# MINOR PROJECT-1

15B19CI591

PROJECT REPORT

## SENTIMENT ANALYSIS AND TOPIC MODELING USING TWEETS RELATED TO DIFFERENT COVID VACCINES

TEAM MEMBERS (**G131**)

| Enrollment Number | Name | Batch |
|---|---|---|
| 19103045 | **Himanshu Chauhan** | B10 |
| 19103046 | **Deepak Kumar Gupta** | B10 |
| 19103048 | **Ritik Rustagi** | B10 |

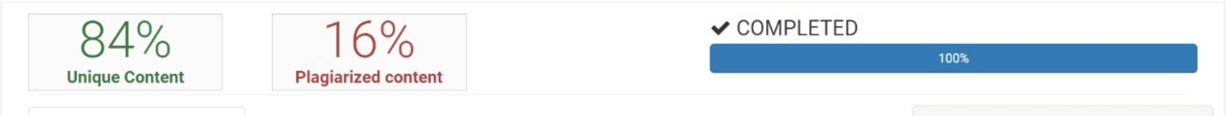
# SELF DECLARATION

We hereby declare the following usage of the open source code and prebuilt libraries in our minor project in 5th Semester with the consent of our supervisor.  We also measure the similarity percentage of pre written source code and our source code and the same is mentioned below. This measurement is true with the best of our knowledge and abilities.

1. List of pre build libraries
    a. Csv
    b. Tweepy
    c. Re
    d. Pandas
    e. Numpy
    f. String
    g. Nltl
    h. Warnings
    i. Os
    j. Matplotlib.pyplot
    k. Textblob
    l. pyLDAvis
    m. Gensim

2. List of pre-built features in libraries or in source code.
    a. pd.read_csv(”)
    b. porterStemmer()
    c. wordNetLemmatizer()
    d. tokenize(tweet)
    e. TextBlob(text).sentiment.subjectivity
    f. TextBlob(text).sentiment.polarity
    g. corpora.Dictionary()
    h. gensim.models.ldamodel.LdaModel()
    i. tweepy.OAuthHandler()
    j. tweepy.API()
    k. auth.et_access_token()
    l. csv.writer()
    m. writerow()
    n. tweepy.Cursor()

3. Percentage of pre written source code and source written by us.



84% Unique Content

16% Plagiarized content

✔ COMPLETED

100%

| Student ID | Student Name | Student signature |
|---|---|---|
| 19103045 | Himanshu Chauhan | HChauhan |
| 19103046 | Deepak Kumar Gupta | deepak |
| 19103048 | Ritik Rustagi | |

**Declaration by Supervisor (To be filled by Supervisor only)**

I, **JASPAL KAUR** (Name of Supervisor) declares that I above submitted a project with Titled **SENTIMENT ANALYSIS AND TOPIC MODELING USING TWEETS RELATED TO DIFFERENT COVID VACCINES**. was conducted under my supervision. The project is original and neither the project was copied from External sources nor it was submitted earlier in JIIT. I authenticate this project.

**Satisfactory work**

(Any Remarks by Supervisor)

Signature (Supervisor)

# ABSTRACT

The words we use to talk about the current epidemiological crisis on social media can inform us on how we are conceptualizing the pandemic and how we are reacting to its development. In recent years twitter has emerged an amazing database of thoughts of people globally. In this project we have crawled tweets from twitter related to different vaccines around the world and analysed them using sentiment analysis to see how positive people are towards different vaccines. To crawl data we used NodeXL and twitter API. And sentiment analysis was done via calculating subjectivity and polarity from the textblob library. After that we used Latent Dirichlet Allocation algorithm to do topic modelling.

The vaccines on which sentiment analysis was done are:-
Zykov-D,Pfizer,Moderna,Janssen,Covishield,Covexin,Corona-Vac,AstraZeneca

Among all these vaccines Covishield, Sputnik 5 and CoronaVac were found to have the highest positivity ratio of 85.22,83.61 and 82.38 respectively.

On doing Topic Modeling using LDA it was found that the top trending words are:-
Vaccine, Pfizer, Moderna , Omicron , Booster shots.
On looking at the words one can analyse that Pfizer and moderna vaccines were tweeted most number of times and Omicron and booster shots were hot topics during the time when data was crawled because of the outbreak of the new variant.

# INTRODUCTION

Covid-19 was officially reported for the first time by Chinese authorities as a virus that originated in the city of Wuhan, Hubei province in China, on 31 December 2019. According to official announcements from the World Health Organization (2020), while reviewing this manuscript the disease is contagious. has infected more than 267 million people worldwide, killing more than 5.2 million.

Although the recent release of various vaccines suggests that we may be facing the final stages of this health crisis, the effects of this long-term pandemic around the world will be seen beyond the actual end of the medical emergency and various aspects of our lives.

The pandemic put the world in a race to get vaccinated. But apart from logistical challenges the world had to face psychological challenges as well. Since the beginning  there have been different opinions on vaccines. In our project we have tried to do sentiment analysis of tweets from around the world on all major vaccines using textblob. We have tried to find out the trending topics related to vaccines using LDA. We crawled tweets from twitter on all major vaccines, pre-processed them and using sentiment analysis found out the subjectivity and polarity of tweets since the tweets are considered to be a good representation of the public opinion.

## RELATED WORK

| Year | Proposed Work Done | Performance Metrics | Dataset Used | Future Work |
|---|---|---|---|---|
| 2021 | A situation analysis in spain among BFHI maternity hospitals in spain | Response rate was 50% (58/116). Mothers suffered greater restrictions in the practices compared to women without COVID-19, with lower rates of companion of choice during labor (84% vs 100%; $p = 0.003$), skin-to-skin contact (32% vs 52%; $p = 0.04$), rooming-in (74% vs 98%; $p < 0.001$), companion of choice during hospital stay (68% vs 90%; $p = 0.006$), and breastfeeding support (78% vs 94%; $p = 0.02$). In COVID-19 mothers Practices were significantly less prevalent compared to normal situations. During delivery A lower accompaniment rate was observed in non-COVID-19 group .(24% vs 47.9%; $p < 0.01$). Hospitals with higher commitment to BFHI practices reported higher rates of skin-to-skin contact (45.2% vs 10.5%; $p = 0.01$) and rooming-in (83.9% vs 57.9%; $p < 0.05$) in COVID mothers. | cross-sectional survey conducted in May 2020 by using online questionnaires. Comparison of normal and pandemic situations and level of commitment to BFHI practices was performed. | To study the impact of COVID-19 on hospitals and learnings learned by hospitals committed to BFHI |
| 2021 | The information encoded in the short texts produced by private internet users on Twitter (the tweets) provides useful clues that in some cases can be used by experts. | Created 32 LDA models, each with a different number of topics. The evaluation of the Cv coherence measure revealed an elbow of the function for $N = 20$ . In addition to this, we selected a model that allowed for a broader analysis of the topics, hence a smaller number oftopics. Based on our previous experience with topic modeling and the coherence value function we selected $N = 12$, together with $N = 32$ for the fine-grained solution, and $N = 64$, our most fine-grained solution. | 1,698,254 tweets from individual users (without retweets), produced between 20.03.2020 and 01.07.2020. | Future research could explore public confidence in existing measures and policies, which are essential. |
| 2020 | To examine COVID-19–related discussions, concerns, and sentiments using tweets posted by Twitter users by LDA | Popular unigrams included “virus,” “lockdown,” and “quarantine.” Popular bigrams included “COVID-19,” “stay home,” “coronavirus,” “social distancing,” and “new cases.” We identified 13 discussion topics and categorized them into 5 different themes: (1) public health measures to slow the spread of COVID-19, (2) social stigma associated with COVID-19, (3) COVID-19 news, cases, and deaths, (4) COVID-19 in the United States, and (5) COVID-19 in the rest of the world | We analyzed 4 million Twitter messages related to the COVID-19 pandemic using a list of 20 hashtags (eg, “coronavirus,” “COVID-19,” “quarantine”) from March 7 to April 21, 2020 | Future studies should further investigate sentiments by examining specific countries and expanding the scope to include other media platforms such as Facebook and Weibo |
| 2020 | This study aimed to examine worldwide trends of four emotions—fear, anger, sadness, and joy—and the narratives underlying those emotions during the COVID-19 pandemic. | public emotions shifted strongly from fear to anger over the course of the pandemic, while sadness and joy also surfaced. Findings from word clouds suggest that fears around shortages of COVID-19 tests and medical supplies became increasingly widespread discussion points. Anger shifted from xenophobia at the beginning of the pandemic to discourse around the stay-at-home notices. Sadness was highlighted by the topics of losing friends and | Over 20 million social media twitter posts made during the early phases of the COVID-19 outbreak from January 28 to April 9, 2020, were collected using “wuhan,” “corona,” “nCov,” and “covid” as search keywords. | Future studies should expanding the scope to include other media platforms such as Facebook and Weibo |

