An analytical insight of discussions and sentiments of Indians on Omicron-driven third wave of COVID-19 using twitter data

Deepika Vatsa Deepika Vatsa (vatsa.deepika.email@gmail.com)
Bennett University
Ashima Yadav Ashima Yadav
Bennett University

Research Article

Keywords: Coronavirus, COVID-19, India, Pandemic, Sentiment analysis, Topic modeling, Twitter

Posted Date: April 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1508291/v2

An analytical insight of discussions and sentiments of Indians on Omicron-driven third wave of COVID-19 using twitter data

Deepika Vatsa[a,*], Ashima Yadav[b]

[a]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, 201310, Uttar Pradesh, India
[b]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, 201310, Uttar Pradesh, India

ARTICLE INFO

Keywords:
Coronavirus
COVID-19
India
Pandemic
Sentiment analysis
Topic modeling
Twitter

ABSTRACT

Microblogging has become one of the most crucial tool for expressing and sharing the opinions and views of everyday life events. Digital channels are being used to monitor public health issues on the Internet. Twitter is a very popular source that provides tweets related to the sentiment of the public during the COVID-19 pandemic. Many researchers have used tweets to monitor the opinion of the people towards the coronavirus vaccine, mental health problems, treatment received by the doctors, impact of lockdown, etc. However, these works were mostly limited to the first and second waves of the pandemic. In this work, we aim to study the impact of the third wave of the pandemic, which started in December 2021 in India. We accomplished this by collecting tweet data set of two months, i.e., December 2021 and January 2022, discussing COVID-19 and having country code as “IN". We employed the Latent Dirichlet Allocation (LDA) technique for topic modeling and labeled each tweet message with the topic words that best describe it. We also utilized sentiment labels for each tweet and analyzed the distribution of different topics across different sentiment labels. This helped us to analyze the perspectives and sentiments of the people with respect to different topic discussions. Our analysis discovered that the two most discussed topics were “precautionary measures" like get well soon, stay safe, wear mask, etc., and “vaccine" where people have discussed about its effectiveness and vaccination drive in India. We found that people mostly had neutral sentiments for the former topic while for the latter, overall sentiment polarity was negative, reflecting peoples’ mistrust in the COVID-19 vaccine.

1. Introduction

Twitter has become one of the popular sources for gathering public opinion on health research. According to the 2019 survey results [1], there were 290.5 million monthly users actively using Twitter, and this count will increase to 340 million by 2024. Hence, Twitter can get real-time opinions and attitudes about people located in different parts of the world. The outbreak of COVID-19 has become a cause of concern for policymakers and scientists. In March 2020, the World Health Organization (WHO) declared COVID-19 as a pandemic [2]. Since then, the dreaded disease has caused a devastating effect on the entire world, resulting in more than 5,878,328 deaths worldwide [3].

Previous studies have applied sentiment analysis during different outbreaks and epidemics. Baker et al. [4] utilized machine learning-based techniques to study the spread of influenza based on Arabic tweets. The authors studied and conducted experiments using several machine learning-based methods like Support Vector Machine, Decision Trees, Naïve Bayes, and K-Nearest Neighbor to analyze around 54,065 influenza-related tweets in Arabic. Culotta et al. [5] detected N1H1 influenza-related tweets and compared the results with the Centers for Disease Control and Prevention by applying different classification methods. Experimental results show that the multiple linear regression model achieved the highest accuracy of 84.3%. Some studies have developed models which collected data for multiple infectious diseases like measles, Ebola, swine flu, listeria, etc., from Twitter [6, 7]. The authors have developed a hybrid model consisting of clue-based lexicons that separated the opinionated text from the factual reports and utilized machine learning classifiers to classify the multiple infections.

*Corresponding author Bennett University, India
vatsa.deepika.email@gmail.com (D. Vatsa); ashimayadavdtu@gmail.com (A. Yadav)
ORCID(s): 0000-0002-2032-8076 (D. Vatsa); 0000-0002-1467-1601 (A. Yadav)

The first case of the COVID-19 virus was reported in December 2019 in Wuhan [8]. Since then, many of us are left with the question of “when will this be over?". With each passing year, we see different variants of the deadly virus, resulting in multiple waves of the pandemic. This creates a huge psychological impact in people’s minds as it has been more than two years, and people cannot socialize due to worldwide curfews that have confined them into their respective homes. Currently, the entire world is going through the third wave of the coronavirus, with its new variant, known as Omicron. The first case of Omicron was reported on 02 December 2021 in India [9]. Past studies have focused on the first and second waves of the virus, where researchers have analyzed the impact of the pandemic on people [10]. However, many people have either gotten tired of taking precautions, staying home, and relying on the digital world or are getting used to it.

Hence, in this paper, we aim to study the state of the mind of the people by analyzing the varying pattern of public sentiments over time during the third wave of the virus (Omicron) among citizens of India. We also identify whether the sentiments of people change after the third wave of the pandemic or not. Since now the third wave is getting over in India, we aim to study the after-effects of this wave on the psychology of Indian citizens. In order to gain insights into the experience of the people and uncover public concerns during the third wave of the virus, we apply the topic modeling technique, which extracts the popular topics that are getting significant attention from the public and study the sentiments associated with each of them. We also show the temporal trend in the sentiments of the people. This study will be helpful to the policymakers and the healthcare professionals as they can take timely actions for the well-being of their citizens during any pandemic.

The main contributions of our work are summarized as follows:

1. To the best of our knowledge, this is the first work which focuses on sentiment analysis and retrieving topic discussions on the Omicron-led third COVID-19 wave in India.
2. We performed LDA-based topic modeling on Twitter data geo-located as India to extract the important topics which were prevalent during the third wave of the pandemic.
3. We also analyzed the sentiment trend across different topics on complete two-month data and week-wise data.
4. Finally, we summarized the prominent topics that gained major public attention during the third wave of COVID-19 pandemic in India. We found that the discussions were majorly negative towards “vaccine" topic and neutral towards “precautionary measures" topic.

The rest of the paper is organized as follows: Section 2 reviews the crucial work on sentiment analysis and topic modeling related to the COVID-19 pandemic. Section 3 discusses the proposed methodology. Section 4 focuses on the experimental results and analysis. Section 5 presents the discussion about the analysis. Section 7 concludes the paper with future remarks.

2. Related Work

This section discusses the previous work related to sentiment analysis and topic modeling. Past studies have focused on applying sentiment analysis to study different diseases and during disease outbreaks in public. We also explore the popular work that utilizes topic modeling techniques to discover abstract topics from large textual documents.

2.1. Sentiment analysis

Basiri et al. [11] studied the sentiment intensities of Twitter users for the coronavirus by fusing deep learning techniques like CNN, BiGRU, FastText, DistilBERT, and one machine learning classifier NBSVM. The study aimed to detect the correlation of the Tweets generated at the pandemic with the news and events that gained significant attention from the public. Although the authors targeted eight different countries for this study, selecting the right keyword for searching information and filtering the tweets was independent of the country. Hence, the study could not provide an accurate estimation of the sentiment trend of the people in each country. Priyadarshini et al. [12] analyzed the psychology of the people during lockdown by performing sentiment analysis on the tweets after two and four weeks of lockdown. The study helped to analyze the mental well-being of the citizens during the lockdown. The results show that the people were optimistic and supported the lockdown strategies imposed by the government.

