International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 09 Issue 11 ǁ 2021 ǁ PP. 62-69

COVID-19 Vaccination Twitter Data Analysis for India

[*1]Biraaj Rout, *Department of Computer Engineering, Ambedkar Institute of Technology, Bangalore, India*
[2]Deeksha Pandit, *Department of Computer Engineering, Ambedkar Institute of Technology, Bangalore, India*
[3]Dr Harish G, *Department of Computer Engineering, Ambedkar Institute of Technology, Bangalore, India*
[4]Dr Smitha Shekar B, *Department of Computer Engineering, Ambedkar Institute of Technology, Bangalore, India*

Abstract:
Vaccination is the foundation of the elimination of contagious diseases like polio, mumps, pertussis etc. yet, vaccines have been customarily linked up to apprehensiveness and hesitancy among the public and the same is the case with the COVID-19 vaccines. Social media platforms have given insights into the definite acceptance of vaccines. Our dataset [1] consists of around 200 thousand tweets gathered from 12 Dec 2020 to 16 Oct 2021. Although the public is concerned about its side effects, unavailability in certain areas, long term effects and so on, the data shows that the public is willing to take the vaccine jabs as quickly as possible and thus the acceptance rate of vaccines is relatively high. Additionally, businesses need to know what their customers or users perceive and anticipate about the product. Our case analysis of vaccine sentiments will help companies know what the customers think about the vaccines, any kind of side effects and also how safe the vaccines are according to the consumers. This way companies could make better products i.e. vaccines, a better way of providing vaccines, work on ways to reduce side effects and so on.

Keywords: *Vader, TextBlob, SentiWordNet, Covid-19, Sentiment, Analysis, Twitter, Vaccine, Covishield, Covaxin, Sputnik V, WordCloud, Python, Matplotlib, Numpy, Panda, Coronavirus, India*

Date of Submission: 29-10-2021 Date of acceptance: 13-11-2021

I. Introduction:

The coronavirus outbreak in the year 2019 has been a disaster for all the countries and people all over the world. The World Health Organization (WHO) labelled COVID-19 as a pandemic [2]. Since then numerous research teams in large pharmaceutical companies and institutions around the world have been coming up with vaccines. In the year 2021, the vaccines were released for the frontline workers and the general public. All over the world, people took to Twitter to express their views about the COVID-19 vaccine. In this paper, the dataset is made using Tweepy to collect the data and tweets were analysed using methods such as sentiment analysis to survey the behaviour and sentiments of people in India. Public response to the COVID-19 vaccination was diverse but the data used concludes that people had positive feedback about the dose. Scrutinizing the data from Twitter helps doctors, health specialists, researchers and lawmakers to understand the public's opinion of the COVID-19 vaccination. Analysing the public's reaction helps make better and safer vaccines, helps the authorities frame better policies and gives estimations about the manufacturing and production of the vaccines.

II. Methods:

2.1 Data Acquisition:

Data was collected from the "COVID-19 All Vaccines Tweets" dataset present in Kaggle with data dating from 12 Dec 2020 to 16 Oct 2021. It consists of 15 columns such as hashtags, user_name, user_description, user_friends, user_created, user_location, user_followers, user_favorites, user_verified, is_retweet, date, text, source, retweets, favourites. Geographical information was available for 70% of the tweets and 17.4k tweets were from verified accounts. A total of 207,000 tweets were extracted in the dataset in the form of comma-separated values (CSV). 79.1k unique vaccine data was collected using the following terms "Oxford/AstraZeneca", "Sinopharm", "Covishield", "Moderna", "Covaxin", "Pfizer/BioNTech", "Sputnik V".

2.2 Preliminaries:

The two major approaches for sentiment analysis:

- Unsupervised approach.
- Supervised machine learning approach.

As there is no pre-labelled dataset, the first approach taken is SentiWordNet, TextBlob and Vader for sentiment analysis.

2.2.1 TextBlob:

TextBlob being a library in python for processing text data enables an API for plummeting into natural language processing (NLP) tasks such as noun phrase extraction, part-of-speech (POS) tagging, classification. sentiment analysis, translation, and more. TextBlob outputs the following two metrics for any input text. A float value of polarity ranges within [−1, 1]. −1 symbolises negative sentiment, 1 symbolises positive sentiments, and 0 means neutral sentiment. Also, a float value of subjectivity lies in the range of [0, 1].