| | | | | |
|---|---|---|---|---|
| | | family members, while topics related to joy included words of gratitude and good health. | | |
| 2020 | We performed content analysis to categorize tweets into appropriate themes and analyzed associated Twitter data. | Eight out of nine (88.9%) G7 world leaders had verified and active Twitter accounts, with a total following of 85.7 million users. Out of a total 203 viral tweets, 166 (82.8%) were classified as 'Informative', of which 48 (28.6%) had weblinks to government-based sources, while 19 (9.4%) were 'Morale-boosting' and 14 (6.9%) were 'Political'. Numbers of followers and viral tweets were not strictly related. | This was a qualitative study with content analysis. Inclusion criteria were as follows: viral tweets from G7 world leaders, attracting a minimum of 500 'likes'; keywords 'COVID-19' or 'coronavirus'; search dates 17 November 2019 to 17 March 2020. | Future work can do in-depth analysis of the reach of these tweets as it was beyond the remit of this study |
| 2020 | To study the propagation of conspiracy theories on social media, we examine the case of #FilmYourHospital | To understand how this campaign manifested on Twitter, we examined the formation of this network over time. displays #FilmYourHospital interactions posted during the first three days of the campaign. Notably, one of the most influential users who triggered the viral spread of this misinformation campaign was @DeAnna4Congress, a verified account for DeAnna Lorraine, a former Republican Congressional candidate who recently ran against Nancy Pelosi for the U.S. House California District 12. | used Netlytic (Gruzd, 2016a) to collect and analyze data, Gephi (Bastian et al., 2009) to visualize the resulting communication network over time, the python library Twarc (Ruest and Milligan, 2016) to check if accounts had been deleted or suspended by Twitter, and the Barometer API (Yang et al., 2019) to assess if an account is automated (exhibiting a bot-like behavior). | future studies should investigate automated cam[aigns with more hashtags and accounts as the research had a relatively smaller number of accounts considered |
| 2020 | retrieved tweets using COVID-19-related keywords, and performed semiautomatic filtering to curate self-reports of positive-tested users. | We identified 203 positive-tested users who reported 1002 symptoms using 668 unique expressions. The most frequently-reported symptoms were fever/pyrexia (66.1%), cough (57.9%), body ache/pain (42.7%), fatigue (42.1%), headache (37.4%), and dyspnea (36.3%) amongst users who reported at least 1 symptom. Mild symptoms, such as anosmia (28.7%) and ageusia (28.1%), were frequently reported on Twitter, but not in clinical studies. | Using keywords social media, communicable diseases, virus diseases, natural language processing, text mining | Future studies could focus upon the variation in symptoms of the virus overtime as this study was conducted in a specific time period. |
| 2020 | Few infodemiology studies have applied network analysis in conjunction with content analysis. This study investigates information transmission networks and news-sharing behaviors regarding COVID-19 on | The network analysis suggests that the spread of information was faster in the Coronavirus network than in the other networks (Corona19, Shinchon, and Daegu). People who used the word "Coronavirus" communicated more frequently with each other. The spread of information was faster, and the diameter value was lower than for those who used other terms. Many of the news items | Korean COVID-19-related Twitter data were collected on February 29, 2020. Our final sample comprised 43,832 users and 78,233 relationships on Twitter. We generated four networks in terms of key | They applied multi-coder methods, but two coders may not be sufficient to ensure reliability. Future studies could apply more coders to guarantee the |

| | | | | |
|---|---|---|---|---|
| | Twitter in Korea. | highlighted the positive roles being played by individuals and groups, directing readers' attention to the crisis. Ethical issues such as deviant behavior among the population and an entertainment frame highlighting celebrity donations also emerged often. There was a significant difference in the use of nonportal (n=14) and portal news (n=26) sites between the four network types. The news frames used in the top sources were similar across the networks (*P*=.89, 95% CI 0.004-0.006). Tweets containing medically framed news articles (mean 7.571, SD 1.988) were found to be more popular than tweets that included news articles adopting nonmedical frames (mean 5.060, SD 2.904; N=40, *P*=.03, 95% CI 0.169-4.852). | issues regarding COVID-19 in Korea. | reliability of the analysis results |
| 2020 | The proposed system aims to fetch tweets from dataset and after applying prepossessing classify into positive and negative tweets and further classify into six emotions using lexical-oriented methods using R. | Using our emotional analysis method the tweets related to COVID-19 have been classified into basic emotion categories3 The number of positive and negative tweets is almost equal. Another classification of emotion shows more tweets are related to trust which shows that people have a lot of positive trust to fight against COVID-19 and policies taken by authorities across the globe. Another majority of tweets shows fear, as the number of patients with COVID-19 is continuously increasing and is spreading at a fast pace across the globe. Till today no proper vaccine and treatment is available so many people have expressed fear. Quite a few tweets also show sadness. | Tweets can be fetched in Real Time also by using TwitterR used in R which provides access to Twitter API. For experimentation purposes, the Dataset used includes Tweets related to COVID-19 between 22nd January 2020 and 15th April 2020 used for classification and is downloaded from TweetBinder website . Data collected consists of around 30,000 tweets in the English language from all over the world | closely view the social media content, many people are also using various emojis and graphics to convey their emotions. Further the system can be extended to consider such graphics and emojis in analyzing the sentiments |
| 2020 | The aim of this study was to increase understanding of public awareness of COVID-19 pandemic trends and uncover meaningful themes of concern posted by Twitter users in the English language during the pandemic. | The results indicate three main aspects of public awareness and concern regarding the COVID-19 pandemic. First, the trend of the spread and symptoms of COVID-19 can be divided into three stages. Second, the results of the sentiment analysis showed that people have a negative outlook toward COVID-19. Third, based on topic modeling, the themes relating to COVID-19 and the outbreak were divided into three categories: the COVID-19 pandemic emergency, how to control COVID-19, and reports on COVID-19. | Data mining was conducted on Twitter to collect a total of 107,990 tweets related to COVID-19 between December 13 and March 9, 2020. | This Twitter data analysis can be used to explain the public awareness and perception of the COVID-19 pandemic. Based on public awareness, the data were divided into three main stages in relation to the timeline of the epidemic. The early or incubation stage (Stage 1) was the phase in which the severity and the spread of COVID-19 began to increase. |

# PROPOSED WORK

## 1. DATA COLLECTION

The data was collected from twitter using the Nodexl application and threw twitter api using python library tweepy.We collected data from 11 different vaccine hashtags(#).Thus we collected a total of 74,796 tweets.

Twitter API

Create an authorization that accesses my developer account using consumer_key, consumer_secret_key,access_token,access_token _secret.

We can pull  last week's data using this algorithm,then we write the header of the spreadsheet and for each tweet matching our hashtags, write relevant info to the spreadsheet.

Using  libraries-

- Csv
- Tweepy
- re

## VACCINES

**AstraZeneca**: WHO has declared  two versions of AstraZeneca for emergency use.On 19 April 2021 WHO globally advisory committee committed vaccine safety statement.

**Booster shots:** Booster shot, vaccines are working well to prevent severe illness, hospitalization, and death, even if it reduces the probability of death.

**CoronaVac:** This vaccine is approved for use in China and authorized for emergency use in more than a dozen other countries. In a trial in Brazil, researchers found it had an efficacy against infections of 50.65%.

**Covaxin**: Covexin, produced  by Bharat Biotech is developed in collaboration with the Indian Council of Medical Research (ICMR) - National Institute of Virology.

**Covishield**: It is a recombinant, replication-deficient chimpanzee adenovirus vector encoding the SARS-CoV-2 Spike (S) glycoprotein.

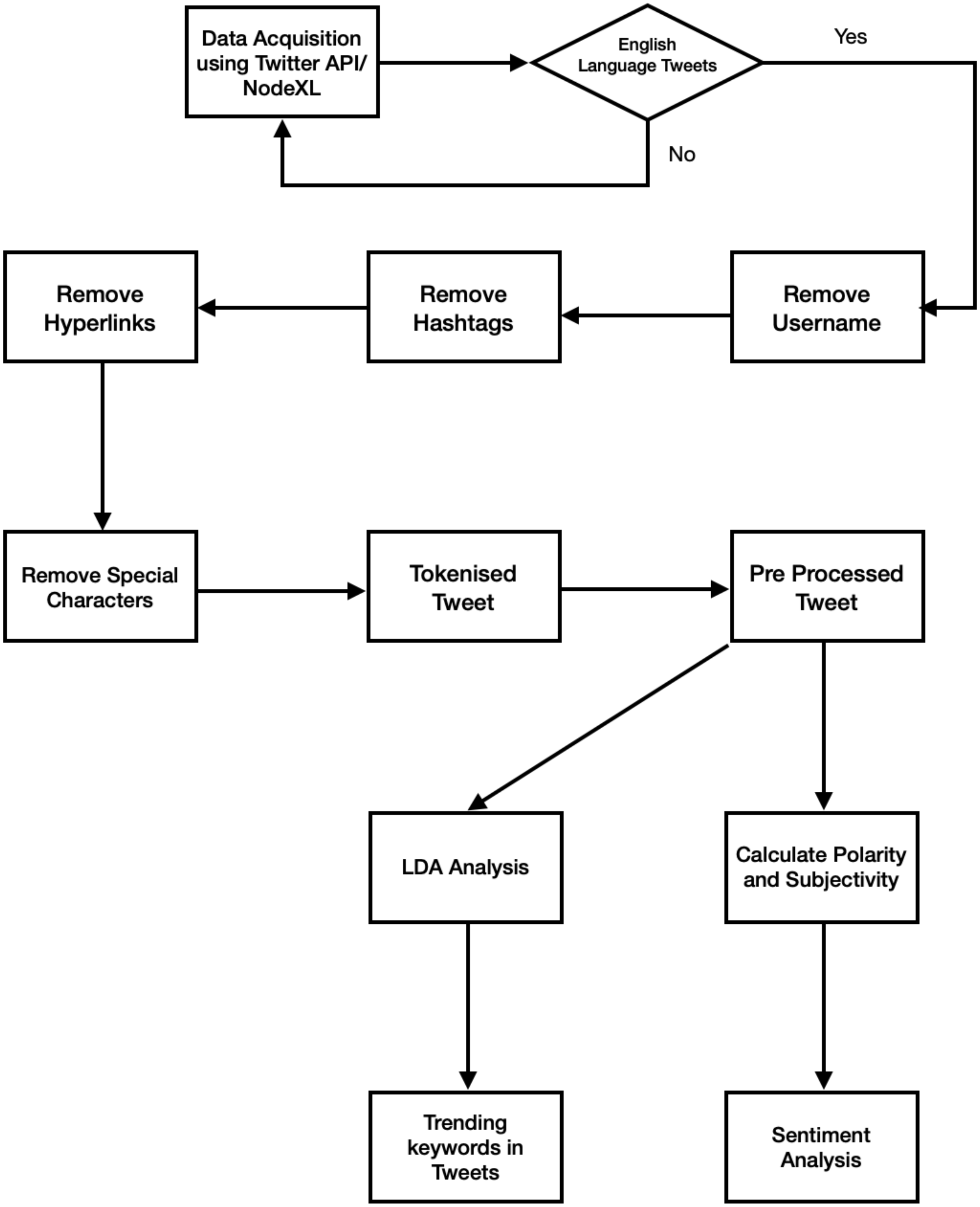
**Janssen**: Janssen Vaccine was averagely effective(63.3%) in clinical trials at preventing laboratory-confirmed COVID-19 infection in people who received the vaccine and had no evidence of being previously infected.