Yousefinaghani et al. [13] extracted the sentiments of the people towards the COVID-19 vaccine by retrieving the tweets and comparing their progression based on time, themes, geographical distribution, and other characteristics. The results show that the people were more interested in discussing about vaccine rejections. People were vaccine-hesitant rather than favoring them or being optimistic towards them. The limitation of their work was that the approach used

by the authors to categorize the sentiments of tweets might have missed some important posts as the entire corpus was not reviewed. Similarly, Liu et al. [14] identified the themes and studied the temporal trends in the COVID-19 vaccine-related tweets in different countries and amongst different states of the US. The study majorly focused on analyzing the sentiments of the citizens before and after the announcement of the Pfizer vaccine. Based on the geographical analysis, the fluctuation patterns of sentiment were influenced by the number of positive cases or reported deaths in that area. Nezhad et al. [15] presented a study that aimed to assess the Persian tweets to analyze the sentiments of Iranian citizens towards the COVID-19 vaccines. The authors compared the sentiments of homegrown vaccine (named Barekat) and imported vaccines like Pfizer, Moderna, AstraZeneca, and Sinopharm. The authors observed that Iranian citizens reflected more positive sentiments towards the imported vaccines as compared to the homegrown vaccine.

Huerta et al. [16] explored the sentiment polarity trend in Massachusetts during the pandemic. The tweets were majorly focusing on increasing the risk in the health of the citizens and anxiety expressions. Das et al. [17] applied CNNs for training the classifier on the COVID-19 tweets. The authors simulated Bayesian regression based model for predicting the future cases of the virus and the recovery rate with respect to the latest scenario.

2.2. Topic modeling

Ridhwan et al. [18] applied emotion analysis using a pre-trained RNN classifier and sentiment analysis using the VADER tool to classify the sentiments into different categories. The authors also applied topic modeling to find the prevalent discussion topics during the COVID-19 pandemic in Singapore. The results show that during the lockdown, the positive sentiment was dominant. However, emotions like fear and joy varied over time due to the developments involved during the pandemic. Chekijian et al. [19] examined the emergency care given to the patients during the pandemic. The authors applied a topic modeling approach to analyze the comments of the patients and uncover the concerns of patient experiences in the hospital. The results show that patients were having many issues regarding their safety, treatment protocols, family or visitors' restrictions, and limitation of testing.

Melton et al. [20] investigated the sentiments of the people towards the COVID-19 vaccine by applying topic modeling on text-based data collected from 13 Reddit communities. The authors applied a topic modeling technique that identified popular topics from the combined dataset and the polarity-wise topics. The polarity analysis was conducted using lexical-based methods, which suggested that most people were showing positive sentiments towards the vaccine-related news. This sentiment remained static over an initial period. However, later on, negative sentiments emerged that were majorly focused on the side-effects of these vaccines as citizens were not confident about them.

Garcia et al. [21] analyzed the tweets of Brazil and the USA for four months by applying different machine learning classifiers: Naïve Bayes, Logistic Regression, Random Forest, Linear SVM, MLP, and AdaBoost. Topic modeling was applied to extract the ten main topics related to both countries, out of which seven topics were common in both. The negative sentiments were prevalent in topics like case statistics, economic impact, and proliferation care. In Portuguese text, the positive sentiment was mainly directed towards politics. Similarly, for English tweets, the highest number of positive messages were written for treatment received by the doctors. The major limitation of their work was that the machine learning-based classifiers failed to generalize and handle the enormous data found on social media platforms.

3. Proposed methodology

3.1. Data Collection and pre-processing

Lopez et al. [22] provided COVID-19 related tweet data set on Github. The authors have continuously collected data set using standard Tweeter API since 22 January 2020. The authors used certain keywords like coronavirus, COVID, mask, vaccine, etc. to collect tweets for the data set [22]. The data set is organised by each hour of the day. It is pre-processed and contains summary details like mentions, hashtags, sentiment scores, Named Entity Recognition (NER) data of tweets. The sentiment scores and NER data are generated using state-of-the-art Twitter Sentiment and Named Entity Recognition (NER) algorithms. For sentiment scores, Cliche's Twitter Sentiment algorithm [23] is used, which is an ensemble model of multiple Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) Networks. The method performed the best in the 11th international workshop on Semantic Evaluation SemEval-2017, where the task was sentiment analysis of Twitter data. For each tweet, the algorithm generates a vector of predicted scores for three sentiment classes: positive, neutral, and negative. Then the tweet is assigned to the sentiment class having highest predicted probability. The algorithm is found to have an accuracy of 75-80% on a sample of COVID-19 tweets, as observed by Rustam et al. [24]. For NER of English language tweets, the authors used the state-of-the-art English NER model provided by Akbik et al. [25].

In India, the first case of the Omicron variant was reported on 2 December 2021 in Bangalore. The Omicron-driven third wave reached its peak on 21 January 2022 with nearly 3,47,000 recorded cases, after which a decline in the number of cases was observed [26]. Thus, in this study for analysis we have used the data of two months, from 1st December 2021 till 31st January 2022. Specifically, we used CSV files from the folder *Summary_Details*, where each CSV file is named as *yyyy_mm_dd_hr_Summary_Details.csv*. For instance, *2022_01_01_00_Summary_Details.csv* file contains data of first hour of 1 January 2022. Each CSV file consists of eight columns having headers as: "Tweet_ID", "Language", "Geolocation_coordinate", "RT", "Likes", "Retweets", "Country", "Date Created". Since we require tweet data of Indians, we pulled out tweet IDs with country code - "IN" and language set to "EN". For the extracted tweet IDs, we then retrieved the corresponding sentiment labels from the CSV files in *Summary_Sentiments* folder in the data set. For topic modeling, we first obtained the tweet message by using the Hydrator application on extracted tweet IDs. The data is preprocessed before applying LDA over it. That is, the data is changed to lower case; then punctuation marks, stop words, and special characters are removed. We also removed hashtags, mentions, and URLs from it. Also, some keywords like Omicron, COVID, COVID19, COVID-19, corona are likely to be present in most of the tweet messages and are thus removed to get crisp and concrete topics. Finally, Lemmatization is performed to get the base word in each tweet message.

3.2. Topic Modeling

Topic modeling or topic discovery is a sub-problem in natural language processing (NLP). The aims is to discover abstract topics discussed in a set of documents, and then classifies any individual document in the set depending upon its relevance to each of the discovered topics. A topic is a set of words taken together to suggest a theme. It is crucial and comes handy when one wants to analyse a huge amount of textual data. It is useful for summarization, similarity estimation, novelty detection, and categorizing a massive collection of documents. In this work, we applied Latent Dirichlet Allocation (LDA) [27] technique on tweet messages to retrieve the topic words.

Assuming our dataset is a collection of multiple documents denoted by $D = d_1, d_2, \ldots, d_n$. Each document d_i is a mixture of different topics, where each topic is a probabilistic mixture over different words that are combined to form a document. Topic modeling is used to explain the hidden information in any document. This can be achieved by grouping the words in such a way that each group represents a topic in a document. Hence, we apply the Latent Dirichlet allocation (LDA), which is a Bayesian hierarchical probabilistic generative model for finding the hidden thematic structure in the unstructured text.

