre-trained NRC Emotion lexicon. They analysed tweets and performed sentiment classification. For Covaxin, 69% were tweeting with positive sentiments while 31% had negative sentiments. For Covishield, 71% tweets had positive sentiments while 29% tweets had negative sentiments associated with it	We also perform sentiment of users towards Indian vaccination drives. However, we build our novel dataset categorized into four classes: Provaccine, Antivaccine, Hesitant and Cognizant, which would be more useful for policy decisions
Radzikowski et al. (2016)	They studied the narratives related to the Measles vaccination drive on Twitter. Those tweets were analysed to identify key terms, connections among such terms, retweet patterns, the structure of the narrative, and connections to the geographical space. They found that the tweets made by news agencies had more effect on the opinion than the tweets by health organization accounts	Our work is also similar in terms of understanding people's opinions on social media. However, we go beyond the conventional sentiment classes and curate an annotated dataset of user attitudes towards Covid-19 vaccines in India
Samuel et al. (2020)	They inspected the general outlook of the public towards the pandemic using tweets related to Covid-19 by using a pre-trained sentiment analysis model. RNN is used along with NLP processing and Sentiment analysis for classic as well as Deep Learning styles. They covered four critical issues: (1)opinions related to the outbreak of Covid-19, (2) the use of tweets for sentiment analysis, (3) they observed that naive bayes are preferable for predicting sentiments with a precision of 91% and an accuracy of 57%	In contrast, we do not use a pre-trained model. Instead, we annotate and build a trained model from scratch using our annotated dataset
Shaban-Nejad et al. (2015)	They proposed a method for layout and build-out of an integrated semantic platform to grasp the understanding of vaccine sentiments, reliance, and attitude using ontologies	In our work, we approach the problem as the design of a machine learning model to automatically flag the user's attitude based on the tweet
Nemes and Kiss (2021)	They designed an emotion prediction model by co-relating words and further labelling the words to entries instead of the usual negative and positive classification. The deep learning model is used for sentiment analysis	We work on classifying a tweet into four categories, namely Provaccine, Antivaccine, Hesitant and Cognizant, moving beyond the typical sentiment classification

Table 2 Existing datasets related to vaccine

Dataset	Authors	Description
Barrier childhood immunization	Pearce et al. (2015)	The Longitudinal study of Australian children data was collected that contain a rich set of children and family information that are instrumental in controlling for the decision around vaccination
Twitter-measles	Radzikowski et al. (2016)	The GeoSocial Gauge system prototype was used to collect data from Twitter using a user-specified set of parameters such as keywords, locations, and time to understand users opinions on measles vaccines
Fear sentiment tweets	Samuel et al. (2020)	Over nine hundred thousand tweets were downloaded using a Twitter API by applying the keyword 'Corona', and the goal was to analyse fear sentiment related to Covid-19

social media, its advantage, and a particular scenario where these models are applicable. More research has been done to enhance further the results that we can infer from the social media data. For instance, Comito et al. (2019) worked on enhancing the existing clustering technique for topic detection using word embedding by using external corpora.

In general, there are many computational approaches to analyse the large information present on the Internet Zatari (2015). One of the most desired approaches, which does not require annotation, is the unsupervised classification Velusamy and Manavalan (2012) where unlabelled data is segregated based on the latent features in the data. An important example is the sentiment lexicon, a collection of words attached to it, which is widely used for sentiment analysis. Likewise, opinion summarization Kim and Kim (2014), is an important applications natural language processing, topic detection Snelson (2016); Liu and Zhang (2012); Comito et al. (2016) is a technique used to discover topics across text documents. The other approach is semi-supervised learning, where the algorithm is trained upon a combination of labelled and unlabelled data. This combination consists of a very small amount of labelled data and a very large amount of unlabelled data. This technique is widely used for speech analysis, Internet content classification, and many more applications.

However, a key contribution of our research is to create a novel dataset by extracting tweets related to COVID-19 Vaccination in India and annotating it into four classes. Further, we apply various machine learning algorithms to this data and compare their performance.

In Table 2, we mention a few datasets on studies related to vaccination. Pearce et al. (2015) collected data about fifteen factors causing barriers and facilitation towards immunization among 5107 children in Australia. They found that children who are not completely immunized need personalized interventions since the factors affecting lack of immunization are heterogeneous. They collected data from in-person questionnaires rather than collecting it from social media. Radzikowski et al. (2016) studied opinions of online users on Twitter towards vaccination drive in the aftermath of the Measles outbreak. They collected 669,136 tweets during a one-month period between 1 Feb to 9 March, 2015. They found that understanding the online discussions would help in formulating the policies for the betterment of people. Samuel et al. (2020) collected tweets related to Covid-19 to understand the fear sentiment among online users, particularly during the peak time of Covid-19 in the United States. As evident from these prior works, there is very little work on curating datasets related to Covid-19 vaccination. Our work fills this gap by proposing a dataset comprising of annotated tweets of online users' sentiment towards Covid-19 vaccination drives before and after the second wave in India. A lot of work has been done on vaccination and its public sentiment. But, in our work, we have focused on a novel data set that is India-specific. Also, there is hardly any work about COVID - 19, its vaccination, and the public sentiment related to it. So our work bridges this gap.

In the online world, knowledge about immunization is disseminated through social media platforms. In this paper, our goal is to explain the impact of Twitter, a prominent social media platform, on vaccination hesitancy and pro-vaccine sentiments. Online platforms exacerbate debates about immunization and have a long-term influence on the people, bringing narratives to combat vaccine reluctance. Therefore, our work endeavours to help policymakers understand the different viewpoints around the Covid-19 vaccination drives.

3 Methodology

We divide this section into two parts; first, we explain our steps to curate our dataset. Then, in the second part, we build classification models to predict users' sentiments towards vaccines.

3.1 Data set preparation

Our key data collection goal is to obtain tweets related to COVID-19 vaccines, and in this section we describe our collection methodology.