**Moderna**: Moderna, a Massachusetts-based vaccine developer, partnered with the National Institutes of Health to develop and test a coronavirus vaccine known as mRNA-1273.In a clinical trial result that the vaccine has more than 90% preventing rate against covid-19.

**Pfizer**: The vaccine works by introducing a molecule to cells, known as messenger RNA (mRNA). This molecule tells cells to make a protein from the virus COVID-19, which is SARS-CoV-2, by which this protein body triggers an immune response. This creates antibodies, as well as longer lasting immunity that fights  SARS-CoV-2 infections.

**Sputnik-V**: The Sputnik V COVID-19 vaccine uses two harmless viruses that deliver the genetic code for our cells to make a protein from the new coronavirus.

## FLOW CHART


Data Acquisition using Twitter API/ NodeXL
English Language Tweets
Yes
No
Remove Username
Remove Hashtags
Remove Hyperlinks
Remove Special Characters
Tokenised Tweet
Pre Processed Tweet
LDA Analysis
Calculate Polarity and Subjectivity
Trending keywords in Tweets
Sentiment Analysis

## RECORD OF CRAWLED TWEETS

| Vaccination Name | HashTag | Date from-to | No of tweets |
|---|---|---|---|
| AstraZeneca | #AstraZeneca | 15-11-2021 to 04-12-2021 | 8600 |
| Booster shots | #Booster shots | 15-11-2021 to 04-12-2021 | 6845 |
| CoronaVac | #CoronaVac | 15-11-2021 to 04-12-2021 | 4751 |
| Covexin | #Covexin | 15-11-2021 to 04-12-2021 | 17200 |
| Covishield | #Covishield | 15-11-2021 to 04-12-2021 | 2200 |
| Janssen | #Janssen | 15-11-2021 to 04-12-2021 | 2700 |
| Moderna | #Moderna | 15-11-2021 to 04-12-2021 | 6800 |
| Pfizer | #Pfizer | 15-11-2021 to 04-12-2021 | 14500 |
| Zykov-D | #Zykov-D | 15-11-2021 to 04-12-2021 | 6800 |
| **Total no of tweets** | | | **68200** |

## 2. DATA PREPROCESSING

https://drive.google.com/file/d/1SMiu7iKh9CL7laGnj0otYazJ-mODj5Aq/view?usp=sharing

- **Libraries used**-

  - **Pandas:** For Data Manipulation and Analysis
  - **Numpy:** To perform mathematical operations on Array
  - **String:** To process standard python strings
  - **Nltl:** Stands for Natural Language ToolKit, for natural language processing
  - **Re:** Stands for Regular Language, to find different pattern in tweets
  - **Warnings:** To warn the developer of unnecessary exception
  - **Os:** For changing the current working directory

- **PreDefined Functions-**

  - **pd.read_csv("Pfizer.csv"):** To read the csv file of raw tweets
  - **porterStemmer():** To find the root word of the tweet
  - **wordNetLemmatizer():** similar to porterStemmer() but it brings context to words
  - **tokenize(tweet):** To break the tweet in individual words

- **UserDefined Function-**

  - **load_data():** To load the desired csv file of raw tweets
  - **remove_pattern(input_txt,pattern):** For removing the @username
  - **clean_tweets(tweet):** It performs multiple functions-
    - Remove the RT keyword from tweet
    - Remove the hyperlinks
    - Remove coma
    - Remove numbers
    - Tokenize Tweet
  - **remove_punct(text):** To remove the ',' symbol from tokenized tweet

➔ rawTweets

➔ preProcessedTweets

## 3. SENTIMENT ANALYSIS

https://drive.google.com/file/d/1SMiu7iKh9CL7laGnj0otYazJ-mODj5Aq/view?usp=sharing

- **Libraries used-**

    - **Matplotlib.pyplot:** For graph plotting
    - **Textblob:** To get the subjectivity and polarity of preprocessed tweets

- **PreDefined Functions-**

    - **TextBlob(text).sentiment.subjectivity:** To find the subjective score of the tweet ranging from [0,1]
    - **TextBlob(text).sentiment.polarity:** Find the polarity of each word of tweet and sum them overall
        - Negative: polarity score < 0
        - Neutral: polarity score = 0
        - Positive: polarity score > 0

- **UserDefined Function-**

    - **getSubjectivity:** To get the subjectivity value of individual tweet
    - **getPolarity:** To get the polarity value of individual tweet
    - **getAnalysis(score):** For classifying the tweets as Negative, Positive or Neutral based on their polarity score

## 4. TOP TRENDING TOPICS

https://drive.google.com/file/d/1lh7VQPM1dvPeQv3z5DA4ZsPAlhoRlRkw/view?usp=sharing

- **Libraries used**:
  - **Pandas:** Pandas is a software library written for the Python programming language to manipulate and analyze data. In particular, it provides data structures and functions to manipulate numerical tables and time series.
  - **Numpy:** NumPy is a Python programming language library, which adds support for large, multi-sided arrays and matrices, as well as a large collection of advanced mathematical functions to work on these components.
  - **pyLDAvis:** Python library for interactive topic model visualization.
  - **Nltk:** stands for natural language toolkit, it is used for tokenizing the data.
  - **Gensim:** Gensim is an open source modeling library and natural language processing, using a modern mathematical learning machine.

## HOW DOES LDA WORK?

Topic Modeling refers to the task of identifying titles that best describe the collection of documents. These titles will only appear during the title modeling process (hence the so-called latent). And one popular method of title matching is known as the Latent Dirichlet Allocation (LDA). Although the word is oral, the meaning of this is very simple.

In short, LDA is considering a consistent set of topics. Each topic represents a set of words. And the LDA's goal is to map all the texts to the subheadings in a way, so that the words in each text are taken seriously by the subject being considered. We will go through this process in such a way that at the end of it you will be comfortable enough to use this method yourself.

