# EXPLORATORY DATA ANALYSIS OF TWEETS IN INDIA RELATED TO COVID

# (FROM 1ST MARCH 2021 TO 31ST AUGUST 2021)

**SUBMITTED BY:**

**AYUSH MITRA**

**CU ROLL NO.:193012-21-0079**

**CU REGISTRATION NUMBER: 012-1111-0700-19**

**GANGOTRI BASU**

**UNIVERSITY ROLL NO.:193012-11-0078**

**CU REGISTRATION NUMBER: 012-1211-0840-19**

**M.MAHABUB A KHODA**

**UNIVERSITY ROLL NO.:193012-21-0111**

**CU REGISTRATION NUMBER: 012-1115-0742-19**

**Under the guidance of,**

**ANTIKA SINHA**

**Assistant Professor, Department of Computer Science,**

**Asutosh College, Kolkata, India**

**A Project Report submitted to the Department of Computer Science, Asutosh College, University of Calcutta in partial fulfilment of the requirements for the Degree of**

**B. Sc. IN COMPUTER SCIENCE (HONOURS)**

**2022**

# EXPLORATORY DATA ANALYSIS OF TWEETS IN INDIA RELATED TO COVID

# (FROM 1ST MARCH 2021 TO 31ST AUGUST 2021)

**SUBMITTED BY:**

**AYUSH MITRA**

**CU ROLL NO.:193012-21-0079**

**CU REGISTRATION NUMBER: 012-1111-0700-19**

**GANGOTRI BASU**

**UNIVERSITY ROLL NO.:193012-11-0078**

**CU REGISTRATION NUMBER: 012-1211-0840-19**

**M.MAHABUB A KHODA**

**UNIVERSITY ROLL NO.:193012-21-0111**

**CU REGISTRATION NUMBER: 012-1115-0742-19**

**Under the guidance of,**

**ANTIKA SINHA**

**Assistant Professor, Department of Computer Science,**

**Asutosh College, Kolkata, India**

**Project Code: CMS-A-CC-6-13-P & CMS-A-CC-6-14-P**

**Paper Code: CMS-A-CC-6-13-P & CMS-A-CC-6-14-P**

**Paper Name: Project Design and Documentation & Project Implementation and Presentation**

**Department of Computer Science, Asutosh College**

**Affiliated under the University of Calcutta**

# Certificate

This is to certify that the project topic entitled "**Exploratory Data Analysis Of Tweets In India Related To Covid (From 1st March 2021 To 31st August 2021)**" has been solely prepared and submitted by **Ayush Mitra** (Roll No: 193012-21-0079, Reg. No.: 012-1111-0700-19), **Gangotri Basu**(Roll No.: 193012-11-0078, Reg. No.: 012-1211-0840-19) and **M.Mahabub A Khoda** (Roll No: 193012-21-0111, Reg. No.: 012-1115-0742-19) under the supervision of **Prof. Antika Sinha**, Assistant Professor, Department of Computer Science, Asutosh College, Kolkata, India in partial fulfilment of Computer Science Project examination conducted by **UNIVERSITY OF CALCUTTA.**

**(PROF. ANTIKA SINHA)**

**Assistant Professor & Supervisor,**

Department of Computer Science,

Asutosh College, Kolkata-700026, India

**(DR. SAMIR MALAKAR)**

**Assistant Professor & Head,**

Department of Computer Science,

Asutosh College, Kolkata-700026, India

**EXAMINER(S)**

# ACKNOWLEDGEMENT

We, Ayush Mitra, Gangotri Basu, M.Mahabub A Khoda, students of Asutosh College affiliated by University of Calcutta, are grateful to our supervisor for this project **Prof. Antika Sinha** for motivating and guiding us throughout the process. It would have been impossible for us without her help.

We're also grateful to the staff members and librarians for providing us the facilities we needed for this project. We also express gratitude for other researchers whose thesis paper helped us to reach the conclusion.