The LDA model utilizes two matrices, namely: $\theta(t_d)$ and $\phi(w_t)$, which are defined as follows:

$$\theta(t_d) = P(\textit{topic } t \mid \textit{document } d) = \textit{Probability distribution of topics in the documents}$$
$$\phi(w_t) = P(\textit{word } w \mid \textit{topic } t) = \textit{Probability distribution of words in the topics}$$

Hence,

$$P(\textit{word } w \mid \textit{document } d) = \phi(w_t) * \phi(w_t)$$

Assuming we have a total number of topics as T, then the probability distribution of words in the documents $P(w|d)$ is explained as below:

$$P(w|d) = \sum_{T} P(t|d) * P(w|t) \tag{1}$$

Where, $*$ represents the dot product and the weights of $\theta(t_d)$ and $\phi(w_t)$ are assigned randomly. The entire process is summarized as below:

1. For every document d, initialize each word randomly to a topic from the distribution of topics based on their assigned weights.
2. For every document d: For every word w in d: Calculate $P(t|d)$ and $P(w|t)$
3. Considering all words and their topics, reassign the topic to the word w based on the dot product of $P(t|d)$ and $P(w|t)$ as shown in Eq (1).
4. Repeat the above step for the entire document until the assigned topics are not changed.

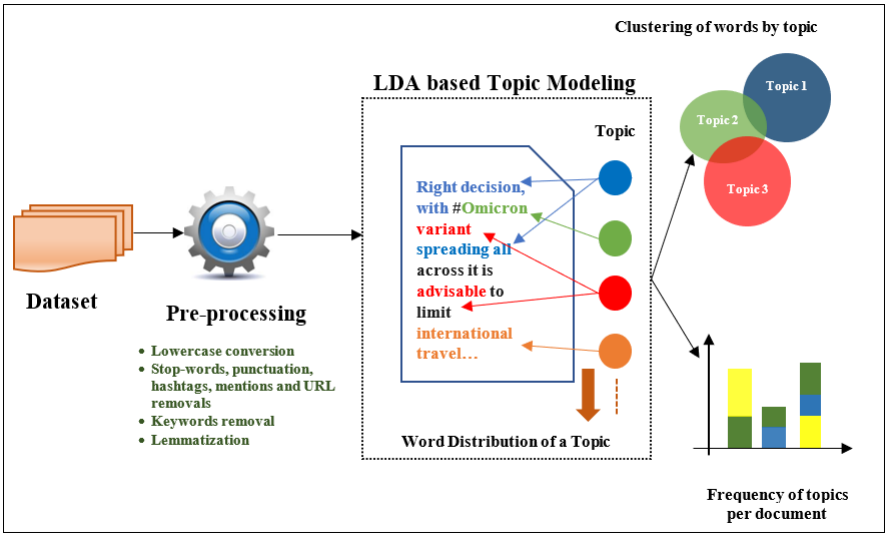


Figure 1: Architecture of the proposed methodology

In topic modeling, we eliminate the list of stop words as they carry no inherent meaning and some repeated keywords. This is done to ensure that the topics generated by the topic modeling approach are not dominated by the stop words or some repeated keywords, and meaningful topics are generated by the algorithm. Figure 1 illustrates the proposed methodology of the work.

4. Experimental Analysis

We have analysed the data set in two ways: firstly, on complete two-month data, and secondly, on weekly data of two months. Analysis of the complete data set gives us an overview of peoples' sentiments and their opinions for the third wave of COVID disease in India. While week-wise analysis of eight weeks data lets us understand the trend and shifts in sentiments and topics discussed by people during the emergence and peak of the third wave in India.

4.1. Complete data analysis

On the complete data set of two months, we have analyzed the retrieved topics, sentiment labels, and topic distribution among sentiments. Table 1 shows the topic words retrieved from the overall tweet data of two months from 1st December 2021 - 31st January 2022. Topic modeling on complete data set yielded six topics. Second table in Table 1 presents the topic themes interpreted by the set of topic words discovered in each topic.

Topic words like "get", "well", "soon", "stay", "safe", "wear", "mask", "up", etc. are categorized under the theme "Precautionary measures". Topic words related to exams like "student", "online", "exam", "sir", "lockdown", etc. are categorized under the theme "Online exam". Statistics related words like "Case", "positive", "report", "update", "test", etc. are categorized as "COVID statistics report". Words like "medicine", "pandemic" when used in conjunction with "artificial", "intelligence", "benefit" are categorized as theme "AI in medicine". Vaccine related words like "vaccine", "dose", vaccination", etc. are categorized as "Vaccine".

Figure 2 shows topic distribution, sentiment distribution, and topic distribution among different sentiment labels on complete twitter data set of two months. As can be seen from the topic bar plot, topic 0 having theme "precautionary measures" was the most discussed while topic 3 having theme "AI in medicine" is the least discussed topic. From the sentiment labels bar plot, out of a total of 42,157 tweets, more than 18,000 tweets carry negative sentiment while 7886 were positive tweets. Thus, the number of negative tweets are more than twice the number of positive tweets. From the sentiments bar plot, we can say that the people of India mostly shared negative or neutral opinion with respect to the third COVID wave in India. Looking at the distribution of topics among different sentiment labels allows us to understand how the sentiments vary for each topic. For instance, out of all topics, topic 1 having theme: "Online exam" carries the most number of negative tweets while topic 3 having theme: "AI in medicine" carries the least number of

Table 1
Topics retrieved from the complete tweet data of two month: December 2021 - January 2022. Second table presents topic themes interpreted by the set of retrieved words in each topic.

	Topic Words
Topic 0	['mask' 'virus' 'get' 'stay' 'well' 'soon' 'people' 'safe' 'take' 'wear']
Topic 1	['exam' 'student' 'online' 'case' 'sir' 'lockdown' 'please' 'pandemic' 'due' 'situation']
Topic 2	['case' 'india' 'variant' 'new' 'virus' 'death' 'positive' 'update' 'report' 'test']
Topic 3	['medicine' 'intelligence' 'artificial' 'great' 'relief' 'pandemic' 'benefit' 'here' 'how' 'world']
Topic 4	['vaccine' 'dose' 'vaccination' 'india' 'year' 'dos' 'crore' 'pm' 'vaccinated' 'via']
Topic 5	['hai' 'nhm' 'employee' 'hp' 'ke' 'pandemic' 'ho' 'india' 'protesting' 'last']

Topic Themes	
Topic 0	Precautionary measures
Topic 1	Online exam
Topic 2	COVID statistics report
Topic 3	AI in medicine
Topic 4	Vaccine
Topic 5	Protest

negative tweets. It shows that Indian students were much affected by the shifting of exams from offline to online mode due to the pandemic. Among all positive tweets, topic 0 theme: “precautionary measure" is the most discussed, while topic 5 (theme: “protest") is the least discussed topic. It shows that people of India were vigilant and were taking proper precautions. The proportion of tweets assigned topic 3 (having theme: “AI in medicine") is almost the same in all three sentiment label categories. Thus people had overall neutral sentiments over topic “AI in medicine".

4.2. Week-wise data analysis

We have sectioned the complete two month data into eight weeks data and analyzed the retrieved topics, sentiment labels, and topic distribution among sentiments on week-wise data set. Figure 3 shows sentiment label bar plots for each week. The trend of distribution of negative, neutral, and positive sentiment tweets is seen to be almost the same in each week with the most number of negative labels, followed by neutral labels, followed by the least number of positive labels. Exceptionally, for week 6, the difference in the number of negative and neutral labels is 17 which can be considered negligible.