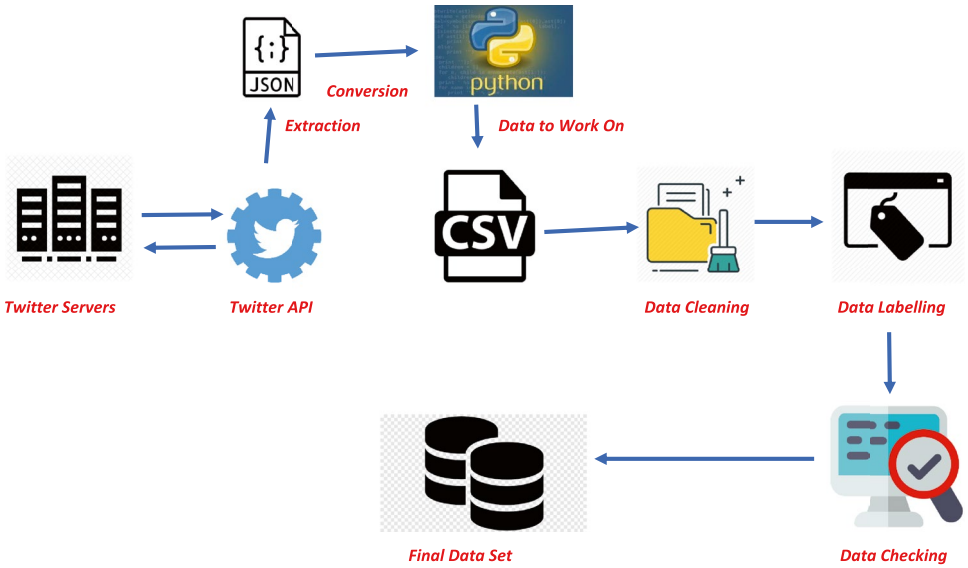


Fig. 2 Data Extraction

Table 3 Keywords used to extract data

Keyword	Reasons/explanation
Covaxine	Indian Vaccine Name
Covidshield	Indian Vaccine Name
Bharat Biotech	Covaxine manufacturer
Serum Institute of India	Covidshield manufacturer
Largest vaccine Drive	Most popular tweet
India vaccine	Getting tweet related to Indian Vaccine
Government India vaccine	Getting tweet related to Indian Vaccine
Modi/BJPVaccine	Getting tweets / trolls related to politics in vaccine

Initially, we started with signing up on Twitter and getting a developer's account[5] so that we could make API requests to extract data. We filled a developer application form and responded to the requirements as asked by Twitter. Once our request is approved, we built a project on a Twitter developer account. We generated Application Programmer Interface keys and access tokens to be used in the data collection process. We wrote a Python code to extract data for our analysis purpose. Figure 2 explains the entire process of data collection. We performed open authentication using the access keys and used Twitter Search API to collect tweets where a given keyword is present.

In Table 3, we enlist the keywords that we used to retrieve tweets related to Covid-19 vaccination. These keywords were decided based on prior research literature (Karafillakis et al. 2021; Tsao et al. 2021; Cinelli et al. 2020; Puri et al. 2020). The two most prominently used vaccines in India are Covaxine from Bharat Biotech and Covidshield from Serum Institute of India, so we searched for these keywords. Besides, the vaccination programme in India is the largest in the world, so keywords like *Largest Vaccine Drive* and *Government India Vaccine* were also used for search. Vaccination drive was also being popularized by the ruling political party and prime minister of India, so keywords like *Modi* and *BJP Vaccine* were also popular.

Given that we focus on vaccination campaigns in India, it was pertinent that we also keep an eye on Covid-19 cases in India. The vaccination campaign started on 16th January 2021 in India. The covid-19 infection cases were significantly low in February and March in 2021. However, subsequently, the second wave started with delta variants and peaked up in the latter half of April 2021 (Ranjan et al. 2021). Since our focus was to capture people's opinion on the Covid-19 vaccination, we collected vaccine-related tweets using the keywords mentioned in Table 3, both before and after the second wave hit India. We collected 5977 and 42,936 tweets before and after the second wave of Covid-19 in India, respectively. This was done to understand and compare the changes in people's opinions towards vaccinations.

[5] https://developer.twitter.com/en/portal.

Table 4 Labels and definitions: The labels were inspired from the study of public opinions on Covid-19 vaccines by Lyu et al. 2021

Category	Description
Pro-vaccine	1. Claiming that they would take the vaccine one it is available 2. Advocating and supporting vaccine/vaccine-associated entities like vaccine trials 3. Believing that vaccine will stop the pandemic 4. Encouraging other people to take vaccine
Anti-vaccine	1. Promoting information about vaccination which are not in support of it 2. Arguing with the facts which are in support of vaccination 3. Believing that an effective vaccine would not be invented soon 4. Believing that vaccine is dangerous
Vaccine-hesitant	1. Claiming that they would like to take the vaccine given that the vaccine is proven safe/effective 2. Asking queries related to COVID-19 vaccine 3. Showing worries about the vaccine's effectiveness
Vaccine-cognizant	1. Vaccine-related news and facts 2. Including vaccine and the commenters' opinions, but the focus is something else

Table 5 Data collection results

Dataset	Total tweets	Annotated tweets	Duration
Before Second Wave	5977	4094	March 2021
After Second Wave	42,936	5000	June 2021
Combined	48,913	9094	N/A

Table 6 Distribution of tweets before and after second wave of Covid-19 in India among four categories

Classes	Before	After
Vaccine-hesitant	545	849
Anti-vaccine	413	274
Pro-vaccine	1340	947
Vaccine-cognizant	1796	2930
Total	4094	5000

3.2 Data annotation

After pre-processing, the next most important step is the annotation of data. To that end, inspired from the study of public opinions on Covid-19 vaccines by Lyu et al. (2021), we classified the labels into four classes, namely, pro-vaccine, anti-vaccine, vaccine-hesitant and vaccine-cognizant. Below, we provide our definitions of these categories, also explained in Table 4:

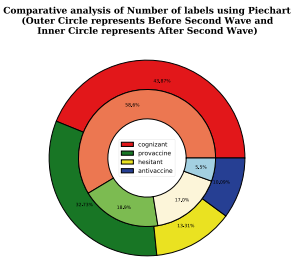
- Pro-vaccine: Tweets indicate people appreciating the vaccine and expressing either being willing to take the vaccine or already having gotten vaccinated.
- Anti-vaccine: Tweets that reveal users' annoyance for the vaccine and their negative attitude towards it.
- Vaccine-hesitant: Tweets which reflects hesitancy of people towards taking the vaccine.
- Vaccine-cognizant: Tweets which were informative regarding COVID-19 vaccines, helping people remain better informed.

In Table 5, we present the number of tweets that we collected before and after the second wave of Covid-19 in India. Further, in Table 6, we present the distribution of annotated tweets before and after the second wave of Covid-19 in India. It is apparent that after the second wave of Covid-19 in India, the number of vaccine-cognizant tweets increased by almost two times. The number of tweets expressing anti-vaccine sentiment reduced by half. This could be because after seeing the havoc caused by the second Covid-19 wave, people's opposition towards vaccination decreased. However, the pro-vaccine sentiment decreased by around 25%, and vaccine-hesitancy increased. These trends appear counter-intuitive. However, these trends could be because of the despair and hopelessness of the people who have seen the extremely adverse impact of the second Covid-19 wave in India.

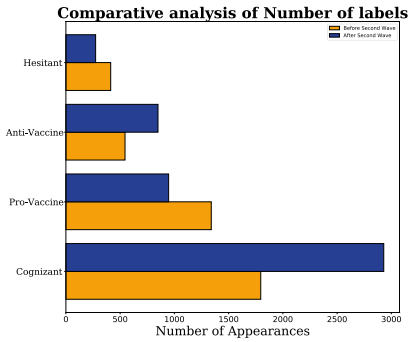
In respect to the data cleaning, we adopted standard approaches as follows: (1) We removed unwanted observations, including duplicate or irrelevant tweets and irrelevant symbols present in the data. (2) User mentions *@user*, where the username of the mentioned Twitter handle is used, were removed since user-user interactions are outside the scope of our work. (3) The resulting text was converted to lowercase letters. (4) Only selected stop words like apostrophes and abbreviations were replaced with spaces. (5) We also tokenized and stemmed the tweets as the last preprocessing step. After cleaning and preprocessing of data, the data was stored in the CSV file.