2.2.2 Vader:

Valence Aware Dictionary and Sentiment Reasoner is a rule and lexicon-based tool for sentiment analysis that is sharply adjusted to opinions expressed in social media like Facebook, Twitter, etc. Vader is memory efficient compared to other machine learning algorithms. VADER's approach helps us to decode and quantify the emotions contained in streaming media such as text, audio or video. The speed-performance trade-off is not critical in Vader. It has F1 Classification Accuracy of 0.96 which beats individual human raters with an accuracy of 0.84 [3][4].

2.2.3 SentiWordNet:

A major application of SentiWordNet is opinion mining in a lexical manner. Objectivity, positivity and negativity are three sentiment scores each synset in SentiWordNet is assigned [5]. Every score exists within 0.0 to 1.0, and their sum is 1.0 for each synset. This means that a synset for all the three categories may have nonzero scores, which would indicate that the corresponding terms have, in the sense indicated by the synset, each of the three opinion-related properties only to a certain stage.

2.3 Methodology:

Here is the description of the end-to-end process starting from data collection to obtaining results and the tools used along with steps taken to analyse the sentiments related to various vaccines in India.

2.3.1 Computational tools or libraries:

Python is used as the basic programming language and its libraries for sentiment analysis. Data were obtained from the Kaggle "COVID-19 All Vaccines Tweets" dataset [1]. For text processing, NLTK [Link] was employed. TextBlob, Vader and Sentiwordnet were used to do sentiment analysis. TextBlob, Vader and Sentiwordnet were adopted for sentiment analysis. Various word clouds were formed using python's "wordcloud" library [Link].

2.3.2 Environment:

This project was performed using a personal laptop with configurations of AMD Ryzen 7 (8 CPU*2ghz), 8 Gb Ram and 512 Mb SSD, 64-bit windows 11.

2.3.3 Sentiment Analysis:

Below is the description of the major step-by-step process taken for sentiment analysis of the Covid-19 vaccine dataset. Figure1 depicts the schematic diagram for various blocks in our sentiment analysis methodology for Covid-19 sentiment analysis.

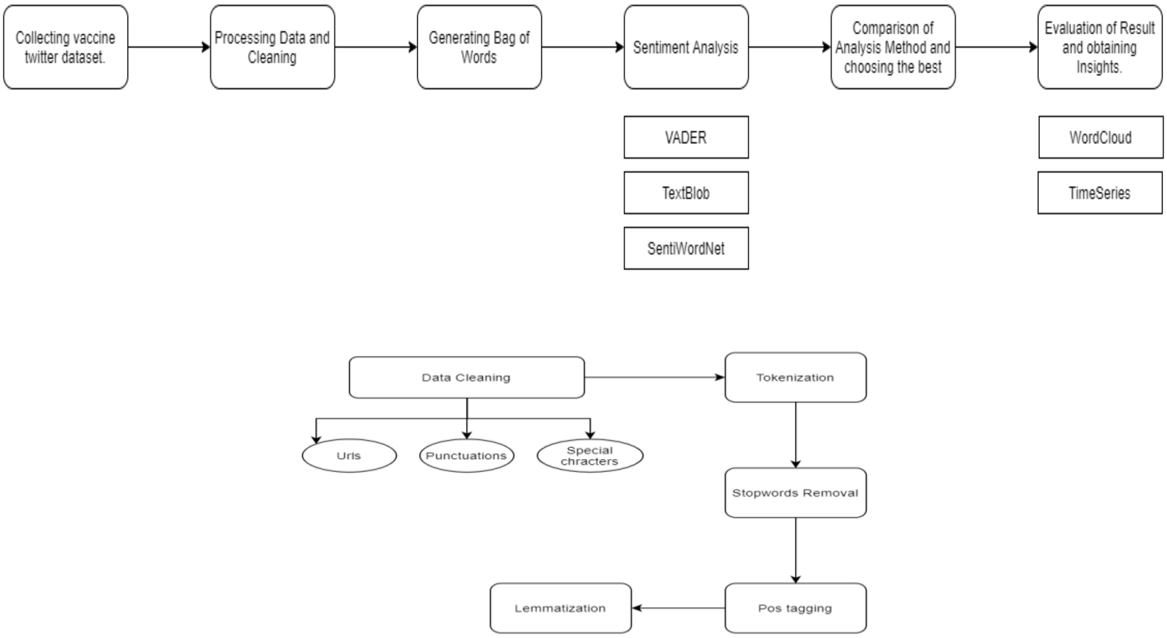


Figure1: Workflow for sentiment analysis on Twitter data using TextBlob and VADER and SENTIWORDNET