At last, Acknowledgment section will be incomplete if we don't show our gratitude for our parents. They supported us in every decision we took and kept motivating us in every way possible.

**(AYUSH MITRA)**
CU Roll No.:193012-21-0079

**(GANGOTRI BASU)**
CU Roll No.:193012-11-0078

**(M.MAHABUB A KHODA)**
CU Roll No.:193012-21-0111

# CONTENT

# ABSTRACT

Exploratory Data Analysis (EDA) is an approach using descriptive statistics and Bar Charts tools to better understand data. It is mainly used to examine data distribution, handling missing values of dataset, detect outliers and anomalies, removing duplicate data, discover pattern and test underlying assumptions. It is a robust first step before application of other statistical methods.

In this report we will show the application of EDA over a collection of twitter datasets over a period of time on the topic COVID, and identify different key words and relations between them and the variation of those relation in six months.

# 1. INTRODUCTION

## 1.1 Domain Description:

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus denoted by the year in which it first appeared. The disease has affected worldwide. All the countries took precautionary measures suggested by WHO (World Health Organization) *[1] to prevent it from spreading more.

In India, the number of cases were at peak through February 2020 to June 2020 and March 2021 to August 2021.These two time periods are referred as covid-19 first wave and second wave respectively. During these bad times people took the path of sharing the local news in social media platforms like Twitter, Facebook, Instagram and so on.

Twitter is a microblogging and social networking service on which users post and interact with messages known as "tweets". Registered users can post, like, and retweet tweets, however, unregistered users have the ability to only read tweets that are publicly available. The information posted by the users during this pandemic reflects deadly effects of Covid including high death rate, Country's economic loss, switch in mode of education system.

Datasets collected from twitter provides vast information unlike other platforms. These collected datasets based on tweets related to covid can help in projects like analysis of Covid.

## 1.2 Motivation:

The primary objective of exploratory data analysis is to uncover the underlying structure. The structure of the various data sets determines the trends, patterns, and relationships among them. A business cannot come to a final conclusion or draw assumptions from a huge quantity of data and rather requires taking an exhaustive look at the data set through an analytical lens.

Therefore, performing an Exploratory Data Analysis allows data scientists to detect errors, debunk assumptions, and much more to ultimately select an appropriate predictive model.

## 1.3 Scope of the work:

As different tweets reflect a collection of words. By analysing these tweets month wise we can find out the relation between various tweets by its words which may not be tweeted by same person or on same time but are related to same topic. And can generate the clusters formed by these words. The objective of this paper is to analyse the collection of tweets and find out the relations between various tweets related to COVID and also to find the hashtags used and userhandles tagged the most.

Key points:

- Collecting tweets of different months based on keywords.
- Removing all the unwanted fields.

[1] The World Health Organization is a specialized agency of the United Nations responsible for international public health. The WHO Constitution states its main objective as "the attainment by all peoples of the highest possible level of health".

- Tokenizing the tweets, analysing the words and formation of clusters and set of user handles and hashtags.

After executing these tasks, we will be able to visualize the relation between various terms in the form of cluster.

| **Keywords Used in Twitter Scrapping** | | |
|---|---|---|
| Covid | Curfew | Covid-19 |
| Covid | Vaccine | Remdesivir |
| COVID | corona | Coronavirus |

## 2. BACKGROUND OF THE RELATED WORKS

In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling and thereby contrasts traditional hypothesis testing. Regarding research in exploratory analysis on Covid and social media:

*Han et al* [1 ] presented a study on the sentiment of the people in China on COVID-19. They categorized COVID-19-related posts into seven topics with further division into 13 more subtopics.

*Usman Naseem et al* [2] aimed to identify the topics and the community sentiment dynamics expressed on Twitter about COVID-19. The research addresses the automatic detection of people's sentiments expressed on Twitter due to COVID-19 and topics mostly discussed by the Twitter users while expressing sentiments about COVID-19.