In Figs. 3a and 3b, we draw pie chart and bar graph of tweets collected before and after Covid-19 second wave. As

Fig. 3 Pie chart and bar chart of tweets before and after second Covid-19 wave

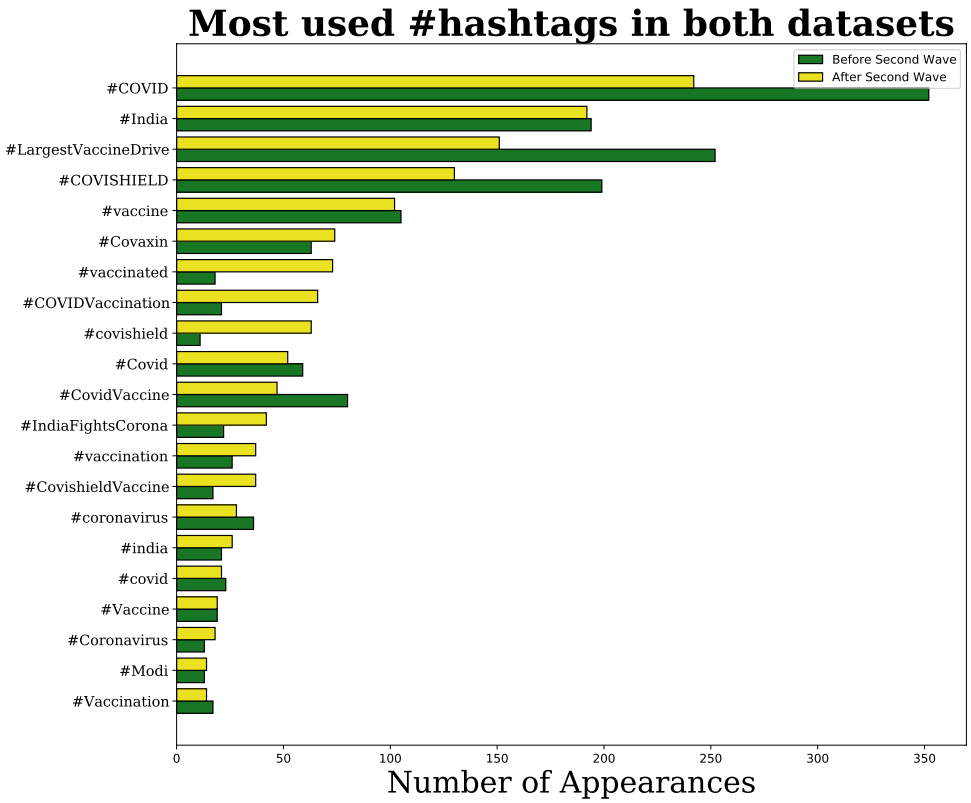


(a) Pie Chart: Pie Chart here represents data labels, the outer circle represents the number of tweets in each label Before second wave and inner circle representing After second wave.



(b) Bar Graph: In the Bar Graph four different bars are used to show the proportional measure of different classes in Before second wave and After second wave.

Fig. 4 Most used hashtags circulated (i) before second Covid-19 wave and (ii) after second Covid-19 wave



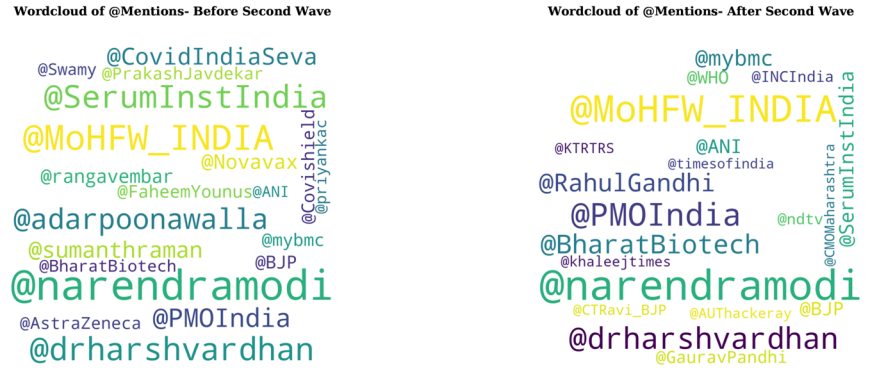
evident from these figures, the tweets belonging to *cognizant* class increased considerably after the second wave of Covid-19, and *hesitancy* decreased.

3.3 Data analysis

Having collected and annotated vaccine-related tweets before and after the second wave hit India, we perform basic data analysis in this section. More specifically, we ask the following research questions.

RQ1: What are the key changes in terms of hashtags frequently circulated before and after the second Covid-19 wave in India?

In Fig. 4, we compare the frequency of popular hashtags before and after the second Covid-19 wave. Before the second Covid-19 wave, the most common hashtags were around the general awareness themes of *#COVID*,



(a) Word cloud of Mentions Before Second Wave

(b) Word cloud of Mentions After Second Wave

Fig. 5 Word cloud and frequency of mentions before and after second Covid-19 wave

#India, and *#VaccineDrive*, and one of the vaccine name *#COVISHIELD* was most frequently occurring. However, after the second Covid-19 wave, vaccine-related hashtags like *#CovidVaccine*, *#vaccinated*, and *#vaccination* started to appear more often.

RQ2: Who are the people citizens looked out for help (by mentioning them) before and after the second Covid-19 wave in India?

We aim to find out people (users) mentioned by users before and after the Covid-19 wave in India. As depicted in Figs. 5a and 5b, there are some common handles like *@narendramodi* (Prime Minister), *@drharshvardhan* (Health Minister) and *@MoHFW_INDIA* (Ministry of Health and Family Welfare) that are mentioned in same proportion. Interestingly, among the two vaccines available in India at that time, *SerumInstIndia* who is the manufacturer of *Covidshield* was more mentioned in before Covid-19 second wave. Subsequently, *@BharatBiotech*, the manufacturer of the *Covaxine* also appeared frequently after Covid-19 second wave.

4 Baseline evaluation

This section presents the results of the baseline machine learning model evaluations performed on the classification of tweets. In the context of our problem, the goal is to classify a given tweet into one of the four classes, namely pro-vaccine, anti-vaccine, vaccine-hesitant, and vaccine-cognizant. Recall that we collected tweets before and after the second wave of Covid-19 in India. Therefore, we have the following four kinds of datasets:

1. Before Second Wave: Set of annotated tweets collected before the start of the second wave of Covid-19 in India.
2. After Second Wave: Set of annotated tweets collected after the start of the second wave of Covid-19 in India.
3. Merged: Set of combined annotated tweets of both before and after the second wave of Covid-19 in India.
4. Merged Back Translated: In order to increase the dataset, we performed a standard technique of *back translation* (Sennrich et al. 2015; Edunov et al. 2018) in which we converted the English tweets into French and then converted the french translation back to English.