Figure 4 shows retrieved topic words, topics bar plot and topic distribution among sentiments for week 1 and week 2. Week 1 corresponds to date range 1st Dec 2021 - 7th Dec 2021 and week 2 corresponds to date range 8th Dec 2021 - 14th Dec 2021. For week 1, topic 0 is the most discussed topic having topic theme interpreted as “new COVID variant". Sentiment distribution over this topic shows approximately twice the number of negative and neutral tweets on this topic than the number of positively labeled tweets. For week 2, the most discussed topic is topic 1, where topic theme is “Vaccine". Sentiment distribution of topic 1 shows that there are more negatively sentiment labeled tweets than positive ones, which reflect that people were not optimistic about the effectiveness of the vaccine developed for COVID virus on the new variant. Topic words in topic 4 and 5 suggest an odd topic, not associated with COVID disease. Topic words suggest the topic theme to be “bail granted to bapuji". It is to note that apart from tweets related to COVID disease, a large number of tweets come from topic 4 and 5 in week 2.

Figure 5 shows the analysis results for week 3 and week 4. Week 3 corresponds to date range 15th Dec 2021 - 21st Dec 2021 and week 4 corresponds to date range 22nd Dec 2021 - 31st Dec 2021. For week 3, the most discussed topic is topic 5 having discussion words as “government", “health", “pandemic", “nhm" (stands for national health mission) suggesting the work done by government employees for public health during pandemic. Again the number of negatively labeled tweets outnumber positively labelled ones for this topic. Other topics discussed are all related to vaccination. The trend of proportion of negative and positive tweets for each topic is the same, i.e., the number of negative tweets are almost twice the number of positive tweets. For week 4, the most discussed topic is topic 2 having topic words as “booster", “vaccine", “dose", “India" suggesting discussion on booster vaccine dose in India for fighting the virus. The

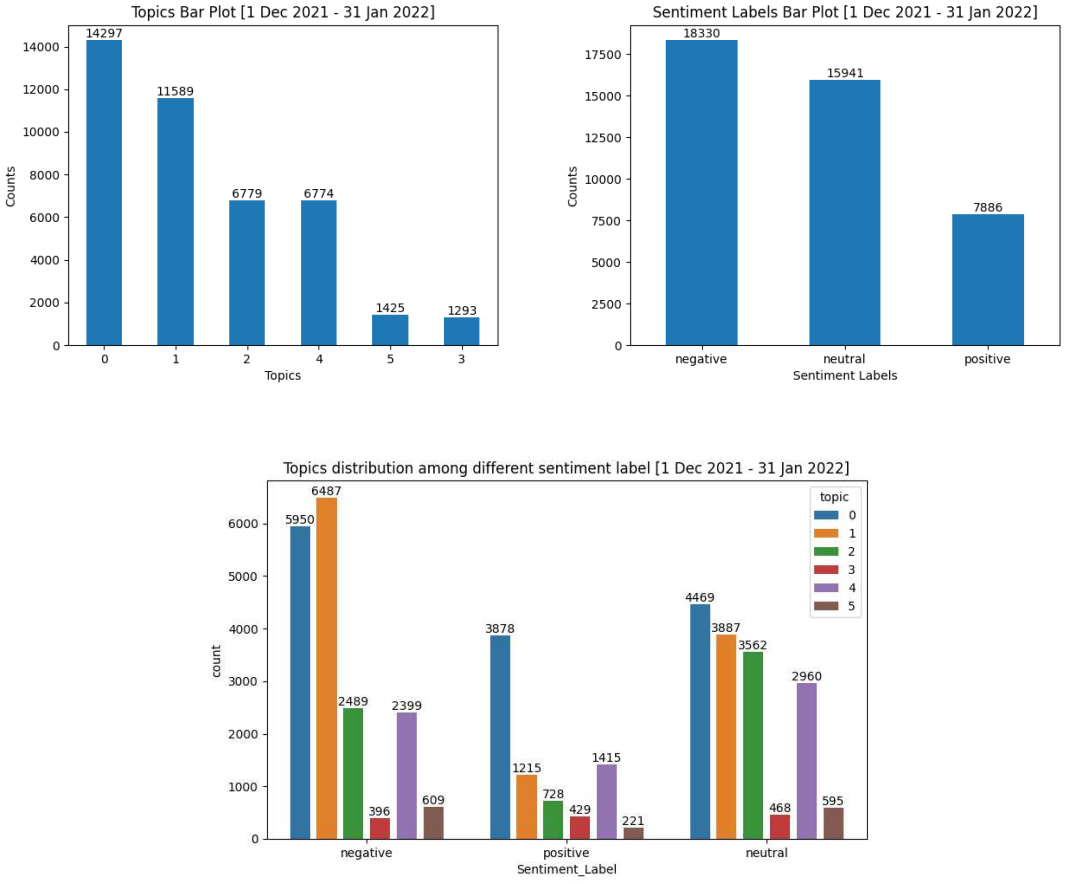


Figure 2: Sentiment and topic distribution of tweet data of two month from 1st Dec 2021 to 31st Jan 2022

number of negative tweets for this topic is 953 which is much higher than the number of positive tweets that is 329. Topic 3 having topic words as “lockdown", “get", “well", “soon", “speedy" “recovery", etc. suggesting topic theme as “precautionary measures and recovery" have comparatively more proportion of positive tweets than other topics.

Figure 6 shows the analysis results for week 5 and week 6. Week 5 corresponds to date range 1st Jan 2022 - 7th Jan 2022 and week 6 corresponds to date range 8th Jan 2022 - 14th Jan 2022. Week 5 marks the beginning of new year 2022 and that was the time when the government of India imposed restrictions in various parts of the country. The most discussed topic in week 5 is topic 2 having topic words as “get", “well", “soon", “lockdown", “new", “year", “positive", “case" suggesting tweets having discussion around lockdown, recovery, new year, and positive cases. While for almost all topics, the tweet sentiment distribution is like the number of negative tweets is around twice or more than twice the number of positive tweets, for topic 2, the number of positive tweets is 636 which is slightly less than the number of negative tweets which is 779. For week 6 also, topic distribution over sentiments is mostly negative than positive, except for topic 4 suggesting topic theme as “recovery" where the number of positive tweets, 86 is slightly less than the number of negative tweets, i.e., 93.

Figure 7 represents analysis results for week 7 and week 8. Week 7 corresponds to date range 15th Jan 2022 - 21st Jan 2022 and week 8 corresponds to date range 22nd Jan 2022 - 31st Jan 2022. For week 7, the most discussed topic is topic 2 having theme “vaccine in India". The trend for all topics is the same that is negative tweets outnumber the positive tweets. For week 8, the most discussed topic is topic 3 having words “pandemic", “vaccine", “India". Looking at the topic distribution among sentiments, topic 4 having “recovery" theme is seen to have around the same proportion for positive and negative tweets. Least number of positive tweets is for topic 5 having topic words related to “elections during pandemic".

