Koushik Reddy Parukola (Spring 2023)

Topic Modeling and Sentiment Analysis on tweets related to COVID-19 Vaccine in India (1328 words)

Introduction:

On 16th of January, 2021, India started giving out COVID-19 shots. India had provided nearly 2.2 billion doses of all current vaccines licensed by the government as of 4 March 2023, including the first, second, and booster doses. The Press Information Bureau reports that two vaccines, Covishield (produced by Serum Institute of India Ltd.) and Covaxin (developed by Bharath Biotech Ltd.), have been used to start the COVID-19 vaccination program in India. This research paper focus on the how people feel and what are the major topics of discussion on the covid vaccination program in India, which is analyzed by sentiment analysis using VADER and topic modeling using LDA. The data for this analysis used is extracted from twitter using hashtags or searching the name of various vaccines given out in India.

Research Question:

1. What were the prominent topics on Twitter in India when the COVID-19 vaccine program initially began?

2. What impression did the COVID immunizations have on people in India in initial stages (positive or negative)?

Data:

For this research we used a pre-existing dataset which was developed from twitter data on the vaccines available in entire world (Pfizer, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, Sputnik V) [DATA]. The dataset initially consists of 16 columns and 2,28,208 data points.

Attribute	Definition
• ID	Specifies the Twitter User ID
• user_name	Specifies the Twitter User Name
• User_location	Specifies the user Location
• User_description	Specifies the user's description in Twitter
• User_created	Specifies the date the user is created in twitter
• User_followers	Specifies the number of followers for the user
• User_friends	Specifies the number of friends for the user
• User_favourites	Specifies the number of favourites of user
• User_verified	Specifies if the user is verified by twitter or not
• Date	Specifies the date of tweet
• Text	Specifies the Tweet's text
• Hashtags	Specifies the Hashtags in the tweet
• Source	Specifies the source of the tweet
• Retweets	Specifies the number of retweets for the tweet
• Favourites	Specifies the number of favourites for the tweet
• Is_retweet	Specifies the number of re-tweets for the tweet

Table 1. Data Variables and its definitions

For our analysis we pre-process the data to extract the tweets only with user_location as India. We then create a subset from the dataset with the tweets from 16th of January 2021 (date when vaccination begun in India) to end of November 2021 using the 'date' attribute, as our analysis focus on the initial reaction of people on the vaccinations. As said, initially the dataset has 2,28,208 data points, after all the pre-processing we are left with 19,181 datapoints for our analysis.

Analysis:

For sentiment analysis, after retrieving the subset of the data, I applied a function to remove all hyperlinks, hashtags, emoticons, pictures, and symbols. Our data will be ready after we drop all the null values from the data frame. Then applying the sentiment analyzer from VADER from NLTK module gives us the sentient scores based on the words used in the text. Using the sentiment scores, we visualize the people reactions.

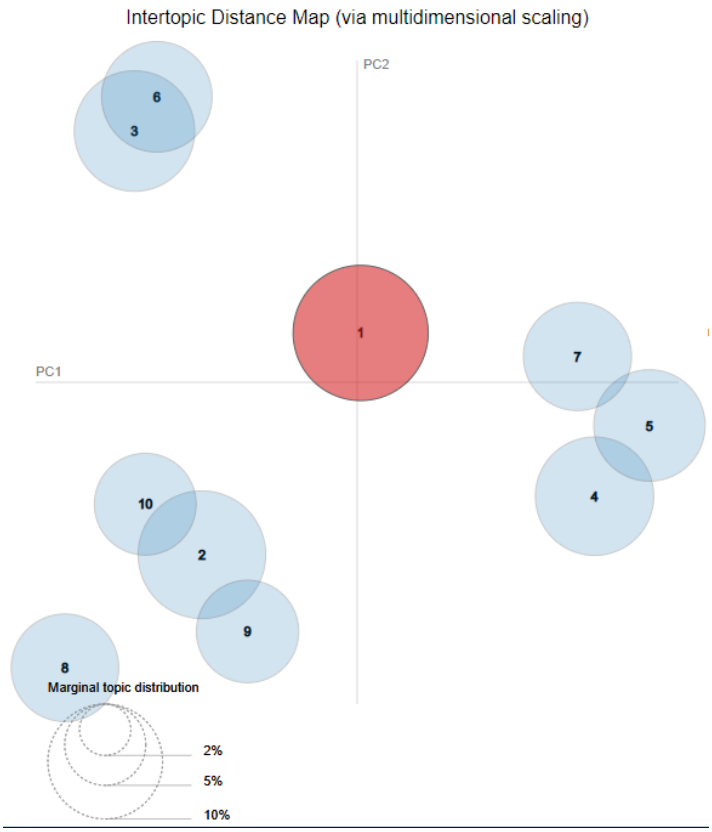
	text	date	month	sentiment_score
0	If we do not witness the same thing in India t...	2021-01-17 06:11:07	1	0.4404
1	has launched a detailed investigation after 2...	2021-01-16 05:47:45	1	-0.4767
2	23 people die in after receiving officials Do...	2021-01-16 05:46:34	1	-0.5994
3	A global rollout suffered a major blow Friday ...	2021-01-16 03:02:45	1	-0.6705
4	29 Dead in Norway after Getting Vaccinated Not...	2021-01-17 10:52:42	1	-0.6486