We performed two types of text vectorization methods on these four datasets, namely, TF-IDF (Term Frequency - Inverse Document Frequency) and BoW (Bag of Words). We experimented with, evaluated, and compared the results of machine learning models on all these four datasets (vectorized using TF-IDF and BoW). We experimented using the following five machine learning algorithms: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbours (KN), and Gradient Boosting (GB). In the sub-sections below, we describe text vectorization and machine learning-based classifiers used.

4.1 Impact of back translation

To improve the efficiency of machine learning algorithms, we performed data augmentation by using the technique of back translation as proposed by Sennrich et al. (2015). Back translation is the process of re-translating content from the target language back to the source language. This work translated the merged dataset comprising annotated tweets before and after the second wave of Covid-19 in India into the French language. Subsequently, we again translated it back to the original English language. In Fig. 6, we depict the accuracy of different machine learning models with and without back translation. We observed that back-translation (shown in orange bars) always increases accuracy in all

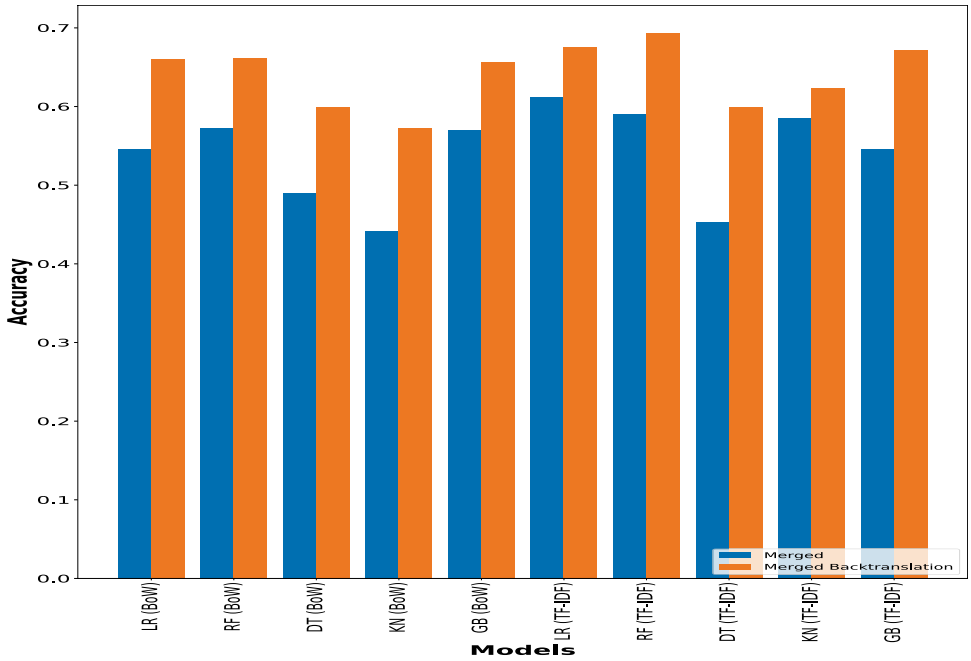
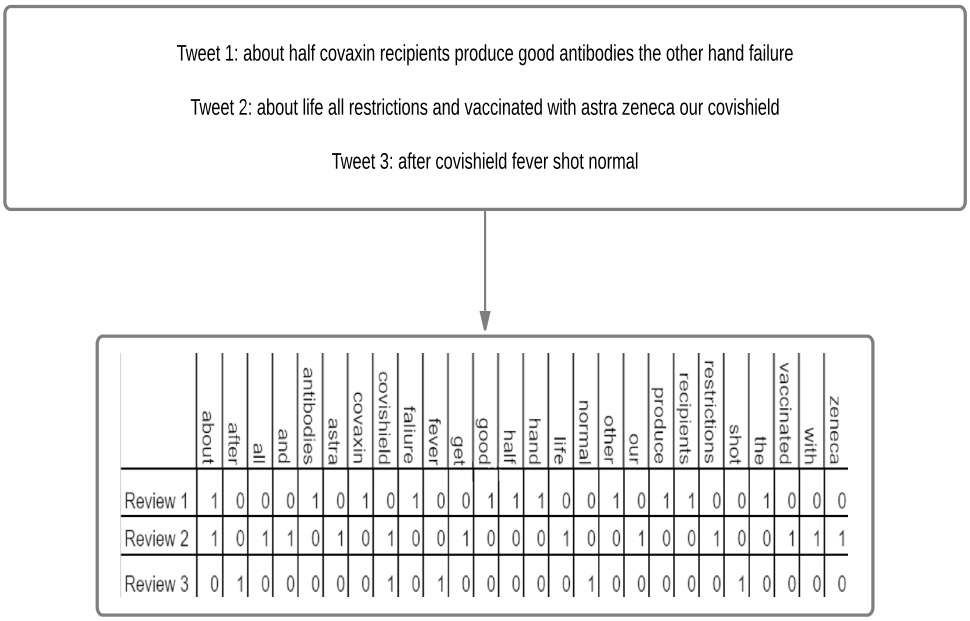


Fig. 6 Effect of backtranslation



	about	after	all	and	antibodies	astra	covaxin	covishield	failure	fever	get	good	half	hand	life	normal	other	our	produce	recipients	restrictions	shot	the	vaccinated	with	zeneca
Review 1	1	0	0	0	1	0	1	0	1	0	0	1	1	1	0	0	1	0	1	1	0	0	1	0	0	0
Review 2	1	0	1	1	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	1	1
Review 3	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0

Fig. 7 Illustration of BoW vectorisation technique

machine learning algorithms by 10%–20%. Without back translation, logistic regression with TF-IDF gave the best accuracy of around 60.4% which increased to 68.3% with random forest classifier when back translation was used with TF-IDF.

4.2 Comparison of vectorization methods

For the baseline evaluation, we leverage two popular baseline vectorization methods, namely, Bag of Words (BoW) (Zhang et al. 2010) and Term Frequency & Inverse Document Frequency (TF-IDF) (Aizawa 2003).