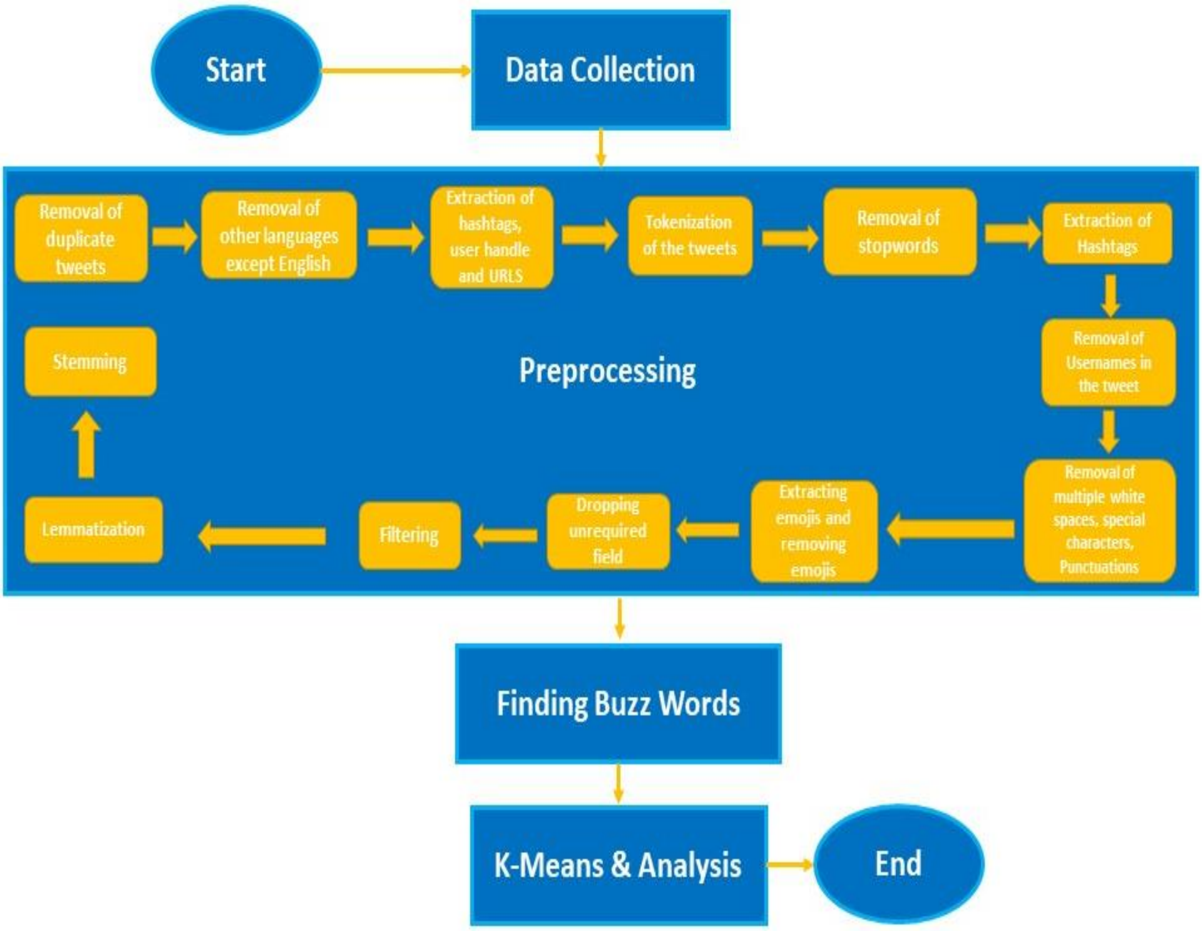
## 3. METHODOLGY

### 3.1 Problem Formulation:

Raw Datasets of tweets with given stroke of keywords the months, March 2021to August 2021, are provided as input.