Table showing the sentiment scores of the tweet along with the date

But in the case of topic modeling, we need much more processing after we remove all hyperlinks, hashtags etc. I used the built-in function of the gensim package to remove stop words and lowercase the Tweet content. The data was then lemmatized and tokenized using spacy, an open-source tool for advanced natural language processing following which we create the data corpus. This corpus is used to for topic modeling using the Latent Dirichlet Allocation (LDA) from gensim module. Finally, we can use the coherence score of this model to see the data and model fit. One common method for determining the optimal parameter values in topic modeling is to use GridSearchCV, which enables the search for the best combination of parameters that provides the best fit for the data. After conducting the parameter search using GridSearchCV, the best number of topics can be selected based on the evaluation metric used. In this case, the number of topics has been set to 10 based on the results of the parameter search. Below are some of the top words from 5 topics out of 10.

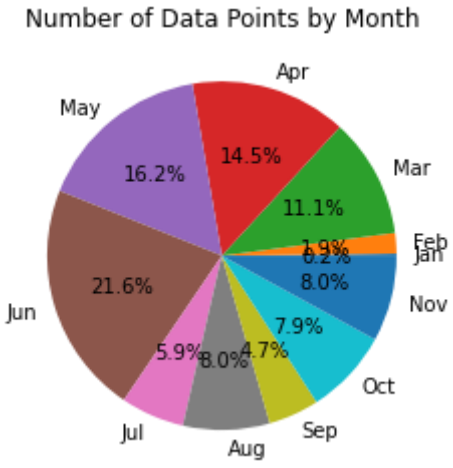
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
trial (0.08)	phase (0.06)	efficacy (0.04)	show (0.04)	datum (0.04)
vaccinate (0.10)	get (0.03)	dose (0.02)	day (0.02)	state (0.02)
approve (0.09)	emergency (0.07)	vaccine (0.05)	approval (0.05)	effective (0.03)
slot (0.03)	work (0.03)	vaccine (0.03)	booster (0.02)	safe (0.02)
take (0.08)	vaccination (0.07)	dose (0.07)	today (0.04)	vaccine (0.03)
get (0.07)	dose (0.06)	production (0.04)	finally (0.03)	second (0.02)
dose (0.09)	vaccine (0.06)	available (0.05)	week (0.02)	administer (0.02)
vaccine (0.13)	covid (0.04)	list (0.02)	kid (0.01)	new (0.01)
amp (0.03)	thank (0.03)	approval (0.02)	world (0.02)	indian (0.02)
vaccine (0.04)	good (0.03)	news (0.02)	government (0.02)	soon (0.02)

Top words in topic 1 to 5



A Inter Topic distance map for all the topics

An Inter-topic distance map is created by calculating the similarity between different topics in a topic model based on the distribution of words in those topics. In this figure we can see some topics clusters colliding with each other, it means that there are some similarities in those topics. So, by reviewing them we can combine them if that are too similar else, we can leave them as they are.

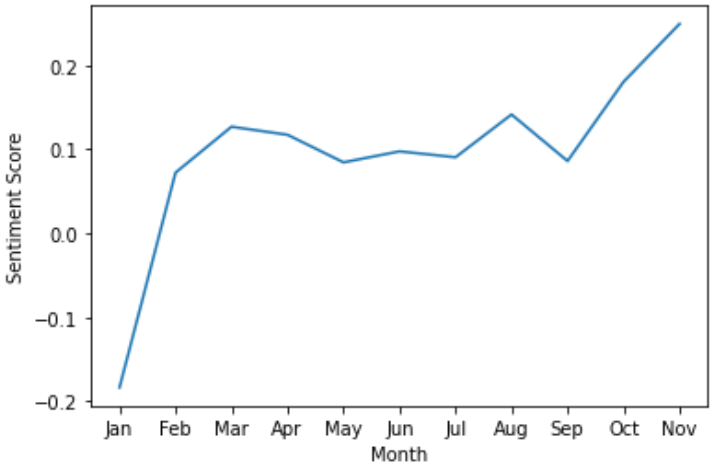


Data distribution in each month

The data in the pie chart above talks about the data from each month remining after all the pre-processing and data cleaning. January is the month with the lest number data points as it only 36 data points contributing 0.2% to the whole data and the most data points are from June with 4,141 data points which contributes about 21.6% to the whole data.