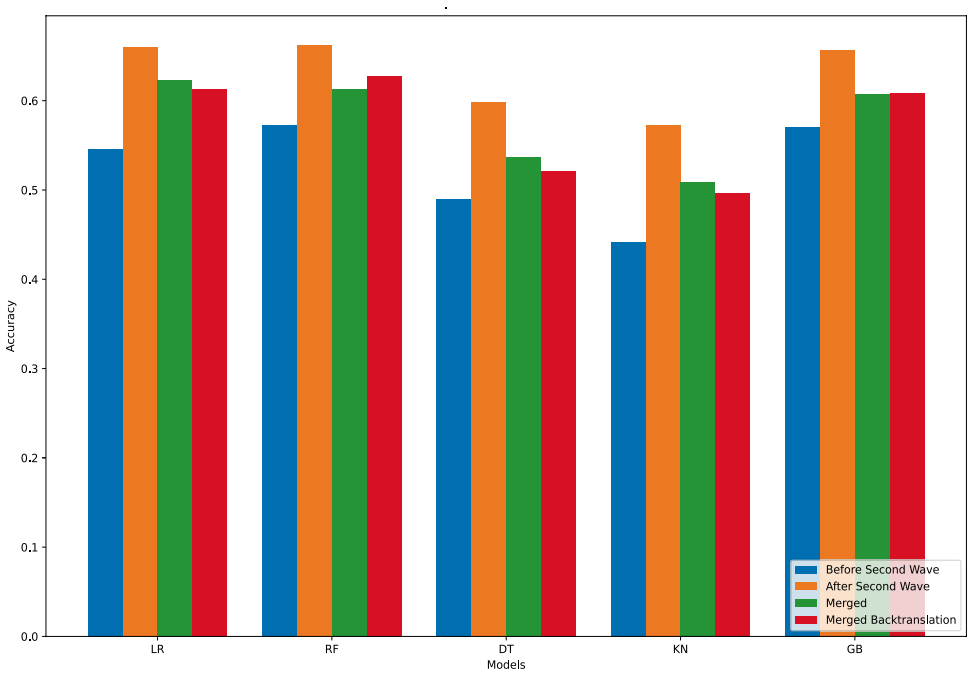
Bag of words (BoW) BoW model takes into account the occurrence of words within the document. In our case, we process tweet text which are written in natural language by users. First, we pre-process our data to convert text to lower case, remove all non-word characters and punctuation. Then we obtain the most frequent words in our text by creating a vocabulary to hold our bag of words. We tokenize each sentence into words; then, we check if the word exists in our vocabulary. To build the bag of words model, we construct

Fig. 8 TF-IDF vectorisation technique

Tweet 1: about half covaxin recipients produce good antibodies the other hand failure
Tweet 2: about life all restrictions and vaccinated with astra zeneca our covishield
Tweet 3: after covishield fever shot normal

- Vocabulary:
  'about', 'half', 'covaxin', 'recipients', 'produce', 'good', 'antibodies', 'the', 'other', 'hand', 'failure', 'life', 'all', 'restrictions ', 'and', 'get', 'vaccinated', 'with', 'astra', 'zeneca', 'our', 'covishield', 'after', 'fever', 'shot', 'normal'
- No of Words in Tweet 1 = 11
- TF for word 'about'
  = (No of times word 'about' appears in Tweet 1)/(No of terms in Tweet 1)
  = 1/ 11
- IDF('about')
  = log(No of documents / No of documents containing the word 'about')
  = 2 / 3
- $(TF\text{-}IDF)_{t,d} = TF_{t,d} * IDF_t$
  = (1/11) * (2/3)
  = 0.0606...

Fig. 9 Comparision of vectorization techniques on all four datasets



a vector to reveal whether a word is a frequent word or not. In Fig. 7, we display bag of words vectorisation technique on three tweets from after second wave dataset.

Term frequency & inverse document frequency (TF-IDF) TF-IDF model works on the intuition that is frequently occurring words are given less importance than rarely occurring words. In other words, the TF-IFDF value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. TF specifies the frequency of a word in the document. IDF is an evaluation of whether a word is rare or frequent across the documents in the entire corpus. A term acquires a high TF-IDF score with a high frequency in a document and low document frequency in the corpus. In Fig. 8, we illustrate TF-IDF vectorization technique on three tweets from after the second wave dataset.

In Fig. 9, we plot accuracy values for five machine learning algorithms, namely logistic regression (LR), random forest (RF), decision tree (DT), K-nearest neighbor (KNN), and gradient boosting (GB). The experiments used all four datasets before the second wave, after the second wave, combined, and merged with back-translation. As evident, all algorithms performed best with annotated tweets after the second wave of Covid-19 in India. This suggests that people's opinion and sentiments can be best predicted after the second wave. It indicates that people are expressing themselves more clearly and definitely after the second wave. On the other hand, the worst accuracy is with the annotated

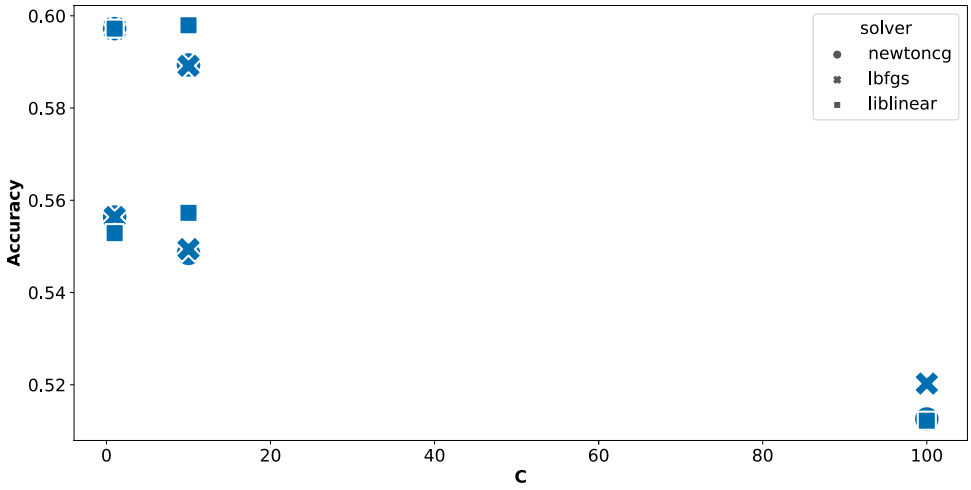


Fig. 10 Results of Hyperparameter tuning in Logistic Regression

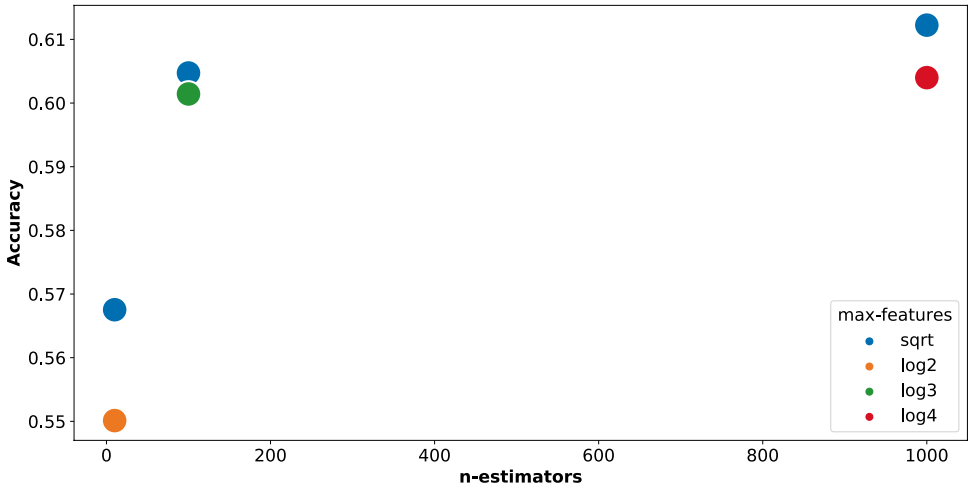


Fig. 11 Results of Hyperparameter tuning in Random Forest

tweets before the second wave of Covid-19, which indicates a lack of clarity in people's expression.