2.3.4 Data Pre-processing:

- **Data Cleaning:** Removal of mentions, hashtags, retweet information, and links/URLs.
- **Removal of Redundancy:** Duplicate texts are removed keeping only single instances.
- **Date Time Conversion:** Since only a particular date is required rather than time information all of the dates and times are converted to only date.
- **Tokenization:** a list of words is created from a single text.
- **Stopwords Elimination:** Removal of the most frequently used English words such as "I", "me", "and", "or" from each text using NLTK stop words list [6][7]. Common words such as "not", "never" and so on were not removed because they would change the emotion predictions.
- **Parts of Speech (POS) tagging:** To denote the part of speech each work in a text corpus is given a unique label and other grammatical classification such as case, number (plural/singular) and tense [8][9].
- **Lemmatization:** Getting root words according to parts of speech.

2.3.5 Sentiment Analysis:

Three sentiment analysis tools are used: TextBlob, Vader and SentiWordNet, to get the sentiments of the tweets. All three have different scoring methods but are used to yield only three sentiments positive, neutral and negative. Classification can be carried out using machine learning and lexicon-based techniques. In the machine learning approach, a labelled dataset is required which can be used to train using algorithms and test the result obtained. Lexicon based method has a polarity associated with each word which helps in predicting the score for a sentence. Lexicon based approach is most suitable for unlabelled data and also is the easiest one for predicting sentiment. As our data is raw and unlabelled, tools that follow the lexicon-based approach are used (Vader, TextBlob and SentiWordNet). Word clouds [10] are generated to visualize the important words based on the frequency of the words initially and also used the log-likelihood ratio to generate the word clouds later which turned out to be more insightful [11].

III. Experimental Results:

In this section, the results are summarized.

3.1 Manual Comparison of Vader, TextBlob and SentiwordNet:

After looking at the counts of positive, neutral and negative sentiments in all three methods using a pie chart in Figure 2 the below points are observed:

- Neutral sentiments are more than positive and negative sentiments on Vader and TextBlob but SentiWordNet has more positive sentiments.
- Overall positive sentiments are higher in number.
- Both Vader and TextBlob results were similar except for results from SentiWordNet.

- Comparing sentiments manually in Figure 3 it is observed that TextBlob performs better in most of the cases so it is used primarily.

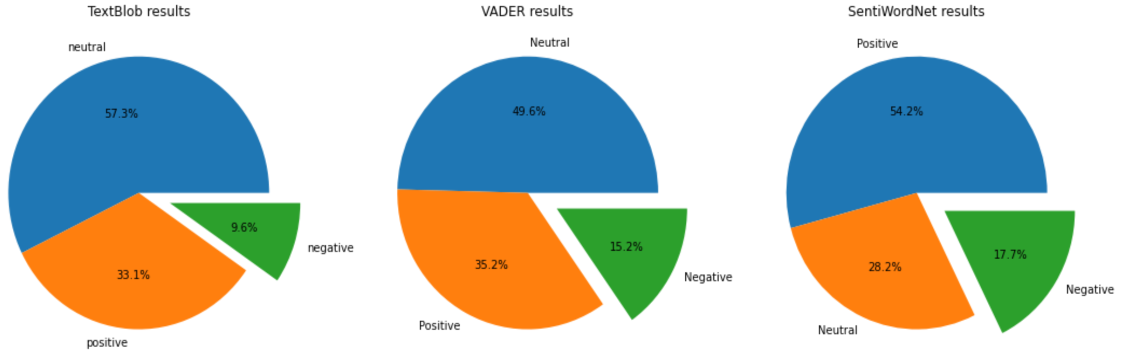
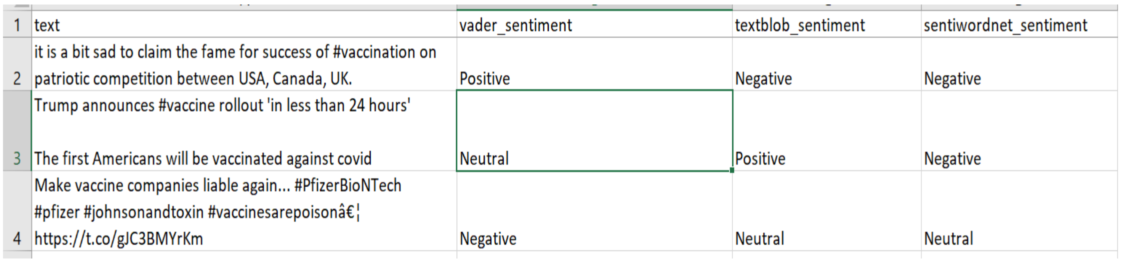