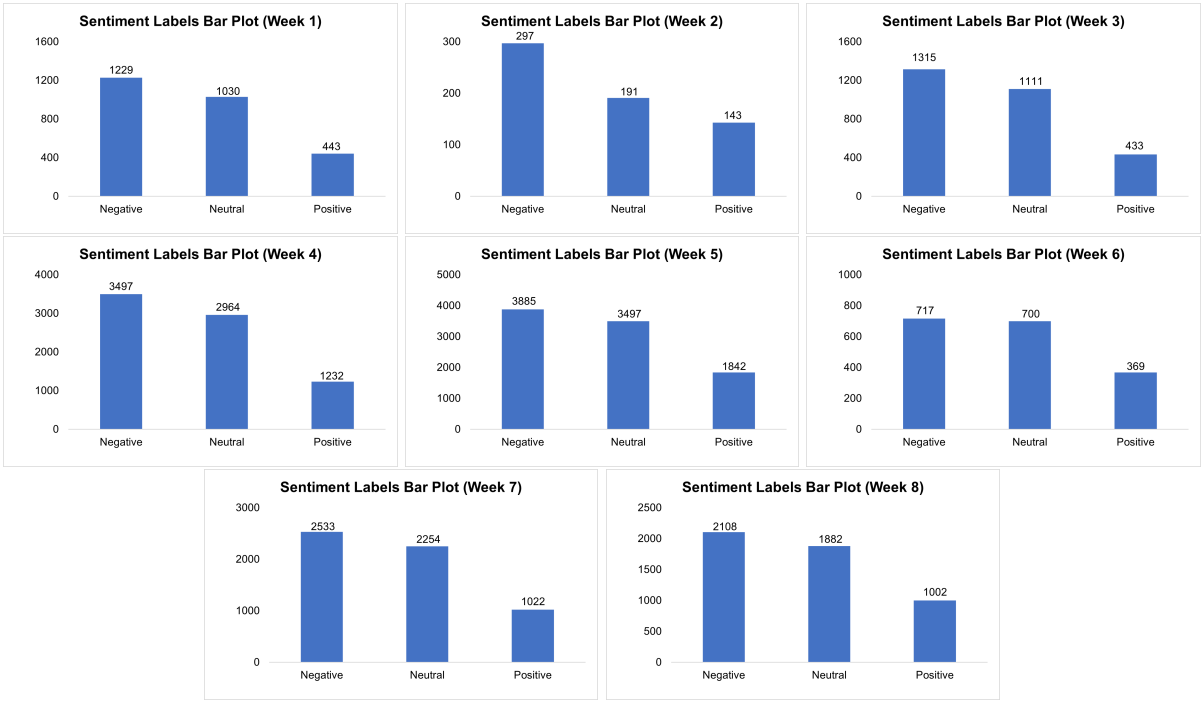


Figure 3: Distribution of sentiment labels across eight weeks starting from 1st December 2021 to 31st January 2022

5. Discussion

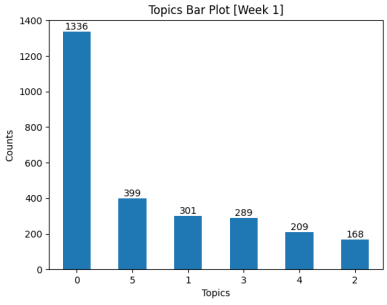
In this work, we have analyzed the Twitter data set related to COVID disease fo two months December 2021 - January 2022. The most discussed topics and their associated number of negative and positive sentiment tweets of week-wise data are summarised in Table 2. Most discussed topics across weeks fall under the themes - “precautionary measures" and “Vaccine". In the table, for all topics, the number of negative tweets is more than triple the number of positive tweets except for weeks 5 and 6, where the difference in the number of negative and positive tweets is not significant. It should be noted that this is due to the fact that in weeks 5 and 6, mostly precautionary measures tweets were posted.

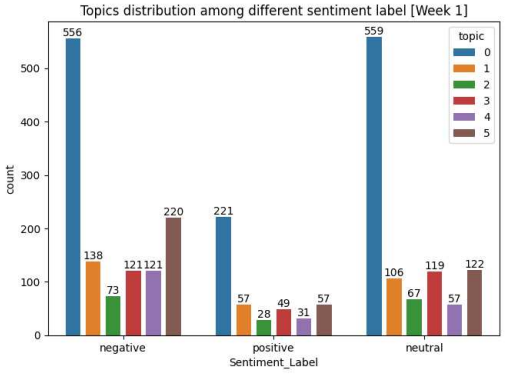
A major challenge that we faced during the analysis of week-wise data was that the extracted words of different topic themes were overlapping each other, thus it was difficult to find concrete topic themes.

The analysis results show that people of India were having more negative sentiments over topics like new COVID variant, vaccines, booster dose, etc. during the third wave of COVID. It is observed that the most of the discussions on twitter across weeks were vaccine-related (See Table 2). Weekly sentiment labels for topic “vaccine" show that the tweets were mainly negative in general. Both COVID vaccine, Covaxin and Covishield had been proven to be effective as in the third wave very few people required hospitalizations those who were vaccinated. Also, vaccination drive of both vaccines, Covaxin and Covishield was started on 16th January 2021 and was done at a quite fast pace. By 1st December 2021, 33% of Indian population (1.39 Billion in 2021) was fully vaccinated and 24% was partially vaccinated with one dose [28]. The negative sentiments of people during the third COVID wave reflect low level of trust of people in the vaccine of COVID disease and the vaccination drive in India.

WEEK 1 [1st Dec - 7th Dec 2021]

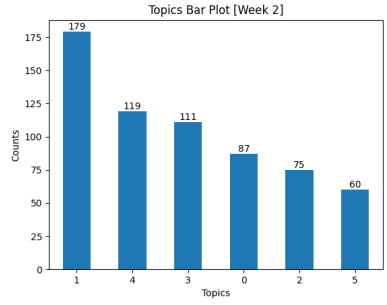
Topic 0	['variant' 'case' 'india' 'new' 'vaccine' 'varient' 'virus' '2' 'people' 'time']
Topic 1	['help' 'sir' 'family' 'college' 'dear' 'please' 'fee' 'job' 'dad' 'gone']
Topic 2	['love' 'life' 'student' 'india' 'pandemic' 'death' 'lok' 'sabha' 'paid' 'off']
Topic 3	['marriage' 'fight' 'mask' 'virus' 'day' 'dcp' 'home' 'without' 'south' 'block']
Topic 4	['nyay' 'family' 'modi' 'victim' 'congress' 'pre' 'compensate' 'demand' 'pandemic' 'death']
Topic 5	['virus' 'death' 'people' 'old' 'india' 'data' 'youth' 'govt' 'lakh' 'due']





WEEK 2 [8th Dec - 14th Dec 2021]

Topic 0	['pandemic' 'leadership' 'management' 'team' 'nhm' 'award' 'response' 'kiit' 'stop' 'line']
Topic 1	['people' 'mask' 'vaccine' 'virus' 'case' '3' 'booster' 'death' 'without' 'new']
Topic 2	['report' 'case' 'variant' 'virus' 'first' 'example' 'fight' 'helping' 'country' 'leader']
Topic 3	['medicine' 'artificial' 'intelligence' 'pandemic' 'great' 'here' 'benefit' 'year' 'mean' 'ayurveda']
Topic 4	['right' 'please' 'human' 'bail' 'sant' 'asharamji' 'bapu' 'shri' 'sir' 'help']
Topic 5	['innocent' 'bapuji' 'government' 'injustice' 'problem' 'long' 'want' 'justice' 'constant' 'veteran']