# EXPERIMENTAL RESULTS AND DISCUSSION

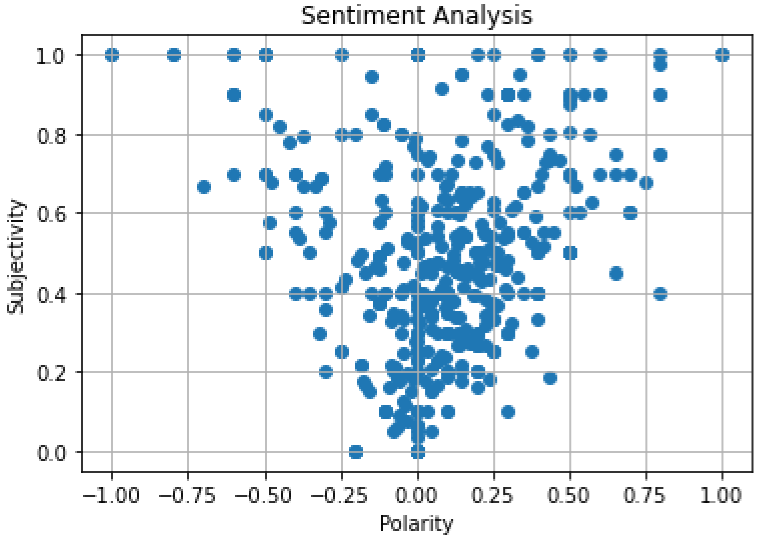
## 1. SENTIMENT ANALYSIS

### Graph

a) Polarity vs Subjectivity

b) Sentiment vs Count

- **Astra Zeneca**



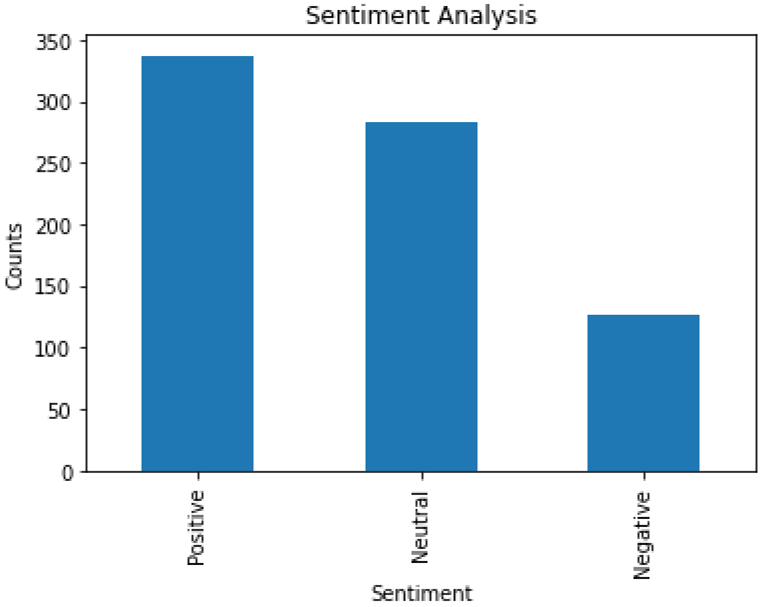
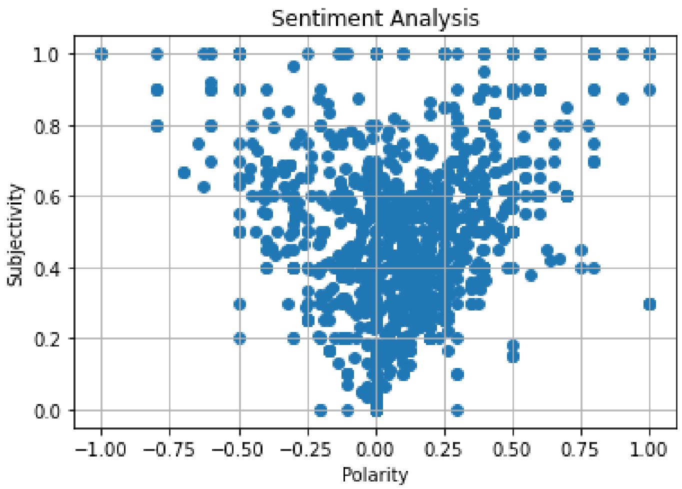


Fig 1.a

Fig 1.b

**Percentage of Positive tweets : 45.1 %**
**Percentage of Negative tweets : 17.0 %**

- **Booster Shots**



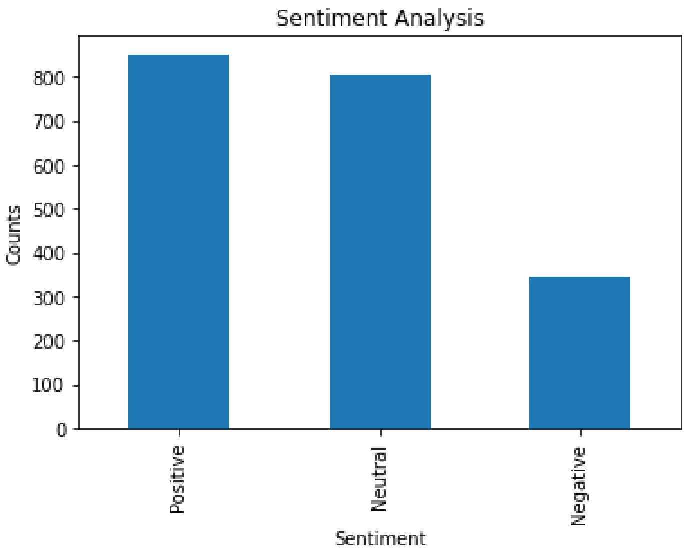
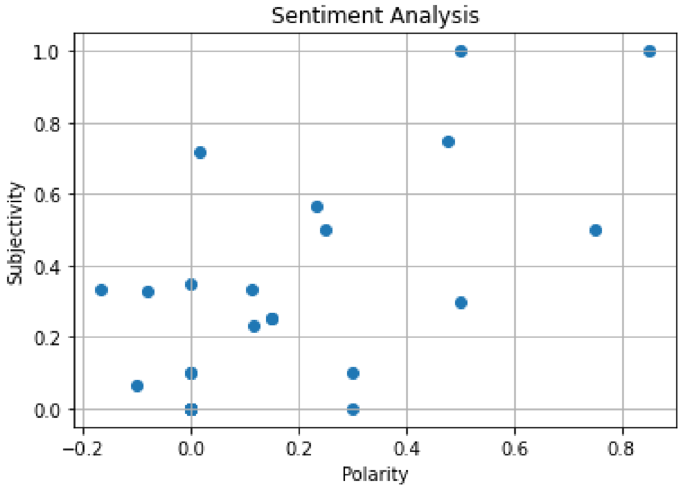