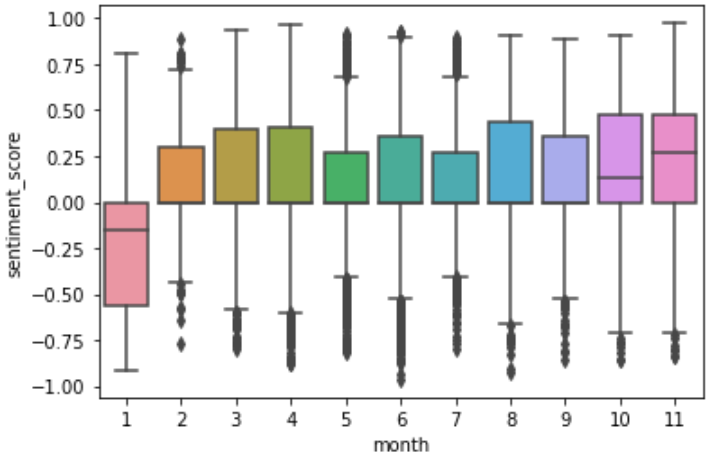
Results:

Sentiment analysis:



A Line plot of Sentiment polarity of tweets over months

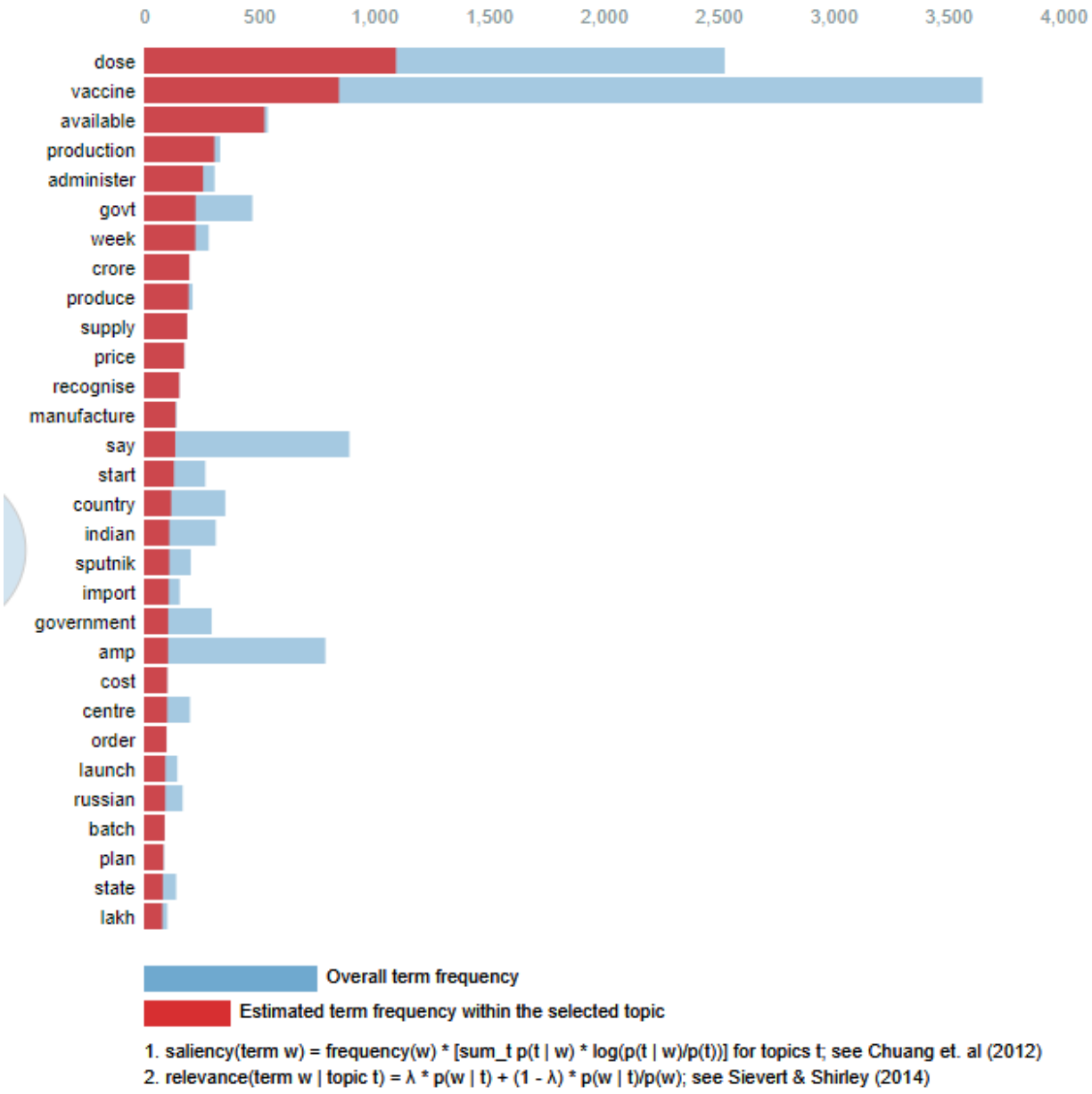
The sentiment scores averaged over each month are shown in the graph above. We can see that over several months, people's opinions toward the COVID vaccination have gradually improved. Since people did not know the vaccine's efficacy, there was a lot of negativities when vaccinations started in January. Additionally, there is a platform called "CoWin" where individuals can schedule an appointment for vaccinations was not fully effective. Numerous appointment cancellations and scheduling errors made people less inclined to support vaccination campaigns. But over the course of just two months, people gradually began to believe that the vaccine had no side effects after witnessing its effectiveness.