4.3 Model construction

In this sub-section, we discuss the different machine learning algorithms used for model construction. For each model, we explain the hyper-tuning process adopted for performance improvement.

Logistic regression (LR) Logistic regression is a classification model. It uses an independent variable to predict the dependent variable. It is similar to linear regression but here the dependent variable is categorical. We used the following three parameters for hyper-parameter tuning, namely, *solver* [newton-cg, lbfgs, liblinear, sag, saga]; *penalty* [none, l1, l2, elasticnet]; and *C* in [100, 10, 1.0, 0.1, 0.01]. The C parameter controls the penalty strength; increasing the regularization strength creates simple models that underfit the data and vice versa. From Fig. 10, we observe that best accuracy of 0.59 is obtained at hyper-parameter values of C: 1.0, penalty: l2, solver: liblinear.

Random forest (RF) A supervised learning algorithm, random forest, is used in regression as well as classification problems. It works by building various decision trees and merging them to get the best accuracy. More trees mean more experimenting and better results, and it will also prevent over-fitting. For hyper-parameter tuning, we focus on two parameters, namely, *n-estimators* [10, 100, 1000] and *max-features* ['sqrt', 'log2']. From Fig. 11, we observe that best accuracy of 0.61 is obtained using max-features: sqrt, n-estimators: 1000.

Decision tree (DT) A Decision Tree classifier is a supervised learning algorithm that is used in classification tasks. Since we are categorizing the tweets into four classes, this is a good algorithm for our task. In the decision tree, every column containing an attribute corresponds to a node in a tree. The root node is the essential attribute, and the leaf node is the class prediction. The algorithm goes through various nodes of the tree to predict the right class. For hyper-parameter tuning, we focus on two most important parameters, namely, *criterion* and *max-depth*. From Fig. 12, we observe that maximum

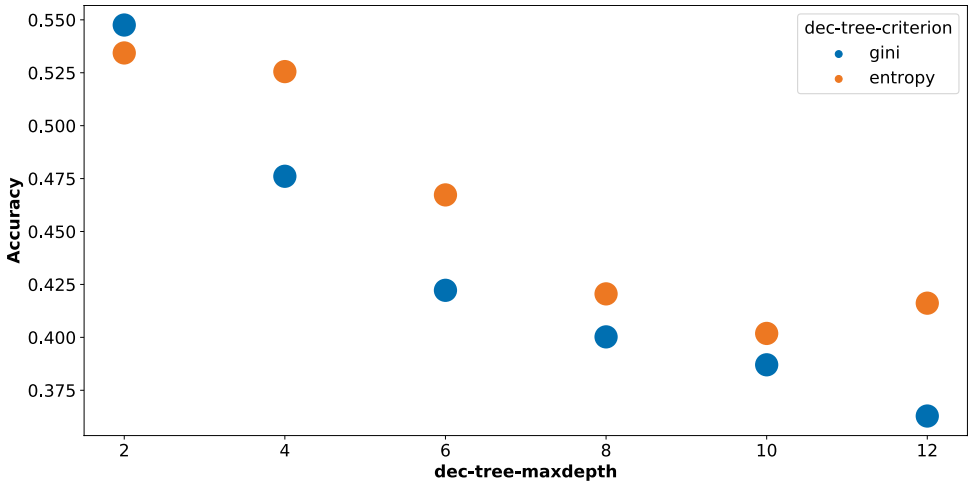


Fig. 12 Results of Hyperparameter tuning in Decision Tree Classifier

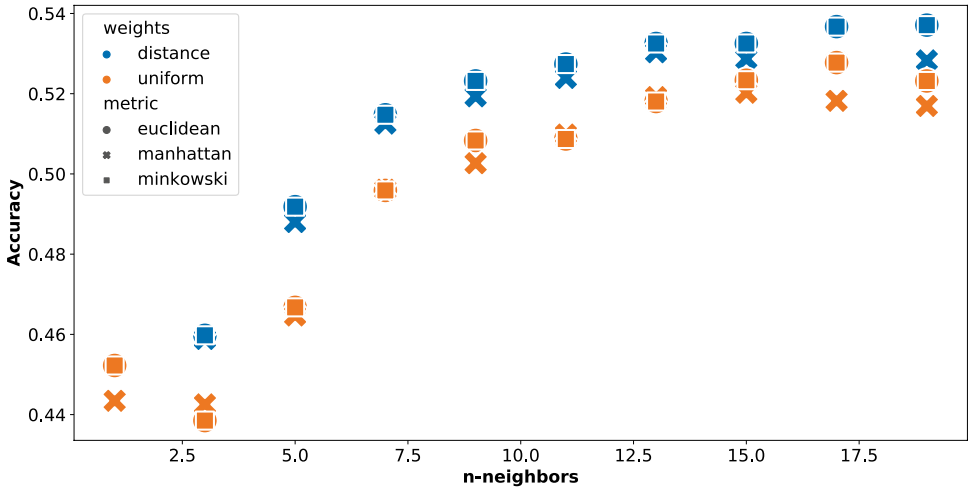


Fig. 13 Results of Hyperparameter tuning in KNN

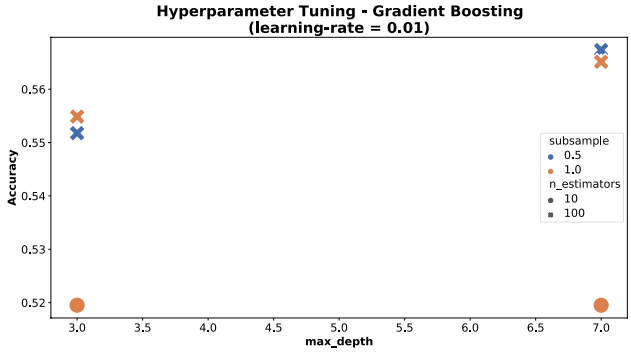
accuracy of 0.55 is obtained using 'gini' as criterion and max-depth of 2.

K-nearest neighbours (KNN) K-Nearest Neighbours algorithm is a classification algorithm. It is nonparametric; in other words, it does not make a presumption on the primary data used. K-nearest neighbours of tweets are fetched before the vaccine data set could be classified. Significant voting of data set in the neighbourhood is used to decide the classification of a tweet. To apply the algorithm, we need to choose a suitable value of k as the outcome is dependent on it.