Figure 2: Pie chart of Vader, TextBlob and SentiWordNet sentiment distribution.



1	text	vader_sentiment	textblob_sentiment	sentiwordnet_sentiment
2	it is a bit sad to claim the fame for success of #vaccination on patriotic competition between USA, Canada, UK.	Positive	Negative	Negative
3	Trump announces #vaccine rollout 'in less than 24 hours' The first Americans will be vaccinated against covid	Neutral	Positive	Negative
4	Make vaccine companies liable again... #PfizerBioNTech #pfizer #johnsonandtoxin #vaccinesarepoisonâ€¦ https://t.co/gJC3BMYrKm	Negative	Neutral	Neutral

Figure 3: Manual comparison of user sentiment results using Vader, TextBlob and SentiWordNet.

3.2 Vaccination Public Sentiment:

In this section, the results of sentiment analysis using TextBlob in various scenarios are depicted.

3.2.1 Analysis of Tweets by Location:

Covishield/Astrazeneca, Covaxin, Sputnik V are vaccines accepted in India and have been in major discussion in all parts of India. So to confirm this assumption filtered tweets (Figure 4, Figure 5, Figure 6) were examined which were related to the above-mentioned vaccines counts by grouping them and found most of the tweets are from Indian locations and the majority of the tweets are from Bangalore, India. Hence the graphs below support our hypothesis.