A box plot of sentiment scores over months

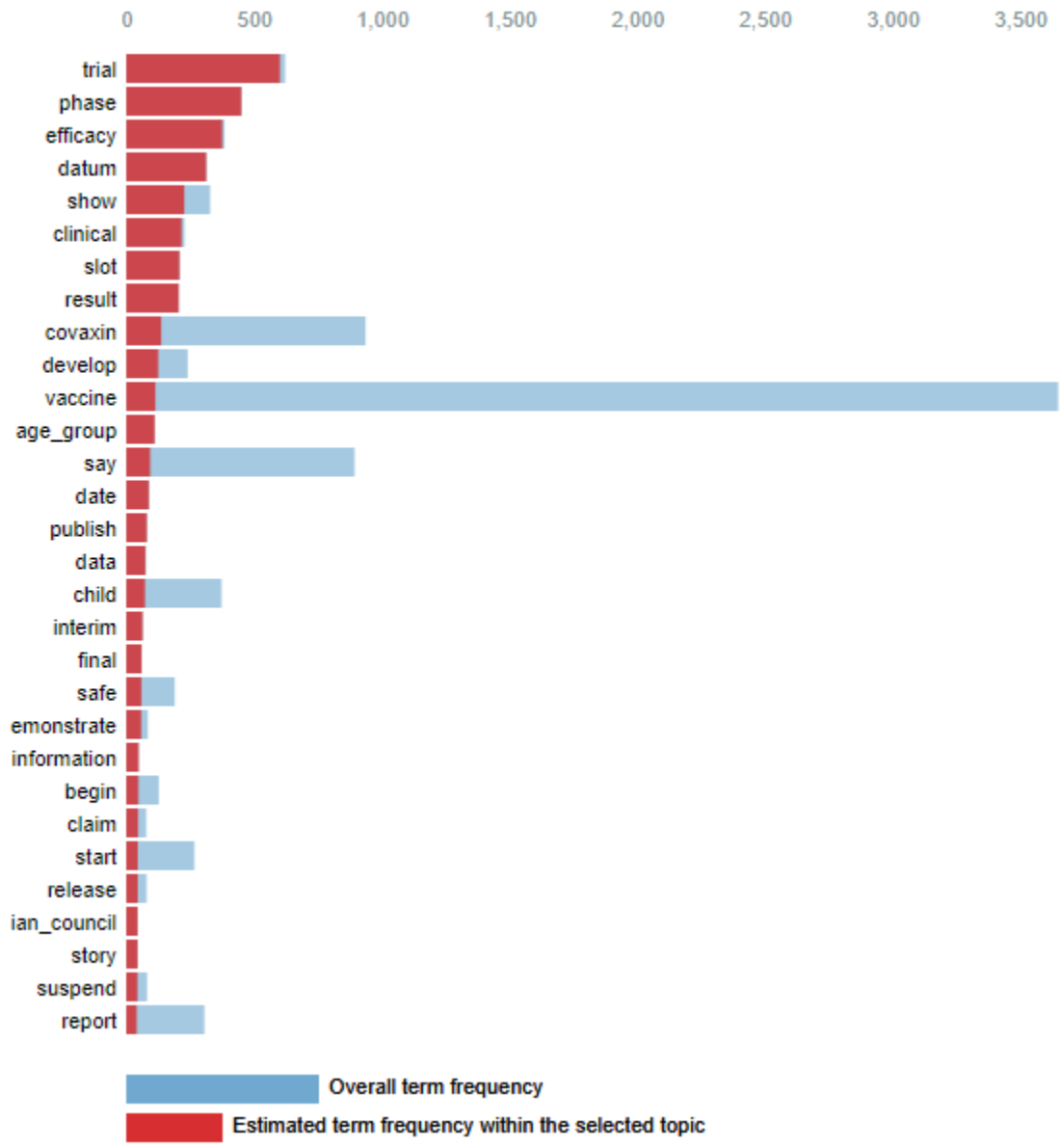
The box plot shows the monthly range of sentiment scores. The top of the box represents the 75th percentile of the data, and the bottom of the box represents the 25th percentile. The whiskers extend from the box to the highest and lowest values. Excluding January, the boxes in all other cases are more skewed towards positive sentiment scores. This shows that there is generally more positive than negative sentiment represented in the data throughout these months.