The analysis is expected to process the data to give the outputs as mentioned below:

- Bar Chart representing buzzwords of each month.
- Clusters of each months using K-means algorithm and density, top features of each of them.


Start
Data Collection
Removal of duplicate tweets
Removal of other languages except English
Extraction of hashtags, user handle and URLS
Tokenization of the tweets
Removal of stopwords
Extraction of Hashtags
Removal of Usernames in the tweet
Removal of multiple white spaces, special characters, Punctuations
Extracting emojis and removing emojis
Dropping unrequired field
Filtering
Lemmatization
Stemming
Preprocessing
Finding Buzz Words
K-Means & Analysis
End

## 3.2 Algorithm Description:

### 3.2.1 Data Collection

Firstly, we collect tweets using the **'twint'** module in python. Tweets are collected based on different keywords from start to end of each month (e.g., 1st March 2021 to 31st March 2021) and are stored in different files (individual files for individual month).
Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API.

### 3.2.2 Pre-processing

Initially, we had 36 fields (or attributes), but after removing the useless fields we have only 7 fields (Created at, user name, tweet, language, reply count, retweet count, like count).

#### 3.2.2.1 Removal of duplicate tweets.

While searching for tweets using different keywords there are chances that different keywords return the same tweet for two or more keywords. So, if a tweet is repeated, we drop the copies of the tweet.
*E.g.,* For the month August, initially there were 3025 tweets after processing we find that within them there are 354 duplicate tweets, so, after removal of these tweets, 2671 tweets are left.

#### 3.2.2.2 Removal of other languages except English

**This method is done as per as following steps:**

i) For every tweet check the language, languages are other than English(‘en’) are put in index_name.

ii) drop index_name using drop function.

E.g., for the month August, before dropping other languages there where 2671 tweets, but after dropping the tweets that doesn’t fall under the English category, tweets left are 1698.

#### 3.2.2.3 Extraction of hashtags, user handle and URLS

From each tweets the hashtags, user handle and URLS are extracted and stored in another file.

**This method is done as per as following steps:**

i) Define function extract_at_the_rates(text,i=0) and extract_https(text,i=0) which will be used to extract the userhandles and URLs.

ii) Find the words starting with ‘@’and ‘https ’and put them in two different lists.

#### 3.2.2.4 Tokenization of the tweets

Separating each word from the whole tweet, so that, each word can be used for analysis.

**This method is done as per as following steps:**

i) Import ‘textblob’ module.

ii) Define function tokenizing_month name(text,i=0).

iii) Take the tokenized words <month_name>[‘tokenized_tweet’][i].

#### 3.2.2.5 Removal of stopwords

After the tweets are tokenized, each word is supposed to be used for analysis, but within them there are certain words which isn’t needed in the analysis *e.g.,* a, an, is, etc.

**This method is done as per as following steps:**

i) For each variable token in list, if length of token is 1, then remove token.

ii) For each token in list, if length of token is 2, then remove token.

iii) For each token in list, if token is one of the pre-defined words, then remove token.

#### 3.2.2.6 Extraction of Hashtags using nfx package

nfx or Neat Text is a simple Natural Language Processing package for cleaning text data and pre-processing text data. It can be used to clean sentences, extract emails, phone numbers, weblinks, and emojis from sentences. It can also be used to set up text pre-processing pipelines.

**This method is done as per as following steps:**

i) Define function extract_hashtags(text,i=0) which will be used to extract hashtags.

ii) Find the words starting with ‘#’and put them in a list.

#### 3.2.2.7 Removal of Usernames in the tweet using lambda function

Python Lambda function is known as the anonymous function that is defined without a name. Python allows us to not declare the function in the standard manner, i.e., by using the def keyword. Rather, the anonymous functions are declared by using the lambda keyword. However, Lambda functions can accept any number of arguments, but they can return only one value in the form of expression.