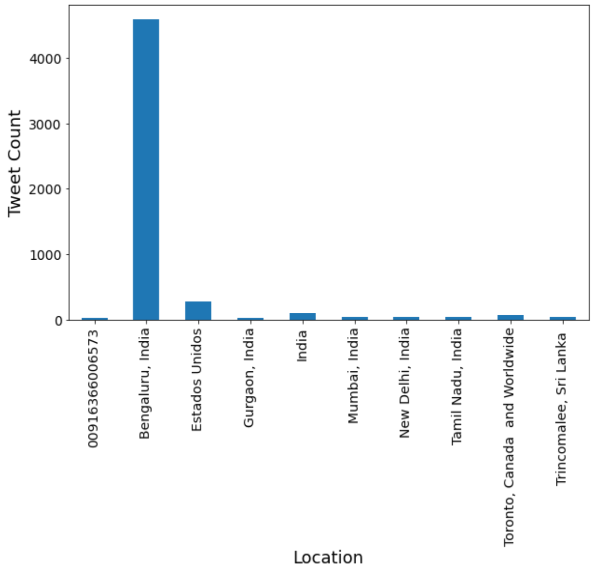


Figure 4: Covishield location graph

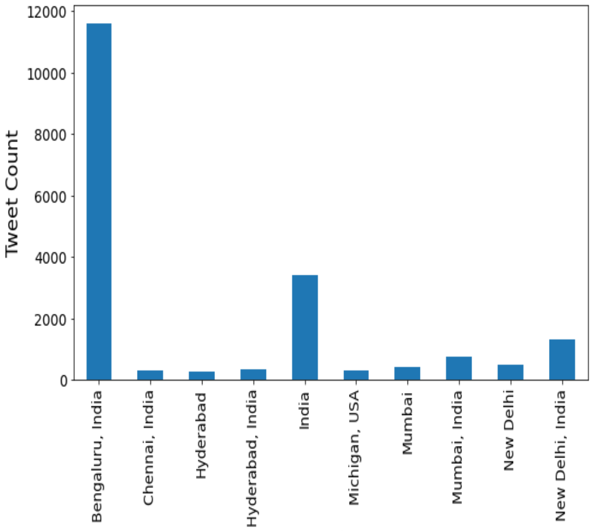


Figure 5: Covaxin location graph

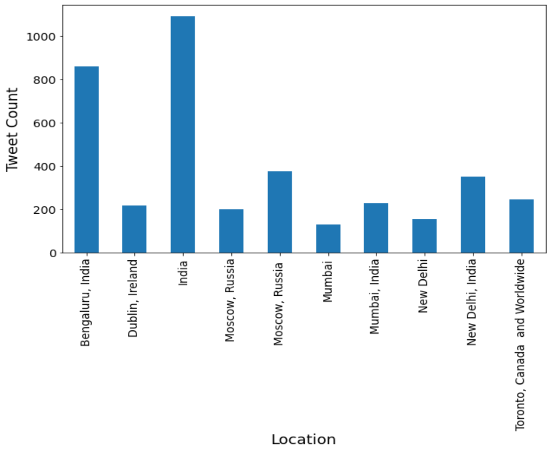


Figure 6: Sputnik V location graph

3.2.2 Timeline for Vaccination Tweets:

Considering the vaccines used in India multiple timeline graphs (Figure 7, Figure 8, Figure 9, Figure 10) were derived with a TextBlob polarity score to check out the positivity rate of sentiments and which dates have a surge in tweets related to a particular vaccine. This helps us to have an insight into important topics on any given day. For instance, Covaxin in figure 8 below, one of the highest peaks is observed on March 1 2021 because the prime minister of India took the first jab of Covaxin that day. Using TextBlob provided sentiment for the timelines and the results for each vaccine timeline are mentioned in Figure7, Figure 8 and Figure 9. Most of the vaccine tweets are on the positive side of the sentiment looking at polarity scores. Covishield has had a more positive response after June 28 as 10.80 crore doses of Covishield were produced in India over the month of June. Also, Covishield was internationally accepted and included in the World Health Organisation list around that time, whereas Covaxin still needs international acceptance which is why its polarity is a little bit on a lower side than that of Covishields response. Sputnik V didn't have much response recorded because it was imported in a small quantity compared to other vaccines in India. A high peak on April 12 for Sputnik V is observed which was due to approval of the Russian vaccine in India. Overall response is more on the positive side than on the negatives which show acceptance of vaccination is on the higher side in India.