Topic Modeling:



Top 30 Most Relevant terms for topic 1

Analyzing the word distribution of the most relevant terms can provide valuable insights into the main ideas and concepts associated with a topic. For instance, in the case of the top 30 most relevant terms from topic 1, it can be inferred that the documents with this topic are likely discussing vaccine availability and the number of doses. This inference is based on the observation that the top words in the distribution for this topic are "dose," and "vaccine," which are the two most frequent terms in the topic. Similarly, to we can examine other topics.



Top 30 Most Relevant Terms for Topic 8

When we analyze the top 30 most relevant terms from topic 8, it suggests that the documents related to this topic are probably focused on the stages of vaccine trials and their effectiveness, based on the occurrence of words such as "trial," "phase," and "efficiency" that are the most frequent terms in this topic's distribution

Conclusion:

The analysis of sentiment scores shows that January had the most negative tweets about the COVID vaccination. This could be due to various reasons, such as issues with vaccination appointments or vaccine availability. However, as the months progressed, we observed a gradual decrease in negative tweets and an increase in positive ones. This could indicate that people were becoming more comfortable with the vaccination process and the availability of vaccines. Furthermore, by using topic modeling, we were able to identify the most important topics discussed in relation to the vaccination drive. These topics included the effects of vaccination, vaccine dose availability, and government

interest in conducting vaccination drives more frequently. It is encouraging to see people getting vaccinated as quickly as possible, and this could be due to increased government efforts to promote and make vaccines accessible. The findings of this analysis can help inform and guide future efforts to promote vaccination and address concerns related to it.

Limitations:

The data used for our analysis consisted of Twitter data on various vaccines from around the world, spanning from May 2020 to November 2021. However, for our specific analysis, we required data only from India and within a specific date range from January 16th 2021 to the end of 2021. Due to the Twitter API blocking third-party apps, we were not able to retrieve data exactly as per our requirements. In case of our LDA topic modeling a low coherence score is due to repeated words in each topic. Using various hashtags can increase the diversity of the data and would have been able to retrieve better topics from the data.

References:

Dataset: Gabriel Preda. (2021). COVID-19 All Vaccines Tweets [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2845240

- Sanjeet Bagcchi. The world's largest COVID-19 vaccination campaign NewsDesk| Volume 21, Issue 3, p323, March 2021. DOI: https://doi.org/10.1016/S1473-3099(21)00081-5

- SV P, Lorenz JM, Ittamalla R, Dhama K, Chakraborty C, Kumar DVS, Mohan T. Twitter-Based Sentiment Analysis and Topic Modeling of Social Media Posts Using Natural Language Processing, to Understand People's Perspectives Regarding COVID-19 Booster Vaccine Shots in India: Crucial to Expanding Vaccination Coverage. Vaccines. 2022; 10(11):1929. https://doi.org/10.3390/vaccines10111929

- Lande, J., Pillay, A., & Chandra, R. (2023). Deep learning for COVID-19 topic modelling via Twitter: Alpha, Delta and Omicron. https://doi.org/10.48550/arXiv.2303.00135