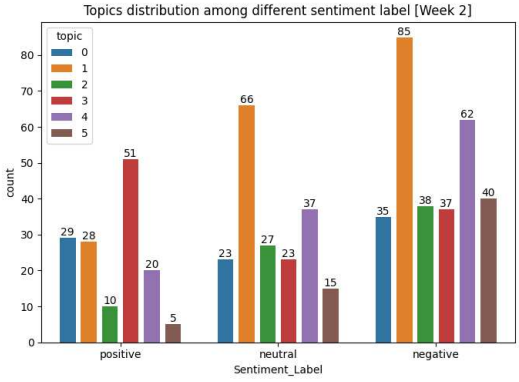
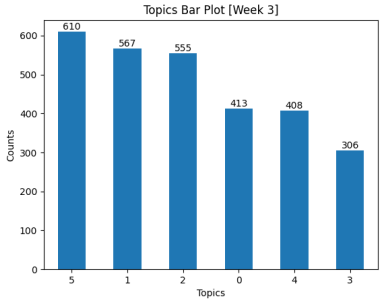
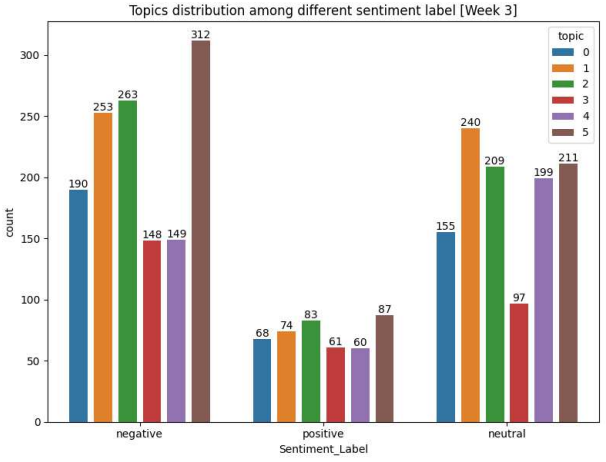


Figure 4: Topic distribution and topic-wise sentiment analysis of tweet data of week 1 and 2 of two month data set [1st Dec 2021 - 31st Jan 2022]

WEEK 3 [15th Dec - 21st Dec 2021]

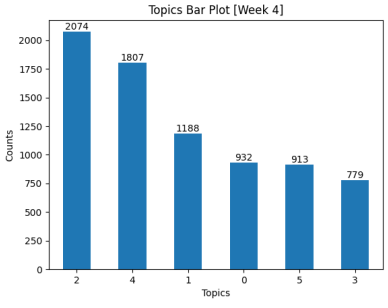
Topic 0	['vaccine' 'people' 'mask' 'need' 'india' 'u' 'careful' 'pandemic' 'jamat' 'time']
Topic 1	['case' 'variant' 'new' 'india' 'virus' 'vaccinated' 'fully' 'mask' 'state' 'report']
Topic 2	['please' 'india' 'also' 'note' 'mask' 'day' 'country' 'exam' 'time' 'vaccination']
Topic 3	['mask' 'time' 'vaccine' 'take' 'people' 'word' 'world' 'get' 'must']
Topic 4	['india' 'marriage' 'vaccine' 'case' 'new' 'time' 'virus' 'variant' 'u' 'tablighi']
Topic 5	['nhm' 'virus' 'worker' 'government' 'stop' 'mask' 'service' 'employee' 'pandemic' 'health']





WEEK 4 [22nd Dec - 31st Dec 2021]

Topic 0	['sir' 'exam' 'student' 'online' 'aktu' 'offline' 'forcing' 'examination' 'want' 'still']
Topic 1	['mask' 'stay' 'safe' 'distancing' 'social' 'wear' 'please' 'home' 'protocol' 'virus']
Topic 2	['vaccine' 'year' 'india' 'hp' 'day' 'booster' 'world' 'dose' 'virus' 'people']
Topic 3	['lockdown' 'soon' 'get' 'well' 'hai' 'recovery' 'mask' 'speedy' 'govt' '7']
Topic 4	['case' 'variant' 'india' 'new' 'election' 'amid' 'virus' 'delhi' 'state' 'night']
Topic 5	['people' 'positive' 'pm' 'must' 'year' 'new' 'relief' 'said' 'modi' 'mask']



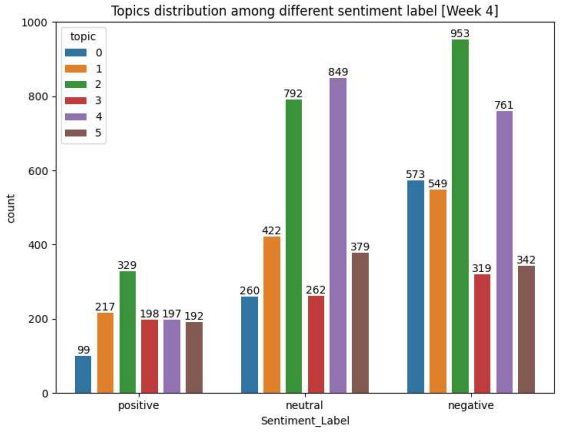
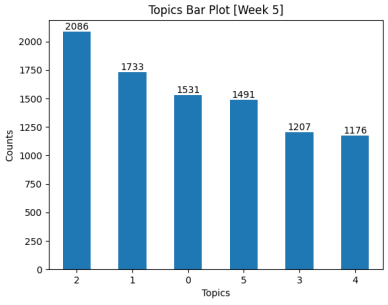
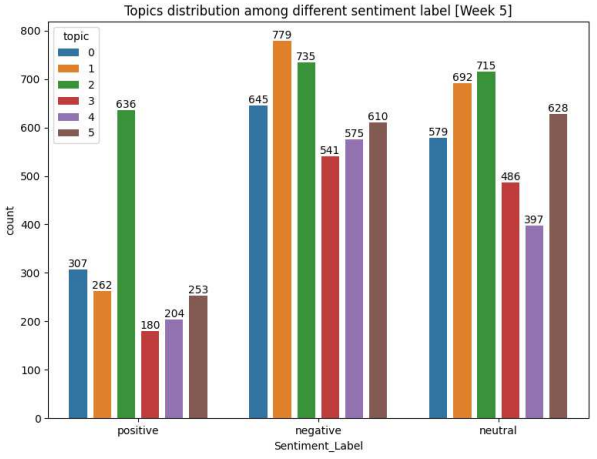


Figure 5: Topic distribution and topic-wise sentiment analysis of tweet data of week 3 and 4 of two month data set [1st Dec 2021 - 31st Jan 2022]

WEEK 5 [1st Jan - 7th Jan 2022]

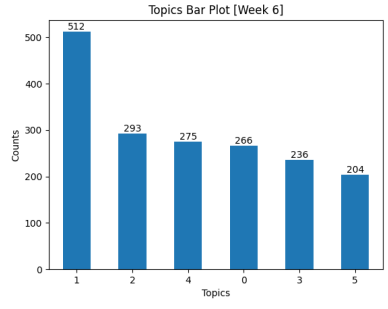
Topic 0	['mask' 'stay' 'safe' 'india' 'please' 'wear' 'home' 'time' 'day' 'variant']
Topic 1	['india' 'vaccine' 'people' 'case' 'positive' 'virus' 'test' 'day' 'go' 'u']
Topic 2	['soon' 'get' 'well' 'lockdown' 'take' 'sir' 'year' 'new' 'case' 'positive']
Topic 3	['case' 'new' '24' 'death' 'last' 'hour' 'india' 'report' 'mask' 'pandemic']
Topic 4	['year' 'college' 'time' 'new' 'people' 'virus' 'case' 'student' 'protocol' 'get']
Topic 5	['vaccine' 'vaccination' 'virus' 'dose' 'people' 'year' 'vaccinated' 'age' 'india' 'first']