#### 3.2.2.8 Removal of multiple white spaces, special characters, Punctuations using nfx package.

**This method is done as per as following steps:**

i) import neattext.functions as nfx.

ii) Do tweet_list_march = tweets_march.apply(nfx.remove_stopwords) to remove stop-words.

iii) Do tweet_list_march = tweet_list_march .apply(nfx.remove_special_characters)

iv) tweet_list_march = tweet_list_march.apply(nfx.remove_punctuations)

#### 3.2.2.9 Extracting emojis and removing emojis

**This method is done as per as following steps:**

i) Import ‘emoji’ and ‘re’ module.

ii) Define remove_emoji(text,i=0) function which will be used to remove emoticons.

iii) Extract emojis from the tweets using get_emoji_regexp() function and put the clean tweet in new_text.

#### 3.2.2.10 Dropping unrequired fields

After saving the clean tweets, we are dropping rest of the fields or column, except tweet, clean_tweet, hashtag, userhandle, URLS, emoji.

#### 3.2.2.11 Filtering

Next, we are removing the non-meaningful words using nfx. We are extending the list of stopwords in nfx and adding few more stopwords of our own (for e.g., amp, gt, etc.). Then, we are removing digits (*E.g.,* Covid and Covid19, are same so it’s reduced to Covid). Also, all the words and characters are changed from uppercase to lowercase. Then we are removing single characters and words with two characters (assuming there’s no meaningful words with two or one characters).

**This method is done as per as following steps:**

For each tweet,

i) Put the tweets without any digits in monthname[clean_tweet][i].

ii) Change all the uppercase tweets using .lower() function.

iii) Remove the new added custom words using .remove_custom_words() function.

#### 3.2.2.12 Lemmatization

Lemmatization usually refers to doing things properly with the use of vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as lemma.

*E.g.,* Vaccines and vaccine are both same word but will be analysed as different word.

This doesn't totally help us to understand the density and the importance of the word in tweets. So, after lemmatization vaccines become vaccine. Making them both the same word and reducing redundancy.

#### 3.2.2.13 Stemming

Stemming is a method of normalization of words in Natural Language Processing. It is a technique in which a set of words in a sentence are converted into a sequence to shorten its lookup. In this method, the words having the same meaning but have some variations according to the context or sentence are normalized.

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning.

### 3.2.3 Finding Buzz Words

Buzzwords are the frequently used words which defines a given situation for a given period of time. For e.g., In the dataset of March 2021, Covid, Corona, Case, Death, Vaccine, Total-these were the frequent words which described the situation in march. A list of is created for all the pre-processed words in a month using **.split()** function. Then the counting of each word is done by checking its frequency in the list and then the value is stored in a dictionary. Then the dictionary is converted into a data frame using pandas library.

### 3.2.4 K-Means & Analysis

After finding the buzz words we need to find the relations between those words for which the K-Means clustering algo is used, to show how the words are related in the form of clusters.