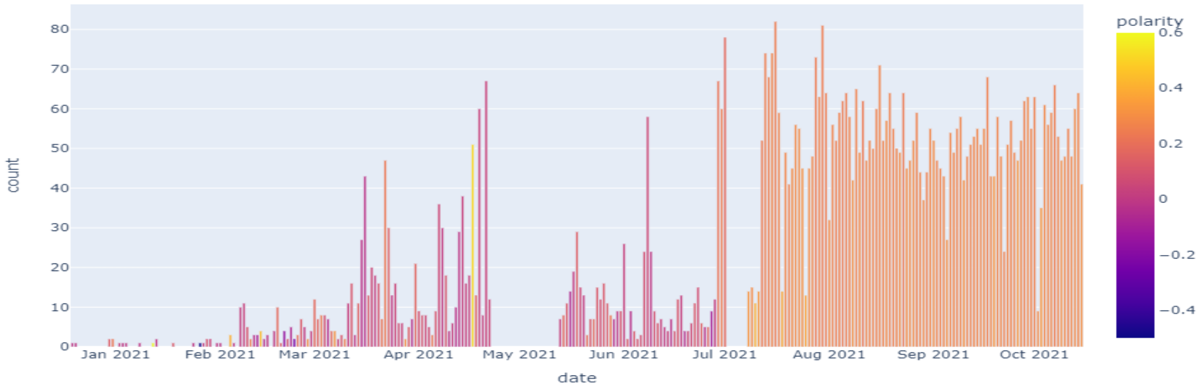


Figure 7: Covishield timeline

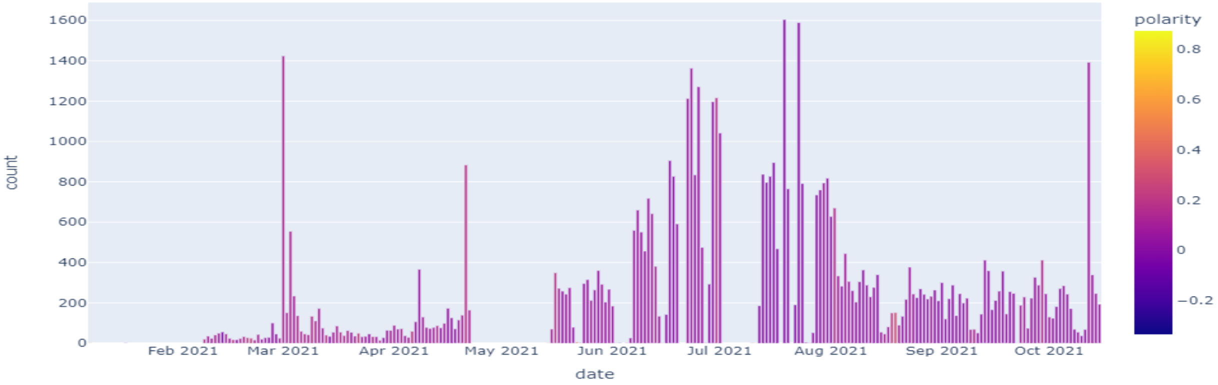


Figure 8: Covaxin timeline

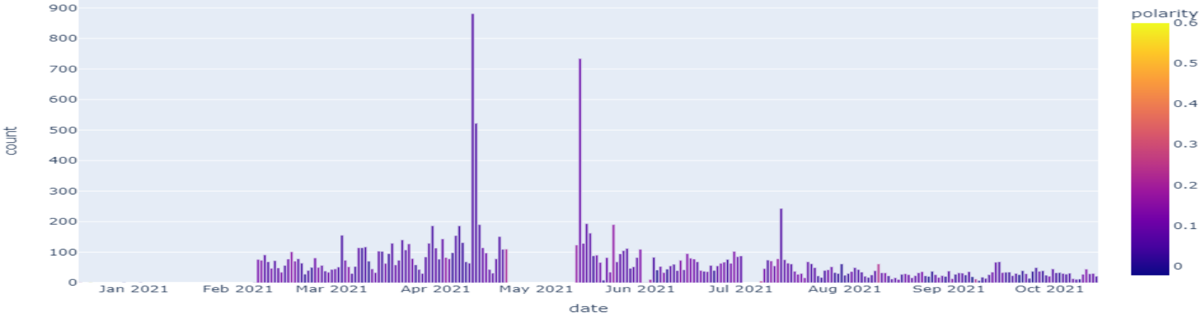


Figure 9: Sputnik V timeline

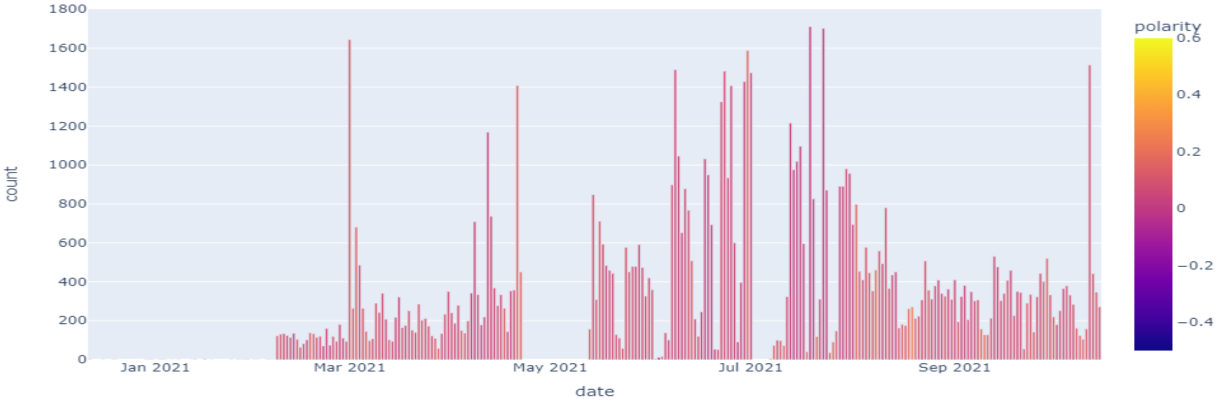


Figure 10: All Indian vaccines timeline

3.2.3 High-Frequency Words:

The categories of the words are positive, neutral and negative. Initially, word clouds were directly generated on the raw text which yielded results shown in Figure 11. Figure 11 shows the top frequently used words in the vaccination tweets. This approach did not produce any useful information. Later, log-likelihood was used for each vaccine available in India for showcasing high occurrence words for positive, negative and neutral sentiments. Figure 13 shows "risk", "blood" and "age" as primary negative words since Covishield had some rare blood clot reactions and it had different reactions