WEEK 6 [8th Jan - 14th Jan 2022]

Topic 0	['exam' 'positive' 'online' 'test' 'case' 'please' 'pandemic' 'time' 'day' 'health']
Topic 1	['get' 'soon' 'well' 'case' 'vaccine' 'student' 'sir' 'take' 'dose' 'mask']
Topic 2	['mask' 'stay' 'lockdown' 'safe' 'wear' 'india' 'please' 'people' 'public' 'without']
Topic 3	['india' 'case' 'vaccine' 'pandemic' 'pm' 'app' 'variant' 'day' '2022' 'mask']
Topic 4	['recovery' 'speedy' 'virus' 'wish' 'get' 'soon' 'wishing' 'well' 'ji' 'positive']
Topic 5	['virus' 'case' 'new' 'day' 'india' 'test' 'today' 'positive' 'take' 'like']



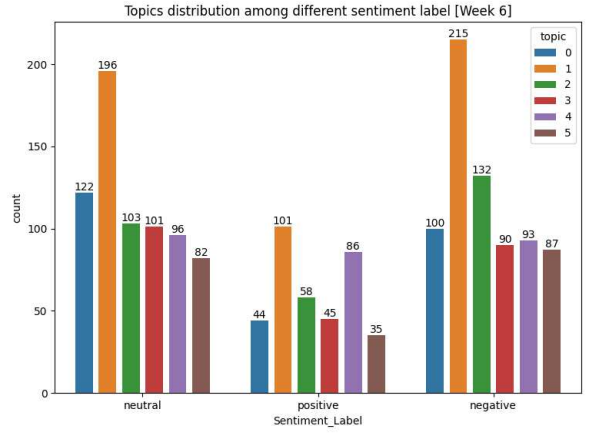
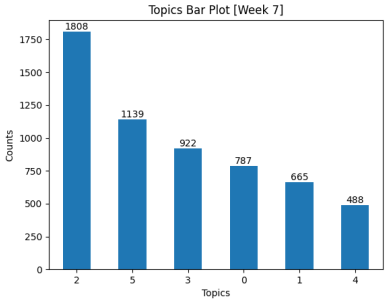
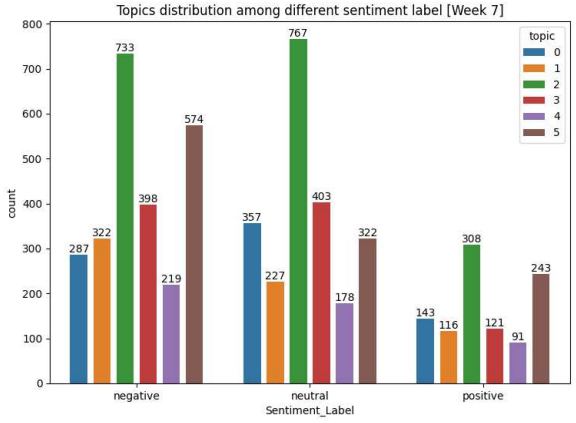


Figure 6: Topic distribution and topic-wise sentiment analysis of tweet data of week 5 and 6 of two month data set [1st Dec 2021 - 31st Jan 2022]

WEEK 7 [15th Jan - 21st Jan 2022]

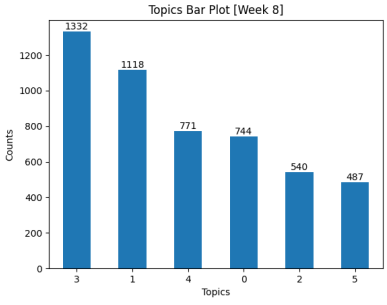
Topic 0	['case' 'exam' 'stay' 'want' 'respected' 'chief' 'minister' 'board' 'increase' 'online']
Topic 1	['sir' 'fast' 'get' 'request' 'pandemic' 'examination' 'wave' 'growing' 'people' 'u']
Topic 2	['vaccine' 'india' 'case' 'vaccination' 'year' 'dos' 'update' 'dose' 'day' 'today']
Topic 3	['lockdown' 'mask' 'vaccine' 'people' 'wear' 'virus' 'vaccinated' 'get' 'pandemic' 'even']
Topic 4	['virus' 'food' 'pandemic' 'agenda' '80' 'world' 'strength' 'free' 'delivered' 'citizen']
Topic 5	['exam' 'student' 'online' 'please' 'soon' 'take' 'well' 'get' 'sir' 'offline']





WEEK 8 [22nd Jan - 31st Jan 2022]

Topic 0	['update' 'school' 'virus' 'new' 'death' 'time' 'must' 'case' 'year' '2022']
Topic 1	['mask' 'exam' 'stay' 'virus' 'please' 'u' 'take' 'safe' 'wear' 'home']
Topic 2	['mask' 'case' 'look' 'state' 'india' 'leadership' 'without' 'forward' 'new' 'united']
Topic 3	['pandemic' 'vaccine' 'india's' 'via' 'fight' 'people' 'doesnt' 'vaccinated' 'exist' 'year']
Topic 4	['soon' 'get' 'well' 'day' 'test' 'please' 'sir' 'positive' 'republic' 'india']
Topic 5	['virus' 'people' 'pandemic' 'new' 'case' 'death' 'election' 'student' 'help' 'protocol']



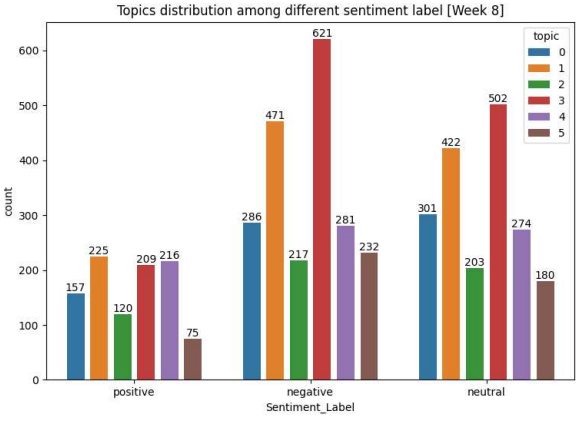


Figure 7: Topic distribution and topic-wise sentiment analysis of tweet data of week 7 and 8 of two month data set [1st Dec 2021 - 31st Jan 2022]

Table 2
Summary of analysis on most discussed topic in each week using week wise data set

Week	Most discussed topic	Negative tweets	Positive tweets
Week 1	New Variant in India	556	221
Week 2	Precautionary measures, Vaccine	85	28
Week 3	Work done by Govt for public health	312	87
Week 4	Vaccine	953	329
Week 5	Precautionary measures, New year, Positive cases	735	636
Week 6	Precautionary measures, Vaccine	215	196
Week 7	Vaccination in India	733	308
Week 8	Vaccine in India	621	209

6. Conclusion

This study identifies the topics and sentiments of Indian citizens about the Omicron-driven third wave of COVID-19 using the Twitter data. Analysis over the two-month Twitter data examines various topics discussed and the changes in these topics over time to understand better the trend of public opinion and perception about the third wave. It also examines the public sentiments about the different topics on the complete and weekly sectioned data set. Among the six distinct topics, the most discussed topics remained “precautionary measures" and “vaccine". While the proportion of negative and positive sentiments over the topic “precautionary measures" is almost similar, negative sentiments outnumber the positive sentiments by a large extent over the latter topic. This suggests that though the third wave was considered to be the “mild" wave of COVID-19 in India, the people of India were still in fear of catching the disease again and thus were taking precautions. It should be noted that the developed COVID-19 vaccines, i.e., Covaxin and Covishield had proven to be quite effective, as from the people who were vaccinated, very few of them required hospitalizations. Also, the vaccination drive taken by the government of India was fast in pace. Still, the negative sentiments toward the topic discussions on “vaccine" reflect a low level of trust of Indians in either the efficacy of the developed vaccine to fight the new variant or the vaccination drive carried out in India or both. A more close attention to the tweets related to vaccine discussion is needed to understand this.