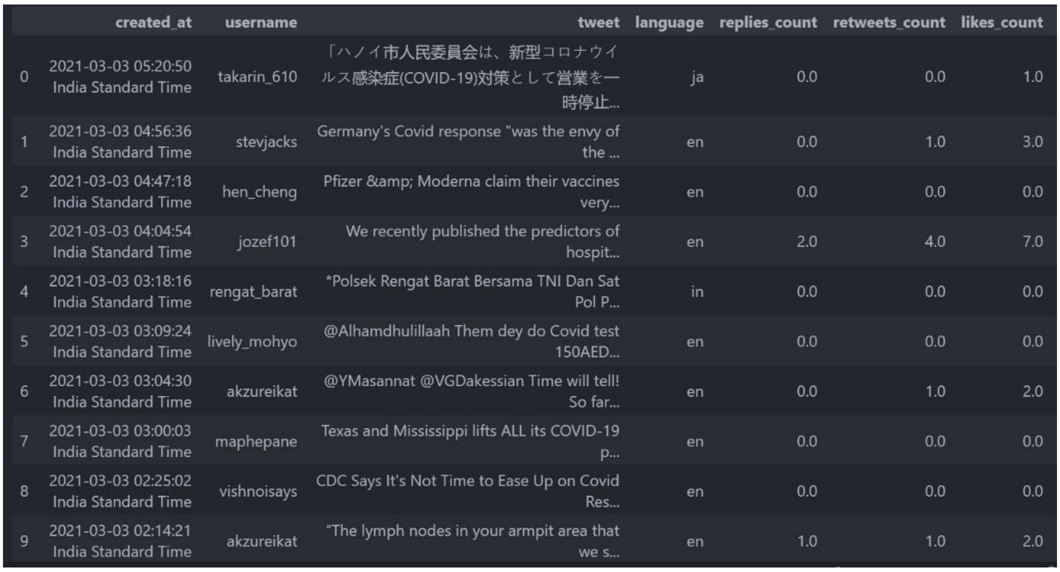
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. Elbow and Silhouette methods are used to find the optimal values of K. Ambiguity arises for the elbow method to pick the value of K. Silhouette analysis can be used to study the separation distance between the resulting clusters and can be considered a better method compared to the Elbow method. That’s why we used Silhouette method.

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.

## 4. Implementation

For the month of March:

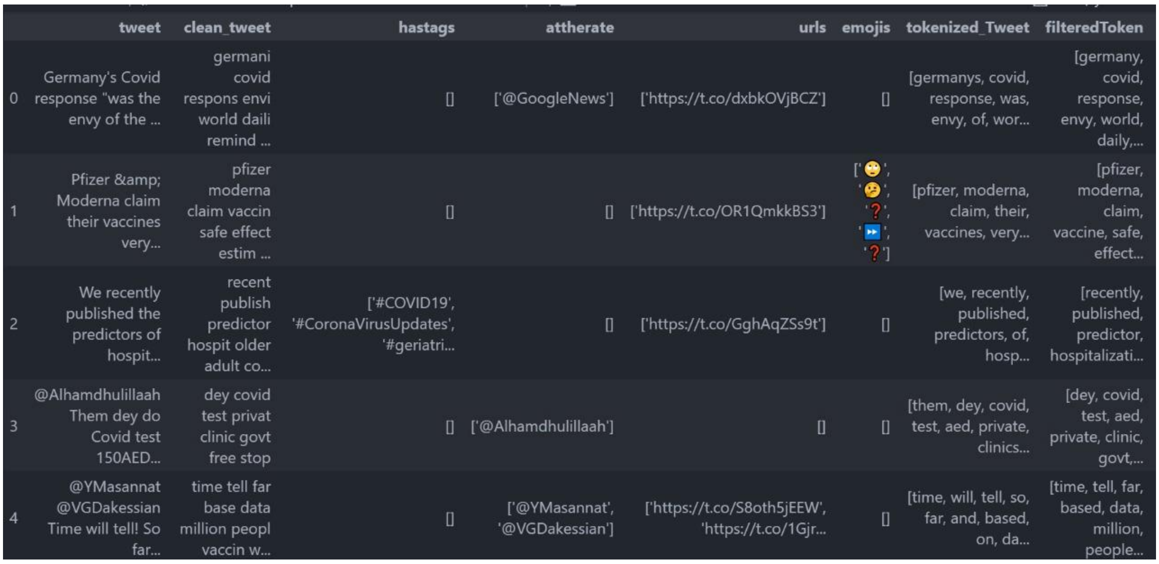
- The raw dataset has 3315 no of tweets. First 10 tweets are shown below using **.head()** function.