Fig 2.a

Fig 2.b

**Percentage of Positive tweets : 42.5 %**
**Percentage of Negative tweets : 17.2 %**

- **CoronaVac**



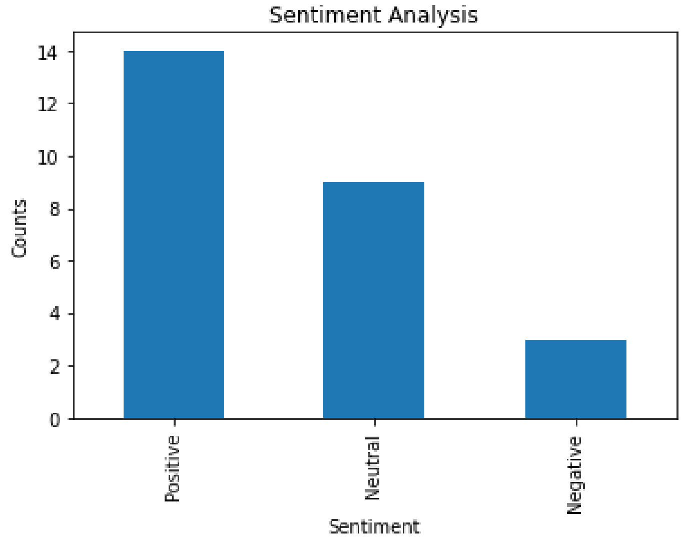


Fig 3.a

Fig 3.b

**Percentage of Positive tweets : 53.8 %**
**Percentage of Negative tweets : 11.5 %**

- **Covaxin**

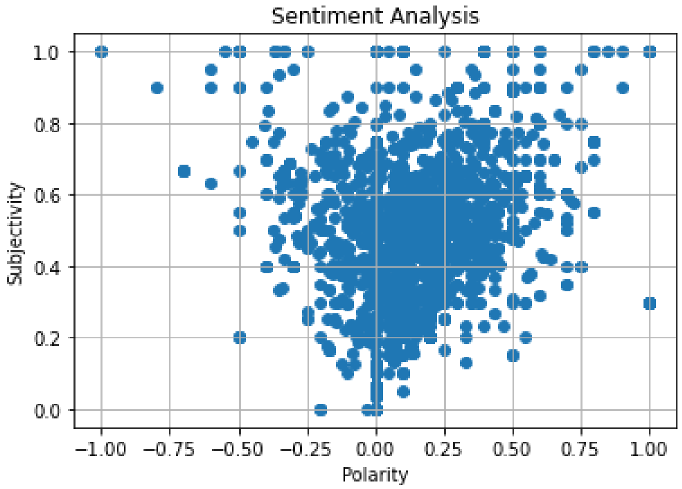


Fig 4.a

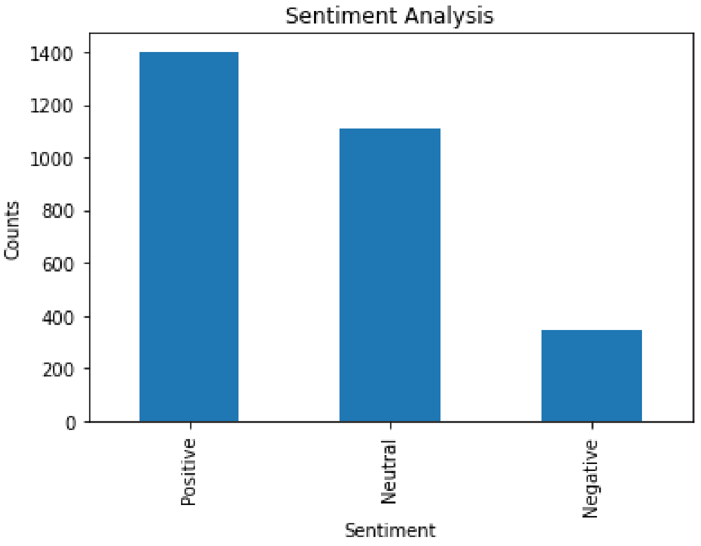


Fig 4.b

**Percentage of Positive tweets : 49.1 %**
**Percentage of Negative tweets : 12.0 %**

- **Covishield**

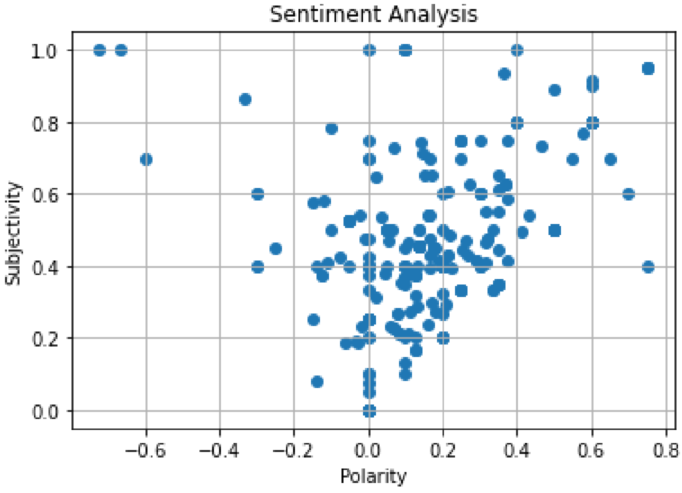


Fig 5.a

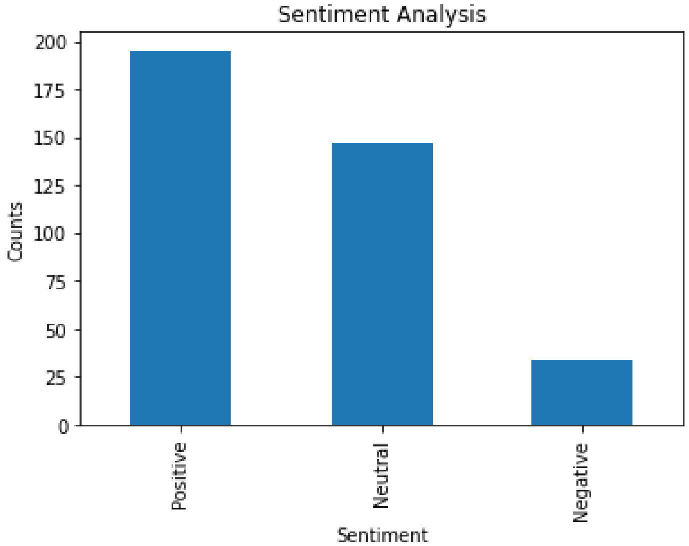


Fig 5.b

**Percentage of Positive tweets : 51.9 %**
**Percentage of Negative tweets : 9.0 %**

- **Janssen**

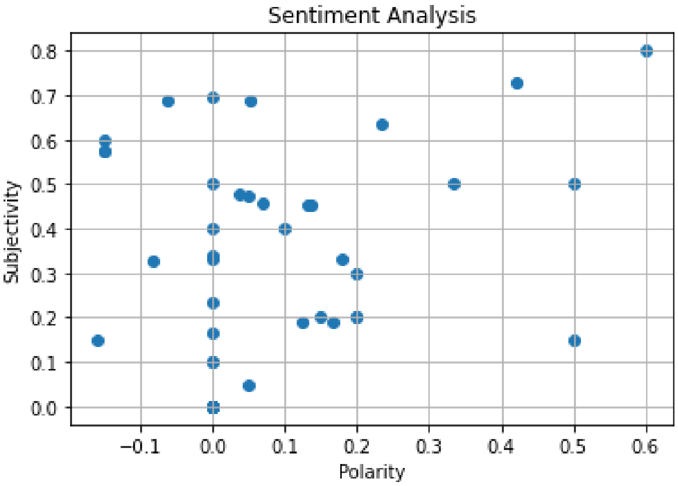


Fig 6.a

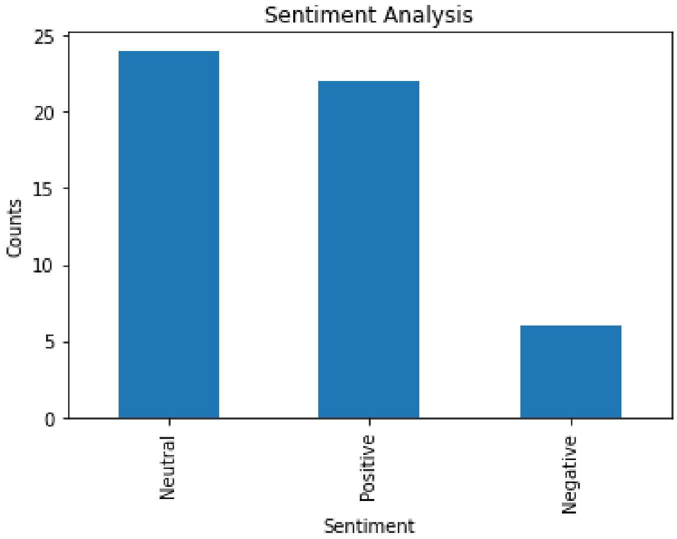


Fig 6.b

**Percentage of Positive tweets : 42.3 %**
**Percentage of Negative tweets : 11.5 %**

- **Moderna**

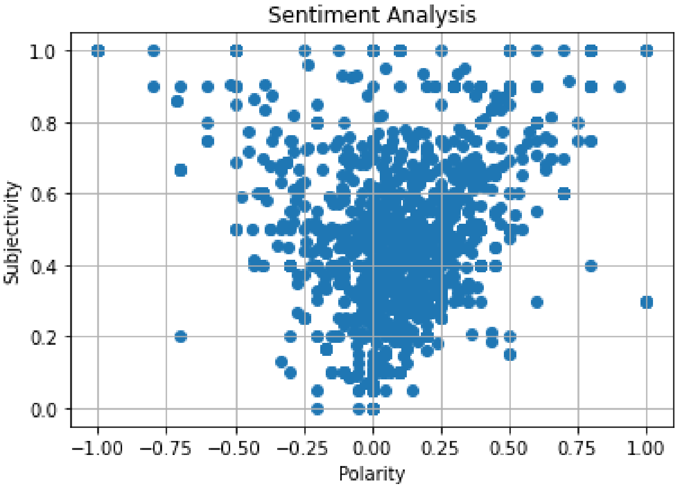


Fig 7.a

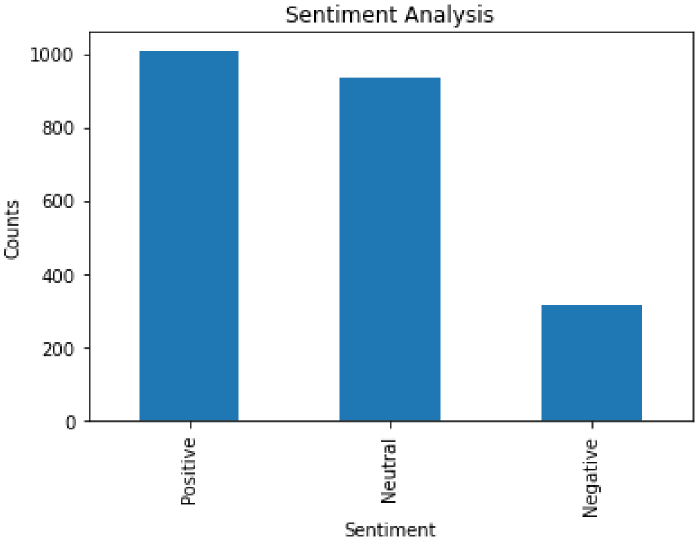
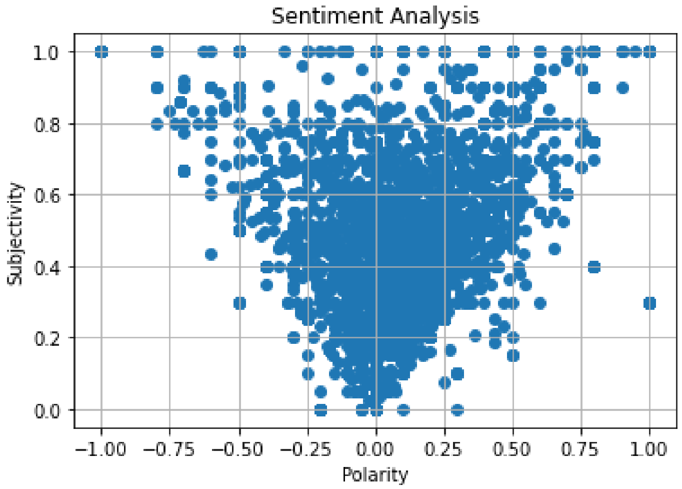