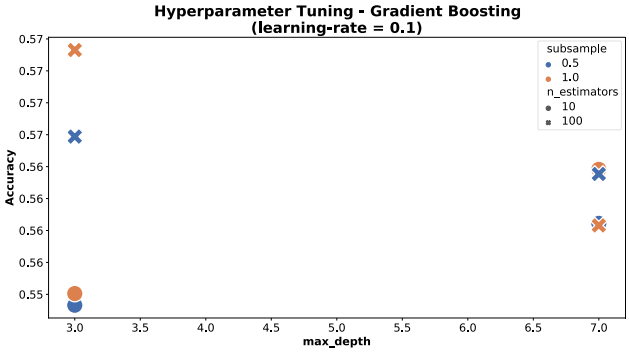
For hyper-parameter tuning, we used the following parameters: *n-neighbors* to determine the best k based on the values we have computed before; *weights* to check whether adding weights is beneficial to the model or not. 'uniform' assigns no weight, and 'distance' weighs points by the inverse of their distances (nearer points have more weight than the farther points); and *metric* which is the distance metric to be used will calculating the similarity. From Fig. 13, it is evident that we got best accuracy of 0.54 using metric: 'euclidean', n-neighbors: 19, and weights: 'distance'.

Gradient boosting Gradient boosting is an algorithm that improves accuracy and reduces over-fitting. Starting with fitting an initial model, it moves to a second model, focusing on accurate predictions at places where the first model went wrong. These two models combined give better results. Target outcomes for each case are set based on the error gradient concerning the prediction. Each model works on minimizing errors of previous models. For hyper-parameter tuning, we used the following parameters, namely, *n-estimators* [10, 100], *learning-rate* [0.01, 0.1], *subsample* [0.5, 1.0] and *max-depth* [3, 7]. From Figure 14, we find that the best accuracy of 0.57 is obtained using learning-rate: 0.1, subsample: 1.0, n-estimators: 100, and max-depth: 3.

In Table 7, we compare the accuracy of different machine learning algorithms and vectorization techniques. A random forest classifier obtained the best average accuracy considering all datasets using the TF-IDF vectorization technique. The accuracy of all machine learning algorithms when



(a) Learning rate = 0.01



(b) Learning rate = 0.1

Fig. 14 Result of Hyperparameter tuning in Gradient Boosting

Table 7 Accuracy of machine learning algorithms using vectorisation techniques with different datasets (D1: Before Second Wave, D2: After Second Wave, D3: Merged, and D4: Merged Back Translation

Models used	D1	D2	D3	D4	Avg
BOW + LR	0.54	0.66	0.62	0.61	0.61
BOW + DT	0.48	0.59	0.53	0.52	0.53
BOW + KNN	0.44	0.57	0.50	0.49	0.50
BOW + RF	0.57	0.66	0.61	0.62	0.61
BOW + GB	0.57	0.65	0.60	0.60	0.61
TF-IDF + LR	**0.61**	0.67	**0.65**	0.63	0.64
TF-IDF + DT	0.45	0.60	0.54	0.54	0.53
TF-IDF + KNN	0.58	0.62	0.561	0.55	0.58
TF-IDF + RF	0.59	**0.69**	0.64	0.64	**0.64**
TF-IDF + GB	0.54	0.67	0.64	**0.65**	0.62

Bold values are the best performing results

Table 8 Results: accuracy of Best Models with different datasets using TF-IDF vectorization technique

Datasets	Best model	Accuracy
Before second wave	Logistic regression	61%
After second wave	Random forest	69%
Merged	Logistic regression	65%
Merged Backtranslation	Gradient boosting	65%

using dataset D2 (annotated tweets after the second wave of Covid-19 in India) is the highest. This shows that people's opinion and sentiments after the second wave of Covid-19 were more predictable with better clarity.

Lastly, in Table 8, we present the accuracy values of the machine learning models that performed the best for that dataset.

To summarize, among all algorithms, the best is random forest using the TF-IDF vectorization technique. The average precision is 0.53, the average recall is 0.45, and the average F1-score is 0.47. This method yields the highest accuracy of 69%.

5 Conclusion and future scope

We address the issue of sentiment analysis of public reaction towards the Covid-19 vaccination drive in India. In the present study, we have built and annotated a novel dataset of tweets categorized into four classes, namely Provaccine, Antivaccine, Hesitant and Cognizant. Furthermore, we captured tweets before the second wave and after the second wave of Covid-19 in India, which helped us understand the variation in the perception of vaccination. In the process of opinion mining and textual analysis, we found that most of the tweets before the second wave (43.8%) and after the second wave (58.6%) belong to the *cognizant* class which shows tweets mostly spreading information toward the vaccination process. Moreover, people have a more positive opinion than those against vaccination, but a significant group of people shows hesitancy toward vaccination. By performing baseline evaluation, we conclude that random forest using TF-IDF vectorization technique gave the best accuracy of 69%. Our dataset can be used for extending research in the field of sentiment analysis for vaccination. This work to analyse the sentiment of different vaccines can be further extended to measure the satisfactory and popularity levels of different vaccines available. It would help the government to choose and order the popular vaccine and wade away the vaccine hesitancy. This work can be used by government authorities to determine and predict the current sentiment of people toward vaccination and accordingly create campaigns to address the concerns and apprehensions of people. In this work, we have only focused on tweets from India, which can be extended to study and analyse the sentiment globally for a

different country that has a significant amount of hesitancy toward vaccination in public.

Acknowledgements We acknowledge the efforts of Vanshika Bagri who made valuable contributions to the work.

Declaration

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aizawa A (2003) An information-theoretic perspective of tf-idf measures. Inf Process Manag 39(1):45–65

Al-Dhuhli I, Ismael S (2013) The impact of social media on consumer buying behaviour. Unpublished Master's Project, Sultan Qaboos University

Asrani P, Eapen MS, Hassan MI, Sohal SS (2021) Implications of the second wave of Covid-19 in India. Lancet Respir Med 9(9):93–94

Bagcchi S (2021) The world's largest Covid-19 vaccination campaign. Lancet Infect Dis 21(3):323

Bridgman A, Merkley E, Loewen PJ, Owen T, Ruths D, Teichmann L, Zhilin O (2020) The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. Harvard Kennedy School Misinformation Review 1(3)

Chaix B, Delamon G, Guillemassé A, Brouard B, Bibault JE (2020) Psychological distress during the covid-19 pandemic in france: a national assessment of at-risk populations. General Psychiatry 33(6)

Chatterjee K, Chatterjee K, Kumar A, Shankar S (2020) Healthcare impact of Covid-19 epidemic in India: A stochastic mathematical model. Medical J Armed Forces India 76(2):147–155

Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. Sci Rep 10(1):1–10

Comito C (2021) How covid-19 information spread in us the role of twitter as early indicator of epidemics. IEEE Trans Serv Comput 1(1):1-1

Comito C, Forestiero A, Pizzuti C (2019) Word embedding based clustering to detect topics in social media. In: 2019 IEEE/WIC/ACM international conference on web intelligence (WI), pp. 192–199. IEEE

Comito C, Pizzuti C, Procopio N (2016) Online clustering for topic detection in social data streams. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI), pp. 362–369 . IEEE