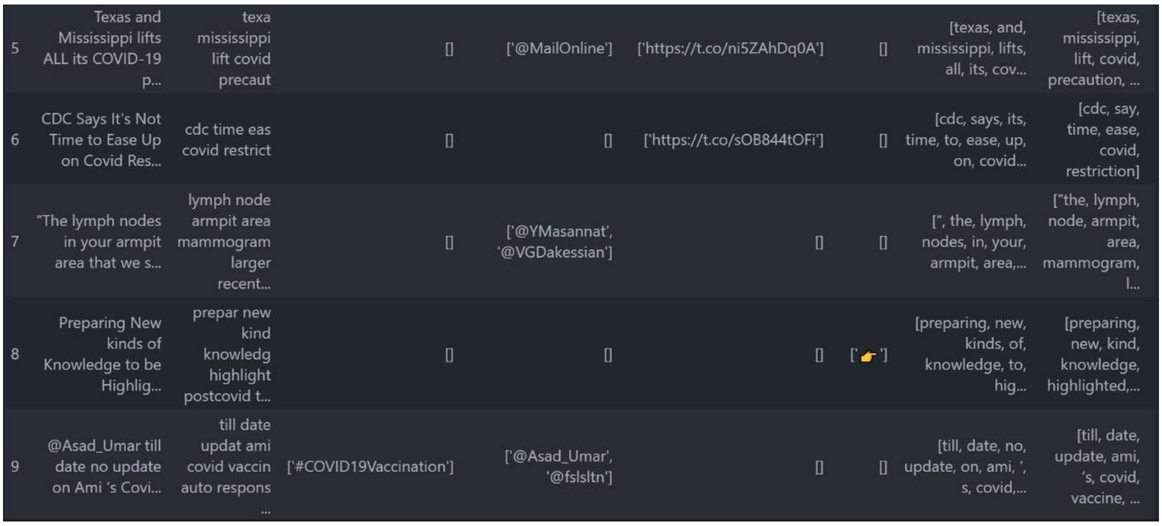
| | created_at | username | tweet | language | replies_count | retweets_count | likes_count |
|---|---|---|---|---|---|---|---|
| 0 | 2021-03-03 05:20:50 India Standard Time | takarin_610 | 「ハノイ市人民委員会は、新型コロナウイルス感染症(COVID-19)対策として営業を一時停止... | ja | 0.0 | 0.0 | 1.0 |
| 1 | 2021-03-03 04:56:36 India Standard Time | stevjacks | Germany's Covid response "was the envy of the ... | en | 0.0 | 1.0 | 3.0 |
| 2 | 2021-03-03 04:47:18 India Standard Time | hen_cheng | Pfizer &amp; Moderna claim their vaccines very... | en | 0.0 | 0.0 | 0.0 |
| 3 | 2021-03-03 04:04:54 India Standard Time | jozef101 | We recently published the predictors of hospit... | en | 2.0 | 4.0 | 7.0 |
| 4 | 2021-03-03 03:18:16 India Standard Time | rengat_barat | *Polsek Rengat Barat Bersama TNI Dan Sat Pol P... | in | 0.0 | 0.0 | 0.0 |
| 5 | 2021-03-03 03:09:24 India Standard Time | lively_mohyo | @Alhamdhulillaah Them dey do Covid test 150AED... | en | 0.0 | 0.0 | 0.0 |
| 6 | 2021-03-03 03:04:30 India Standard Time | akzureikat | @YMasannat @VGDakessian Time will tell! So far... | en | 0.0 | 1.0 | 2.0 |
| 7 | 2021-03-03 03:00:03 India Standard Time | maphepane | Texas and Mississippi lifts ALL its COVID-19 p... | en | 0.0 | 0.0 | 0.0 |
| 8 | 2021-03-03 02:25:02 India Standard Time | vishnoisays | CDC Says It's Not Time to Ease Up on Covid Res... | en | 0.0 | 0.0 | 0.0 |
| 9 | 2021-03-03 02:14:21 India Standard Time | akzureikat | "The lymph nodes in your armpit area that we s... | en | 1.0 | 1.0 | 2.0 |