Fig 7.b

**Percentage of Positive tweets : 44.7 %**
**Percentage of Negative tweets : 14.0 %**

- **Pfizer**



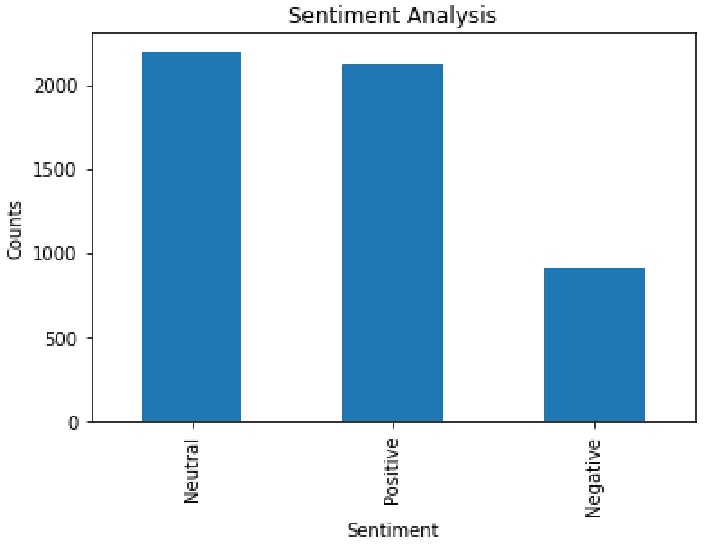
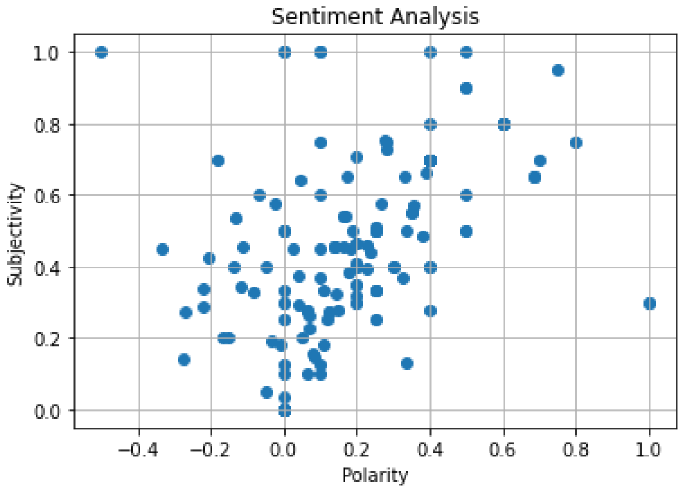


Fig 8.a

Fig 8.b

**Percentage of Positive tweets : 40.6 %**
**Percentage of Negative tweets : 17.4 %**

- **Sputnik-V**



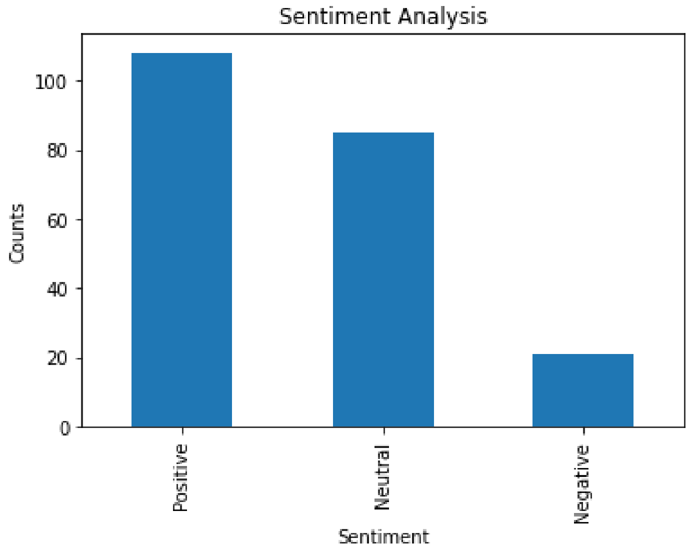


Fig 9.a

Fig 9.b

**Percentage of positive tweets : 50.5 %**
**Percentage of Negative tweets : 9.8 %**

## 2. TOPIC MODELING

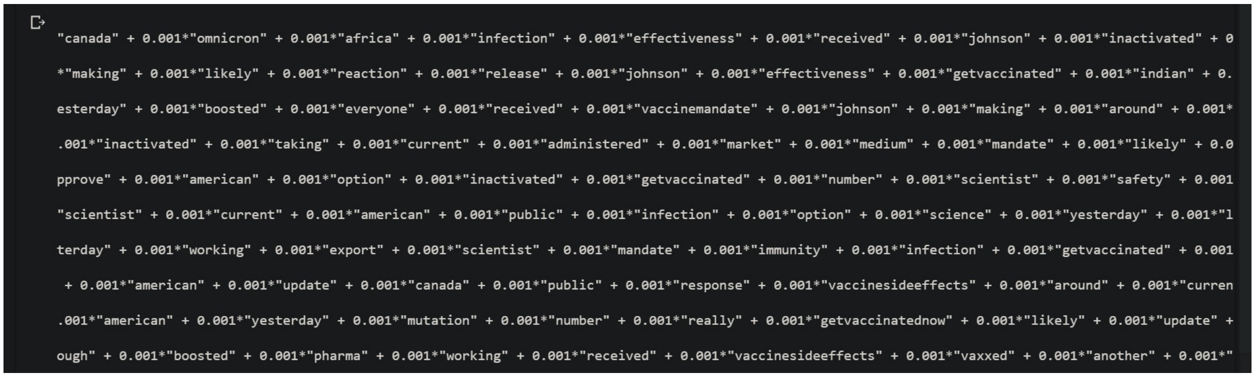
LDA was applied on all tweets of all vaccines individually in order to characterize the tweets into 10 topics and then the perplexity of the resultant topics was noted. It was found that the perplexity of the tweets of vaccines had a small variance.

| VACCINE | PERPLEXITY |
|---|---|
| Astra-zenica | -8.099867285842116 |
| CORONA VAC | -7.651921815635716 |
| COVEXIN | -5.346398360829340 |
| COVISHIELD | -7.707547288424992 |
| JANSSEN | -8.701069847912026 |
| MODERNA | -7.406033452180177 |
| PFIZER | -7.863407843028444 |
| Sputnik | -6.377221925936096 |

| Sno | Top 10 words from topic 1 | Top 10 words from topic 2 | Top 10 words from topic 3 | Top 10 words from topic 4 | Top 10 words from topic 5 |
|---|---|---|---|---|---|
| 1 AstraZeneca | '0.068*"astrazeneca" + 0.039*"vaccine" + 0.021*"pfizer" + 0.016*"booster" + 0.015*"moderna" + 0.006*"omicron" + 0.005*"covidvaccine" + 0.005*"trigger" + 0.005*"people" + 0.004*"vaccination"') | '0.070*"astrazeneca" + 0.039*"vaccine" + 0.024*"pfizer" + 0.018*"moderna" + 0.013*"booster" + 0.007*"omicron" + 0.006*"people" + 0.005*"vaccination" + 0.005*"variant" + 0.004*"effect"' | '0.037*"astrazeneca" + 0.022*"vaccine" + 0.013*"moderna" + 0.013*"pfizer" + 0.009*"booster" + 0.006*"omicron" + 0.004*"people" + 0.004*"trigger" + 0.004*"variant" + 0.003*"covidvaccine"' | '0.075*"astrazeneca" + 0.041*"vaccine" + 0.023*"pfizer" + 0.016*"booster" + 0.014*"moderna" + 0.007*"people" + 0.006*"omicron" + 0.006*"variant" + 0.006*"trigger" + 0.005*"vaccination"' | 0.052*"astrazeneca" + 0.036*"vaccine" + 0.021*"pfizer" + 0.011*"moderna" + 0.009*"booster" + 0.006*"trigger" + 0.005*"omicron" + 0.004*"covidvaccine" + 0.004*"vaccination" + 0.004*"variant"' |