Cuello-Garcia C, Pérez-Gaxiola G, van Amelsvoort L (2020) Social media can have an impact on how we manage and investigate the Covid-19 pandemic. J Clin Epidemiol 127:198–201

Dror AA, Eisenbach N, Taiber S, Morozov NG, Mizrachi M, Zigron A, Srouji S, Sela E (2020) Vaccine hesitancy: the next challenge in the fight against Covid-19. Eur J Epidemiol 35(8):775–779

Dubey AD (2021) Public sentiment analysis of covid-19 vaccination drive in india. Available at SSRN 3772401

Edunov S, Ott M, Auli M, Grangier D (2018) Understanding back-translation at scale. arXiv preprint arXiv:1808.09381

Foy BH, Wahl B, Mehta K, Shet A, Menon GI, Britto C (2021) Comparing Covid-19 vaccine allocation strategies in India: A mathematical modelling study. Int J Infect Dis 103:431–438

Goel A, Gupta L (2020) Social media in the times of Covid-19. J Clinical Rheumatol 26(6): 220-223

Heath PT, Galiza EP, Baxter DN, Boffito M, Browne D, Burns F, Chadwick DR, Clark R, Cosgrove C, Galloway J et al (2021) Safety and efficacy of nvx-cov2373 Covid-19 vaccine. N Engl J Med 385(13):1172–1183

Holingue C, Kalb LG, Riehm KE, Bennett D, Kapteyn A, Veldhuis CB, Johnson RM, Fallin MD, Kreuter F, Stuart EA et al (2020) Mental distress in the united states at the beginning of the Covid-19 pandemic. Am J Public Health 110(11):1628–1634

Jeyanathan M, Afkhami S, Smaill F, Miller MS, Lichty BD, Xing Z (2020) Immunological considerations for Covid-19 vaccine strategies. Nat Rev Immunol 20(10):615–632

Karafillakis E, Martin S, Simas C, Olsson K, Takacs J, Dada S, Larson HJ (2021) Methods for social media monitoring related to vaccination: systematic scoping review. JMIR Public Health Surveill 7(2):17149

Kaur SP, Gupta V (2020) Covid-19 vaccine: A comprehensive status report. Virus Res, 198114

Khan KS, Mamun MA, Griffiths MD, Ullah I (2020) The mental health impact of the covid-19 pandemic across different cohorts. Int J Mental Health Addict, 1–7

Kim S, Kim N (2014) A study on the effect of using sentiment lexicon in opinion classification. J Intell Inf Syst 20(1):133–148

Kostoff RN, Briggs MB, Porter AL, Spandidos DA, Tsatsakis A (2020) [comment] Covid-19 vaccine safety. Int J Mol Med 46(5):1599–1602

Lipsitch M, Dean NE (2020) Understanding Covid-19 vaccine efficacy. Science 370(6518):763–765

Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis,415–463 (2012)

Lyu H, Wang J, Wu W, Duong V, Zhang X, Dye TD, Luo J (2021) Social media study of public opinions on potential covid-19 vaccines: informing dissent, disparities, and dissemination. Intell Med 2(1):1-12

Machingaidze S, Wiysonge CS (2021) Understanding Covid-19 vaccine hesitancy. Nat Med 27(8):1338–1339

Malik AA, McFadden SM, Elharake J, Omer SB (2020) Determinants of Covid-19 vaccine acceptance in the us. EClinicalMedicine 26:100495

Marois G, Muttarak R, Scherbov S (2020) Assessing the potential impact of covid-19 on life expectancy. PLoS ONE 15(9):0238678

McKibbin W, Fernando R, et al. (2020) The economic impact of covid-19. Econ Time of COVID-19 45(10.1162)

Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465

Murphy J, Vallières F, Bentall RP, Shevlin M, McBride O, Hartman TK, McKay R, Bennett K, Mason L, Gibson-Miller J et al (2021) Psychological characteristics associated with Covid-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. Nat Commun 12(1):1–15

Nemes L, Kiss A (2021) Social media sentiment analysis based on covid-19. J Inform Telecommun 5(1):1–15

Pearce A, Marshall H, Bedford H, Lynch J (2015) Barriers to childhood immunisation: Findings from the longitudinal study of australian children. Vaccine 33(29):3377–3383

Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, Guillem-Saiz P, Bermejo-Higuera JR, Guillem-Saiz J, Sicilia-Montalvo JA, Machío-Regidor F (2021) Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019. Vaccines 9(1):28

Puri N, Coomes EA, Haghbayan H, Gunaratne K (2020) Social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases. Human Vaccines Immunotherapeutics 16(11):2586–2593

Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL (2016) The measles vaccination narrative in twitter: a quantitative analysis. JMIR Public Health Surveill 2(1):5059

Ranjan R, Sharma A, Verma MK (2021) Characterization of the second wave of Covid-19 in India. medRxiv

Samuel J, Ali G, Rahman M, Esawi E, Samuel Y et al (2020) Covid-19 public sentiment insights and machine learning for tweets classification. Information 11(6):314

Schoen H, Gayo-Avello D, Metaxas PT, Mustafaraj E, Strohmaier M, Gloor P (2013) The power of prediction with social media. Internet Research 23(5):528-543

Sennrich R, Haddow B, Birch A (2015) Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709

Shaban-Nejad A, Menon S, Buckeridge D (2015) A semantic web platform for online vaccine sentiment surveillance. Online J Public Health Inform 7(1)

Snelson CL (2016) Qualitative and mixed methods social media research: A review of the literature. Int J Qual Methods 15(1):1609406915624574

Thangaraj JWV, Yadav P, Kumar CG, Shete A, Nyayanit DA, Rani DS, Kumar A, Kumar MS, Sabarinathan R, Kumar VS, et al. (2021) Predominance of delta variant among the Covid-19 vaccinated and unvaccinated individuals, India, may 2021. J Infection 84(1):94-118

Toll M, Li A (2021) Vaccine sentiments and under-vaccination: Attitudes and behaviour around measles, mumps, and rubella vaccine (mmr) in an australian cohort. Vaccine 39(4):751–759

Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA (2021) What social media told us in the time of covid-19: a scoping review. The Lancet Digital Health

Velusamy K, Manavalan R (2012) Performance analysis of unsupervised classification based on optimization. Int J Comput Appl 975:8887

Woolf SH, Chapman DA, Sabo RT, Weinberger DM, Hill L (2020) Excess deaths from Covid-19 and other causes, march-april 2020. JAMA 324(5):510–513

Xiong J, Lipsitz O, Nasri F, Lui LM, Gill H, Phan L, Chen-Li D, Iacobucci M, Ho R, Majeed A, et al. (2020) Impact of covid-19 pandemic on mental health in the general population: A systematic review. J Affect Disorders 227(1):55-64

Young SD (2018) Social media as a new vital sign: commentary. J Med Internet Res 20(4):8563

Zatari T (2015) Data mining in social media. Int J Sci Eng Res 6(7):152–154

Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1(1–4):43–52