for varying age groups. Figure 14 shows "effective", "free", "good" as positive words as it was free in most of the states of India and a result by Bharath BioNTech showed it was also effective against delta variant of covid-19. Sputnik V in figure 15 has "fake", "propaganda" as negative frequent words since the news was spread that Sputnik V was part of Russian publicity which created fear among vaccine takers. Thus negative, positive and neutral words for each Indian vaccine using these word clouds were found.

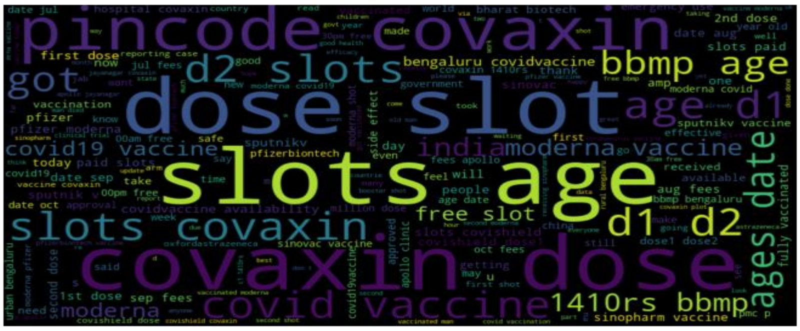


Figure 11: All countries vaccines wordcloud without any loglikelihood.

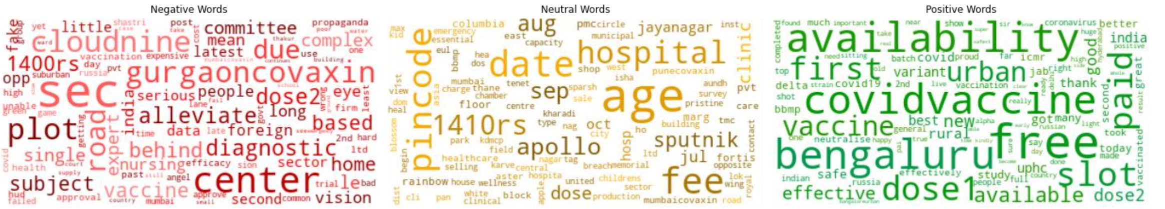

Negative Words
Neutral Words
Positive Words

Figure 12: All Indian vaccines sentiment wordcloud with loglikelihood

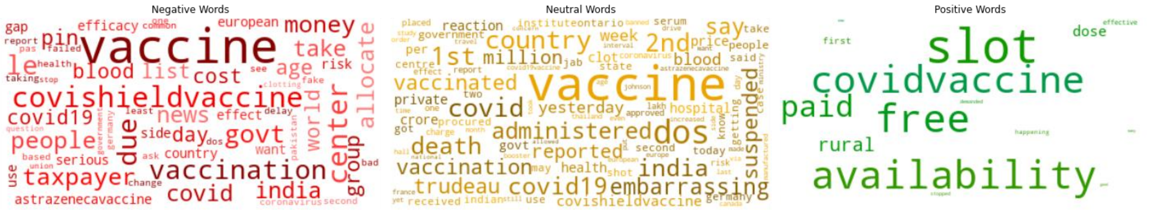

Negative Words
Neutral Words
Positive Words

Figure 13: Covishield wordcloud with loglikelihood.