| | | | | | |
|---|---|---|---|---|---|
| 2<br>Booster Shots | '0.047*"booster shots" + 0.031*"booster" + 0.018*"vaccine" + 0.011*"omicron" + 0.010*"vaccinated" + 0.010*"people" + 0.008*"variant" + 0.006*"vaccination" + 0.005*"omicronvariant" + 0.005*"pfizer"' | '0.041*"boostershots" + 0.030*"booster" + 0.026*"vaccine" + 0.009*"omicron" + 0.009*"variant" + 0.008*"vaccination" + 0.007*"people" + 0.006*"vaccinated" + 0.005*"covidvaccine" + 0.005*"omicronvariant"' | '0.083*"boostershots" + 0.039*"booster" + 0.022*"vaccine" + 0.015*"omicron" + 0.010*"vaccinated" + 0.009*"people" + 0.007*"covidvaccine" + 0.007*"variant" + 0.006*"omicronvariant" + 0.006*"covidvariant"' | '0.030*"boostershots" + 0.024*"booster" + 0.016*"vaccine" + 0.010*"omicron" + 0.008*"people" + 0.007*"variant" + 0.007*"vaccinated" + 0.005*"vaccination" + 0.005*"covidvaccine" + 0.005*"pfizer"' | '0.045*"boostershots" + 0.027*"booster" + 0.014*"vaccine" + 0.013*"omicron" + 0.010*"variant" + 0.007*"vaccination" + 0.006*"people" + 0.006*"vaccinated" + 0.006*"pfizer" + 0.005*"omicronvariant"' |
| 3<br>CoronaVac | '0.012*"coronavac" + 0.009*"sinopharm" + 0.009*"vaccine" + 0.008*"sinovac" + 0.007*"omicron" + 0.007*"contest" + 0.007*"astrazeneca" + 0.006*"billion" + 0.006*"thursdaythoughts" + 0.006*"nobody"' | '0.048*"coronavac" + 0.018*"vaccine" + 0.015*"sinopharm" + 0.012*"sinovac" + 0.011*"nobody" + 0.011*"omicron" + 0.011*"thursdaythoughts" + 0.008*"astrazeneca" + 0.008*"vaccinated" + 0.008*"pfizer"' | '0.017*"coronavac" + 0.013*"vaccine" + 0.011*"sinopharm" + 0.008*"nobody" + 0.008*"sinovac" + 0.008*"contest" + 0.008*"omicron" + 0.008*"thursdaythoughts" + 0.007*"astrazeneca" + 0.007*"growthmindset"' | '0.031*"coronavac" + 0.014*"vaccine" + 0.011*"sinopharm" + 0.010*"omicron" + 0.009*"astrazeneca" + 0.009*"contest" + 0.009*"sinovac" + 0.008*"thursdaythoughts" + 0.007*"nobody" + 0.007*"sputnikv"' | '0.065*"coronavac" + 0.020*"vaccine" + 0.013*"nobody" + 0.013*"sinopharm" + 0.012*"omicron" + 0.012*"sinovac" + 0.009*"contest" + 0.009*"astrazeneca" + 0.007*"covexin" + 0.007*"thursdaythoughts"' |
| 4<br>covexin | '0.082*"covexin" + 0.045*"vaccine" + 0.012*"variant" + 0.010*"effective" + 0.009*"omicron" + 0.009*"traditional" + 0.008*"approved" + 0.007*"approve" + 0.007*"pfizer" + 0.006*"approval"'), (1, | '0.076*"covexin" + 0.034*"vaccine" + 0.008*"effective" + 0.008*"approved" + 0.008*"variant" + 0.007*"approve" + 0.006*"omicron" + 0.006*"traditional" + 0.006*"inactivated" + 0.005*"people"' | 0.050*"covexin" + 0.059*"vaccine" + 0.011*"variant" + 0.011*"traditional" + 0.009*"effective" + 0.009*"approved" + 0.009*"people" + 0.008*"approve" + 0.008*"pfizer" + 0.007*"omicron"' | '0.050*"covexin" + 0.018*"vaccine" + 0.007*"approved" + 0.006*"effective" + 0.005*"variant" + 0.005*"approve" + 0.005*"traditional" + 0.004*"people" + 0.004*"pfizer" + 0.004*"approval"' | '0.055*"covexin" + 0.029*"vaccine" + 0.010*"variant" + 0.007*"approved" + 0.007*"traditional" + 0.006*"omicron" + 0.006*"inactivated" + 0.006*"effective" + 0.006*"approve" + 0.005*"pfizer"' |
| 5<br>Covisheild | '0.075*"covishield" + 0.036*"vaccine" + 0.019*"covaxin" + 0.017*"booster" + 0.014*"vaccinated" + 0.012*"omicron" + 0.010*"effective" + 0.010*"variant" + 0.009*"approval" + 0.008*"effectiveness"' | '0.063*"covishield" + 0.037*"vaccine" + 0.019*"covaxin" + 0.017*"booster" + 0.011*"omicron" + 0.010*"variant" + 0.010*"vaccinated" + 0.009*"approval" + 0.008*"vaccination" + 0.007*"institute"' | '0.061*"covishield" + 0.044*"vaccine" + 0.019*"covexin" + 0.017*"booster" + 0.014*"omicron" + 0.012*"vaccinated" + 0.011*"approval" + 0.010*"variant" + 0.009*"institute" + 0.008*"effective"' | '0.060*"covishield" + 0.024*"vaccine" + 0.017*"covexin" + 0.014*"booster" + 0.011*"omicron" + 0.011*"vaccinated" + 0.010*"approval" + 0.009*"variant" + 0.009*"institute" + 0.007*"effective"' | '0.070*"covishield" + 0.034*"vaccine" + 0.019*"covexin" + 0.015*"booster" + 0.013*"omicron" + 0.009*"variant" + 0.009*"vaccinated" + 0.008*"approval" + 0.007*"vaccination" + 0.007*"effective"' |

| | | | | | |
|---|---|---|---|---|---|
| 6<br>Janseen | '0.033*"janssen" + 0.029*"vaccine" + 0.014*"booster" + 0.010*"astrazeneca" + 0.009*"pfizer" + 0.008*"moderna" + 0.008*"johnson" + 0.005*"vaccination" + 0.005*"pharma" + 0.005*"searching"' | '0.046*"janssen" + 0.042*"vaccine" + 0.020*"booster" + 0.013*"johnson" + 0.013*"moderna" + 0.012*"astrazeneca" + 0.009*"pfizer" + 0.007*"leiden" + 0.007*"pharma" + 0.006*"received"' | '0.043*"janssen" + 0.038*"vaccine" + 0.015*"booster" + 0.015*"moderna" + 0.014*"pfizer" + 0.014*"johnson" + 0.012*"astrazeneca" + 0.006*"people" + 0.006*"netherlands" + 0.006*"received"' | '0.040*"janssen" + 0.031*"vaccine" + 0.020*"moderna" + 0.014*"booster" + 0.012*"pfizer" + 0.010*"astrazeneca" + 0.009*"johnson" + 0.006*"searching" + 0.006*"pharma" + 0.006*"netherlands"' | '0.044*"janssen" + 0.033*"vaccine" + 0.019*"booster" + 0.015*"astrazeneca" + 0.010*"pfizer" + 0.010*"johnson" + 0.010*"moderna" + 0.007*"vaccination" + 0.006*"covidvaccine" + 0.006*"searching"' |
| 7<br>Moderna | '0.058*"moderna" + 0.028*"vaccine" + 0.025*"pfizer" + 0.024*"booster" + 0.021*"omicron" + 0.011*"variant" + 0.006*"effective" + 0.005*"vaccinated" + 0.004*"omicronvariant" + 0.004*"astrazeneca"') | '0.021*"moderna" + 0.016*"vaccine" + 0.010*"omicron" + 0.009*"booster" + 0.008*"pfizer" + 0.005*"variant" + 0.003*"effective" + 0.002*"vaccinated" + 0.002*"omicronvariant" + 0.002*"people" | '0.086*"moderna" + 0.053*"vaccine" + 0.030*"pfizer" + 0.026*"booster" + 0.021*"omicron" + 0.016*"variant" + 0.007*"effective" + 0.007*"astrazeneca" + 0.004*"coronavirus" + 0.004*"vaccinated"' | '0.059*"moderna" + 0.047*"vaccine" + 0.027*"omicron" + 0.024*"booster" + 0.017*"pfizer" + 0.009*"variant" + 0.006*"effective" + 0.006*"omicronvariant" + 0.004*"covidvaccine" + 0.004*"people"' | '0.032*"moderna" + 0.019*"vaccine" + 0.014*"booster" + 0.014*"omicron" + 0.012*"pfizer" + 0.007*"variant" + 0.004*"effective" + 0.003*"astrazeneca" + 0.003*"vaccination" + 0.002*"omicronvariant"' |
| 8<br>Pfizer | '0.068*"pfizer" + 0.033*"vaccine" + 0.016*"booster" + 0.013*"moderna" + 0.007*"omicron" + 0.007*"variant" + 0.005*"people" + 0.004*"astrazeneca" + 0.004*"vaccination" + 0.003*"covidvaccine"' | '0.094*"pfizer" + 0.041*"vaccine" + 0.018*"moderna" + 0.016*"booster" + 0.008*"omicron" + 0.008*"vaccinated" + 0.007*"variant" + 0.006*"vaccination" + 0.004*"people" + 0.004*"astrazeneca"' | '0.060*"pfizer" + 0.045*"vaccine" + 0.014*"booster" + 0.013*"moderna" + 0.007*"omicron" + 0.007*"people" + 0.006*"variant" + 0.005*"vaccinated" + 0.005*"astrazeneca" + 0.004*"vaccination"' | '0.035*"pfizer" + 0.020*"vaccine" + 0.013*"booster" + 0.007*"moderna" + 0.006*"omicron" + 0.005*"variant" + 0.004*"people" + 0.003*"vaccinated" + 0.003*"covidvaccine" + 0.003*"vaccination"' | '0.072*"pfizer" + 0.032*"vaccine" + 0.019*"booster" + 0.013*"moderna" + 0.010*"omicron" + 0.006*"variant" + 0.005*"astrazeneca" + 0.005*"covidvaccine" + 0.004*"people" + 0.004*"vaccination"' |
| 9<br>Sputnik-V | '0.044*"sputnikv" + 0.035*"vaccine" + 0.013*"pfizer" + 0.010*"sputnik" + 0.009*"omicron" + 0.008*"booster" + 0.008*"russian" + 0.007*"vaccinated" + 0.007*"people" + 0.006*"variant"' | '0.044*"sputnikv" + 0.035*"vaccine" + 0.013*"pfizer" + 0.010*"sputnik" + 0.009*"omicron" + 0.008*"booster" + 0.008*"russian" + 0.007*"vaccinated" + 0.007*"people" + 0.006*"variant"' | '0.059*"sputnikv" + 0.056*"vaccine" + 0.017*"pfizer" + 0.016*"sputnik" + 0.011*"booster" + 0.010*"omicron" + 0.010*"vaccinated" + 0.009*"moderna" + 0.007*"russian" + 0.007*"sinopharm"' | '0.064*"sputnikv" + 0.027*"vaccine" + 0.013*"pfizer" + 0.012*"sputnik" + 0.010*"booster" + 0.009*"omicron" + 0.007*"vaccinated" + 0.007*"moderna" + 0.007*"people" + 0.007*"russian"' | '0.028*"sputnikv" + 0.022*"vaccine" + 0.010*"pfizer" + 0.006*"sputnik" + 0.006*"omicron" + 0.005*"booster" + 0.005*"vaccinated" + 0.004*"vaccination" + 0.004*"russian" + 0.004*"people"' |



```
"canada" + 0.001*"omnicron" + 0.001*"africa" + 0.001*"infection" + 0.001*"effectiveness" + 0.001*"received" + 0.001*"johnson" + 0.001*"inactivated" + 0
*"making" + 0.001*"likely" + 0.001*"reaction" + 0.001*"release" + 0.001*"johnson" + 0.001*"effectiveness" + 0.001*"getvaccinated" + 0.001*"indian" + 0.
esterday" + 0.001*"boosted" + 0.001*"everyone" + 0.001*"received" + 0.001*"vaccinemandate" + 0.001*"johnson" + 0.001*"making" + 0.001*"around" + 0.001*
.001*"inactivated" + 0.001*"taking" + 0.001*"current" + 0.001*"administered" + 0.001*"market" + 0.001*"medium" + 0.001*"mandate" + 0.001*"likely" + 0.0
pprove" + 0.001*"american" + 0.001*"option" + 0.001*"inactivated" + 0.001*"getvaccinated" + 0.001*"number" + 0.001*"scientist" + 0.001*"safety" + 0.001
"scientist" + 0.001*"current" + 0.001*"american" + 0.001*"public" + 0.001*"infection" + 0.001*"option" + 0.001*"science" + 0.001*"yesterday" + 0.001*"l
terday" + 0.001*"working" + 0.001*"export" + 0.001*"scientist" + 0.001*"mandate" + 0.001*"immunity" + 0.001*"infection" + 0.001*"getvaccinated" + 0.001
 + 0.001*"american" + 0.001*"update" + 0.001*"canada" + 0.001*"public" + 0.001*"response" + 0.001*"vaccinesideeffects" + 0.001*"around" + 0.001*"curren
.001*"american" + 0.001*"yesterday" + 0.001*"mutation" + 0.001*"number" + 0.001*"really" + 0.001*"getvaccinatednow" + 0.001*"likely" + 0.001*"update" +
ough" + 0.001*"boosted" + 0.001*"pharma" + 0.001*"working" + 0.001*"received" + 0.001*"vaccinesideeffects" + 0.001*"vaxxed" + 0.001*"another" + 0.001*"
```