Such kind of study is extremely helpful for public health agencies to understand the major concerns of people and their varied reactions to different issues. In this study, public health agencies can refer to the distinct topic themes and their associated number of negative and positive tweets to understand the concerns of the general public of India and provide them with more information needed in case of topics where most people have reacted negatively.

In the future, we could expand this work to explore the emotion pattern and behavioral changes of people surrounding the third wave of pandemic across different countries.

7. Declarations

Ethical Approval and Consent to participate: Not applicable

Human and Animal Ethics: Not applicable

Consent for publication: Not applicable

Availability of supporting data:

The datasets analysed during the current study are available in the “COVID19_Tweets_Dataset” repository at https://github.com/lopezbec/COVID19_Tweets_Dataset

Competing interests: The authors declare that they have no competing interests

Funding: Not applicable

Authors’ contributions: DV conceptualized the idea, extracted and interpreted the data set, performed methodological experiments, and analysed results. AY wrote the first draft. All authors interpreted the results, edited and approved the final manuscript.

Acknowledgments: Not applicable

Authors' information: DV is currently an Assistant Professor in the School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Uttar Pradesh, India. She received her Ph.D. degree in Computational Biology from the Indian Institute of Technology Delhi, India. Her current research interest includes Systems biology, natural language processing, machine learning, and sentiment analysis.

AY is currently an Assistant Professor in the School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Uttar Pradesh, India. She received her Ph.D. degree in Deep learning from Delhi Technological University, New Delhi, India. She has been awarded the "Research Excellence Award" by Delhi Technological University, Delhi, India, in the years 2020 and 2021. Her current research interest includes deep learning, natural language processing, machine learning, fake news detection, Deepfakes identification, Emotion Recognition, and sentiment analysis.

References

[1] Statista: Number of Twitter users worldwide from 2019 to 2024, https://www.statista.com/statistics/303681/twitter-users-worldwide/. Accessed: 22-02-2022.

[2] WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020, https://tinyurl.com/2p8fytjc. Accessed: 22-02-2022.

[3] WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/. Accessed: 22-02-2022.

[4] Q. B. Baker, F. Shatnawi, S. Rawashdeh, M. Al-Smadi, Y. Jararweh, Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets, JUCS - Journal of Universal Computer Science 26 (2020) 50–70.

[5] A. Culotta, Towards Detecting Influenza Epidemics by Analyzing Twitter Messages, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 115–122. URL: https://doi.org/10.1145/1964858.1964874. doi:10.1145/1964858.1964874.

[6] K.-W. Fu, H. Liang, N. Saroha, Z. T. H. Tse, P. Ip, I. C.-H. Fung, How people react to Zika virus outbreaks on Twitter? A computational content analysis, Am. J. Infect. Control 44 (2016) 1700–1702.

[7] H. Liang, I. C.-H. Fung, Z. T. H. Tse, J. Yin, C.-H. Chan, L. E. Pechta, B. J. Smith, R. D. Marquez-Lameda, M. I. Meltzer, K. M. Lubell, K.-W. Fu, How did Ebola information spread on twitter: broadcasting or viral spreading?, BMC Public Health 19 (2019) 438.

[8] Coronavirus disease (COVID-19) pandemic, https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov. Accessed: 22-02-2022.

[9] India's first Omicron cases detected in Karnataka, https://www.hindustantimes.com/india-news/indias-first-omicron-cases-detected-in-karnataka-101638445884205.html. Accessed: 23-02-2022.

[10] A. Yadav, D. K. Vishwakarma, A Language-Independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis, ACM Trans. Internet Technol. 22 (2021).

[11] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, U. R. Acharrya, A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets, Knowledge-Based Systems 228 (2021) 107242.

[12] I. Priyadarshini, P. Mohanty, R. Kumar, R. Sharma, V. Puri, P. K. Singh, A study on the sentiments and psychology of twitter users during COVID-19 lockdown period, Multimed. Tools Appl. (2021) 1–23.

[13] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, S. Sharif, An analysis of COVID-19 vaccine sentiments and opinions on twitter, International Journal of Infectious Diseases 108 (2021) 256–262.

[14] S. Liu, J. Liu, Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis, Vaccine 39 (2021) 5499–5505.

[15] Z. Bokaee Nezhad, M. A. Deihimi, Twitter sentiment analysis from Iran about COVID 19 vaccine, Diabetes & Metabolic Syndrome: Clinical Research & Reviews 16 (2022) 102367.

[16] D. Thorpe Huerta, J. B. Hawkins, J. S. Brownstein, Y. Hswen, Exploring discussions of health and risk and public sentiment in Massachusetts during COVID-19 pandemic mandate implementation: A Twitter analysis, SSM Popul. Health 15 (2021) 100851.

[17] S. Das, A. K. Kolya, Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network, Evol. Intell. (2021) 1–22.

[18] K. Mohamed Ridhwan, C. A. Hargreaves, Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore, International Journal of Information Management Data Insights 1 (2021) 100021.

[19] S. Chekijian, H. Li, S. Fodeh, Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the COVID-19 pandemic, Health Technol. (Berl.) 11 (2021) 1073–1082.

[20] C. A. Melton, O. A. Olusanya, N. Ammar, A. Shaban-Nejad, Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence, Journal of Infection and Public Health 14 (2021) 1505–1512. Special Issue on COVID-19 – Vaccine, Variants and New Waves.

[21] K. Garcia, L. Berton, Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA, Applied Soft Computing 101 (2021) 107057.

[22] C. E. Lopez, C. Gallemore, An augmented multilingual Twitter dataset for studying the COVID-19 infodemic, Soc. Netw. Anal. Min. 11 (2021) 102.

[23] M. Cliche, BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 573–580. URL: https://aclanthology.org/S17-2094. doi:10.18653/v1/S17-2094.

[24] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, G. S. Choi, A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis, PLOS ONE 16 (2021) 1–23.

[25] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 54–59. URL: https://aclanthology.org/N19-4010. doi:10.18653/v1/N19-4010.

[26] Omicron peaked on January 21 with 3,47,000 daily cases, urlhttps://economictimes.indiatimes.com/news/india/omicron-peaked-on-january-21-with-347000-daily-cases/articleshow/89335010.cms.

[27] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[28] E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, L. Rodés-Guirao, A global database of COVID-19 vaccinations, Nature Human Behaviour 5 (2021) 947–953.