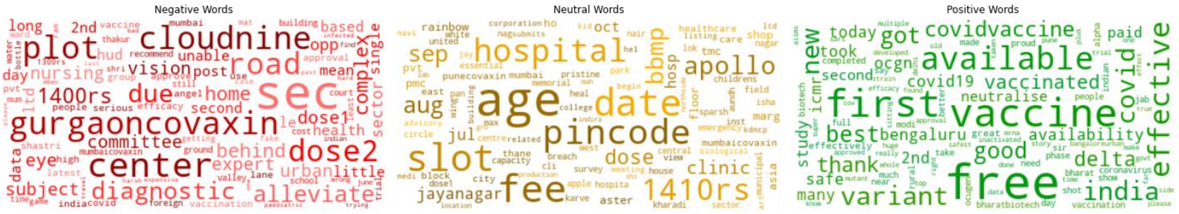

Negative Words
Neutral Words
Positive Words

Figure 14: Covaxin sentiment wordcloud with loglikelihood.

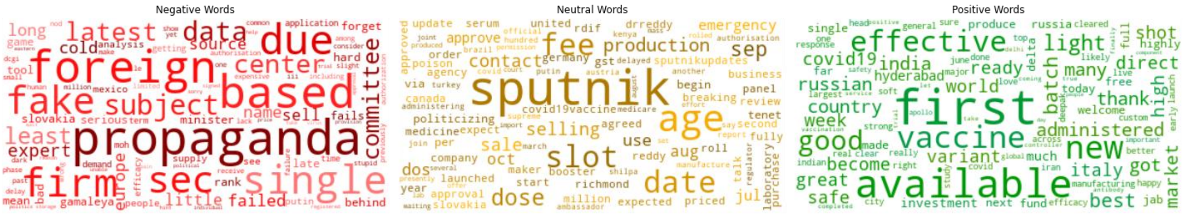

Negative Words
Neutral Words
Positive Words

Figure 15: Sputnik V Sentiment wordcloud with loglikelihood.

This study aimed to extract vaccine sentiments from Twitter datasets and analyse Indian vaccine tweets with three analysing parameters that are positive, neutral and negative. The language used was English so pre-processing using the NLTK library was smooth. Most of the sentiments were neutral, positive opinions weighed more than negative. Looking at the figure it is seen that around 33.1% of tweets were positive and 9.6% were negative sentiments obtained using TextBlob. Although the accuracy with which sentiments are predicted is not perfect, it does provide a good idea of how people feel about vaccination drives in India. Lexicon based is well suited for unlabelled data as useful insights were gained easily without much manual work. Although a machine learning approach could yield better results, manual labelling of the dataset is needed which is a time-consuming process. In a situation like a pandemic, it is very important to have a quick view of what's happening around the world and how people are reacting to it. So, here our approach will be useful and help companies and the government to tackle various issues related to the distribution of vaccines and immediate reactions.

In future, this project can be improved with more information rather than limited data available now as the process of vaccination is quite recent. Datasets can be manually labelled and a machine learning model can be created to predict with high accuracy and the trained model can be used to predict sentiments accurately if situations like this appear further.

Resources:

[1]. Gabriel Preda. (2021, February). COVID-19 All Vaccines Tweets. Version 103. Retrieved October 16, 2021 from https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets/version/103

[2]. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed on 16 October 2021)

[3]. Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review*: Vol. 1 : No. 4 , Article 7. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss4/7

[4]. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[5]. Esuli, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

[6]. Sarica S, Luo J (2021) Stopwords in technical language processing. PLoS ONE 16(8): e0254937. https://doi.org/10.1371/journal.pone.0254937

[7]. NLTK Corpora https://www.nltk.org/nltk_data/ (accessed on 16 October 2021)

[8]. Stefan Heid, Marcel Wever, & Eyke Hüllermeier. (2021). Reliable Part-of-Speech Tagging of Historical Corpora through Set-Valued Prediction.

[9]. Màrquez, L., Padró, L., & Rodríguez, H. (2000). A Machine Learning Approach to POS Tagging. Machine Learning, 39, 59-91.

[10]. Ostapenko, R. I. (2020). "Word cloud" as a graphical supplement to a scientific article (Editor-in-chief's column). Economic consultant, 32 (4), 4. doi: 10.46224/ecoc.2020.4.1

[11]. Comparing Word Form Counts. Available online: https://wordhoard.northwestern.edu/userman/analysis-comparewords.html#loglike (accessed on 16 October 2021).