## CONCLUSION

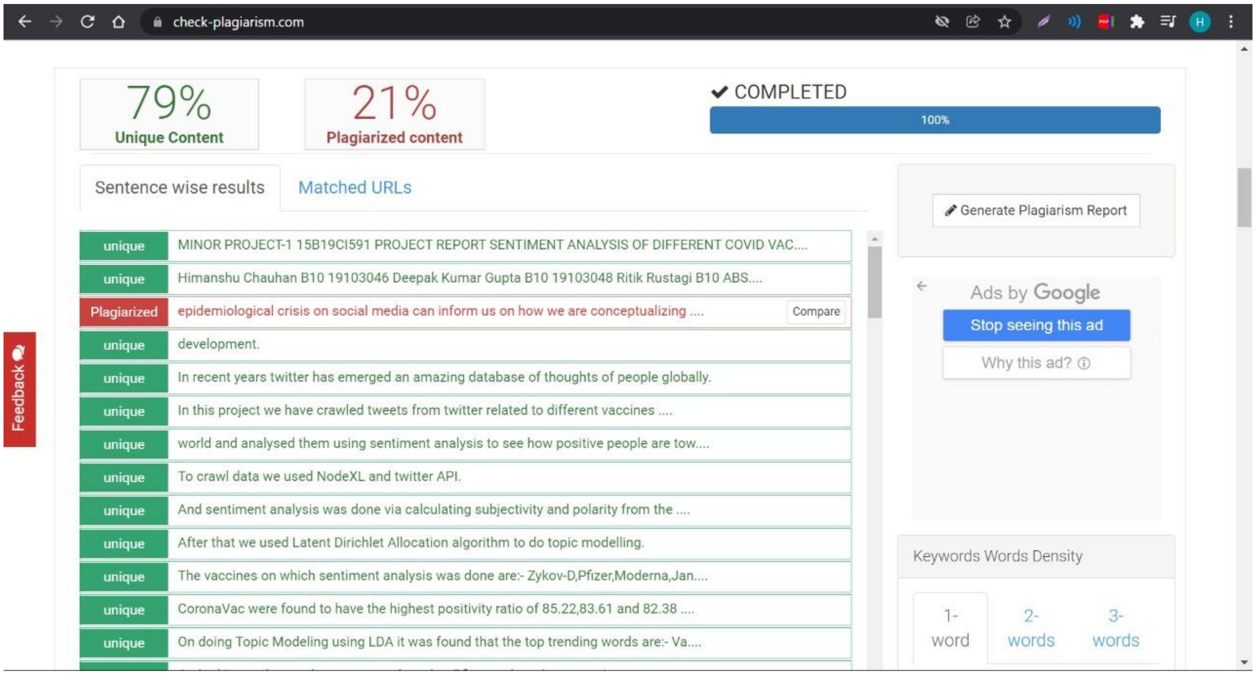
Based on the Sentiment Analysis,

| Vaccine Name | Positive Tweet (%) | Negative Tweet (%) | Relative Ratio | Percent Positivity (%) |
|---|---|---|---|---|
| **AstraZeneca** | 45.1 | 17 | 2.65 | 72.62 |
| **Booster Shots** | 42.5 | 17.2 | 2.47 | 71.18 |
| **CoronaVac** | 53.8 | 11.5 | 4.67 | 82.38 |
| **Covexin** | 49.1 | 12 | 4.09 | 80.36 |
| **Covishield** | 51.9 | 9 | 5.76 | 85.22 |
| **Janssen** | 42.3 | 11.5 | 3.67 | 78.62 |
| **Moderna** | 44.7 | 14 | 3.19 | 76.14 |
| **Pfizer** | 40.6 | 17.4 | 2.33 | 70 |
| **Sputnik-V** | 50 | 9.8 | 5.10 | 83.61 |

Relative ranking based on people's sentiment-

1. **Covishield**
2. **Sputnik-V**
3. **CoronaVac**
4. **Covexin**
5. **Janssen**
6. **Moderna**
7. **AstraZeneca**
8. **Booster Shots**
9. **Pfizer**

# PLAGIARISM REPORT



79%
Unique Content

21%
Plagiarized content

✓ COMPLETED

100%

Sentence wise results | Matched URLs

| | |
|---|---|
| unique | MINOR PROJECT-1 15B19CI591 PROJECT REPORT SENTIMENT ANALYSIS OF DIFFERENT COVID VAC.... |
| unique | Himanshu Chauhan B10 19103046 Deepak Kumar Gupta B10 19103048 Ritik Rustagi B10 ABS.... |
| Plagiarized | epidemiological crisis on social media can inform us on how we are conceptualizing .... |
| unique | development. |
| unique | In recent years twitter has emerged an amazing database of thoughts of people globally. |
| unique | In this project we have crawled tweets from twitter related to different vaccines .... |
| unique | world and analysed them using sentiment analysis to see how positive people are tow.... |
| unique | To crawl data we used NodeXL and twitter API. |
| unique | And sentiment analysis was done via calculating subjectivity and polarity from the .... |
| unique | After that we used Latent Dirichlet Allocation algorithm to do topic modelling. |
| unique | The vaccines on which sentiment analysis was done are:- Zykov-D,Pfizer,Moderna,Jan.... |
| unique | CoronaVac were found to have the highest positivity ratio of 85.22,83.61 and 82.38 .... |
| unique | On doing Topic Modeling using LDA it was found that the top trending words are:- Va.... |

Keywords Words Density

1-word | 2-words | 3-words

# REFERENCES

https://docs.tweepy.org/en/stable/api.html

https://docs.tweepy.org/en/stable/api.html

https://github.com/bisguzar/twitter-scraper

https://apify.com/vdrmota/twitter-scraper

https://www.geeksforgeeks.org/python-for-data-science/

https://www.geeksforgeeks.org/top-10-python-libraries-for-data-science-in-2021/

https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning

https://www.geeksforgeeks.org/python-api-followers-in-tweepy/

https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

https://scholar.google.co.in/scholar?q=corona+research+paper+using+twitter&hl=en&as_sdt=0&as_vis=1&oi=scholart

https://www.geeksforgeeks.org/latent-dirichlet-allocation/

https://www.geeksforgeeks.org/twitter-sentiment-analysis-using-python/

https://www.youtube.com/watch?v=U8m5ug9Q54M

https://www.youtube.com/watch?v=U8m5ug9Q54M

https://www.youtube.com/watch?v=3KaffTIZ5II&t=153s