*Figure 1:First 10 Rows of the Raw Dataset of March*

- After pre-processing the datasets (removing duplicate tweets, languages except English, extracting hashtags, userhandles, URLs, tokenizing the tweets, removing stopwords, extracting hashtags, removing usernames, multiple white spaces, special characters, and punctuations, extracting and removing emojis, dropping unrequired fields, filtering, lemmatization and stemming),2452 tweets will be left. First 10 tweets are shown below using **.head()** function.



| | tweet | clean_tweet | hastags | attherate | urls | emojis | tokenized_Tweet | filteredToken |
|---|---|---|---|---|---|---|---|---|
| 0 | Germany's Covid response "was the envy of the ... | germani covid respons envi world daili remind ... | [] | ['@GoogleNews'] | ['https://t.co/dxbkOVjBCZ'] | [] | [germanys, covid, response, was, envy, of, wor... | [germany, covid, response, envy, world, daily,... |
| 1 | Pfizer &amp; Moderna claim their vaccines very... | pfizer moderna claim vaccin safe effect estim ... | [] | [] | ['https://t.co/OR1QmkkBS3'] | ['🙄', '🤔', '❓', '⏩', '❓'] | [pfizer, moderna, claim, their, vaccines, very... | [pfizer, moderna, claim, vaccine, safe, effect... |
| 2 | We recently published the predictors of hospit... | recent publish predictor hospit older adult co... | ['#COVID19', '#CoronaVirusUpdates', '#geriatri... | [] | ['https://t.co/GghAqZSs9t'] | [] | [we, recently, published, predictors, of, hosp... | [recently, published, predictor, hospitalizati... |
| 3 | @Alhamdhulillaah Them dey do Covid test 150AED... | dey covid test privat clinic govt free stop | [] | ['@Alhamdhulillaah'] | [] | [] | [them, dey, covid, test, aed, private, clinics... | [dey, covid, test, aed, private, clinic, govt,... |
| 4 | @YMasannat @VGDakessian Time will tell! So far... | time tell far base data million peopl vaccin w... | [] | ['@YMasannat', '@VGDakessian'] | ['https://t.co/S8oth5jEEW', 'https://t.co/1Gjr... | [] | [time, will, tell, so, far, and, based, on, da... | [time, tell, far, based, data, million, people... |



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | Texas and Mississippi lifts ALL its COVID-19 p... | texa mississippi lift covid precaut | [] | ['@MailOnline'] | ['https://t.co/ni5ZAhDq0A'] | [] | [texas, and, mississippi, lifts, all, its, cov... | [texas, mississippi, lift, covid, precaution, ... |
| 6 | CDC Says It's Not Time to Ease Up on Covid Res... | cdc time eas covid restrict | [] | [] | ['https://t.co/sOB844tOFi'] | [] | [cdc, says, its, time, to, ease, up, on, covid... | [cdc, say, time, ease, covid, restriction] |
| 7 | "The lymph nodes in your armpit area that we s... | lymph node armpit area mammogram larger recent... | [] | ['@YMasannat', '@VGDakessian'] | [] | [] | [", the, lymph, nodes, in, your, armpit, area,... | ["the, lymph, node, armpit, area, mammogram, l... |
| 8 | Preparing New kinds of Knowledge to be Highlig... | prepar new kind knowledg highlight postcovid t... | [] | [] | [] | ['👉'] | [preparing, new, kinds, of, knowledge, to, hig... | [preparing, new, kind, knowledge, highlighted,... |
| 9 | @Asad_Umar till date no update on Ami 's Covi... | till date updat ami covid vaccin auto respons ... | ['#COVID19Vaccination'] | ['@Asad_Umar', '@fslsltn'] | [] | [] | [till, date, no, update, on, ami, ', s, covid,... | [till, date, update, ami, 's, covid, vaccine, ... |

*Figure 2: First 10 Rows of the Pre-processed Dataset of March*

- After finding the buzzwords based on the pre-processed dataset, we get the below mentioned Bar Charts:

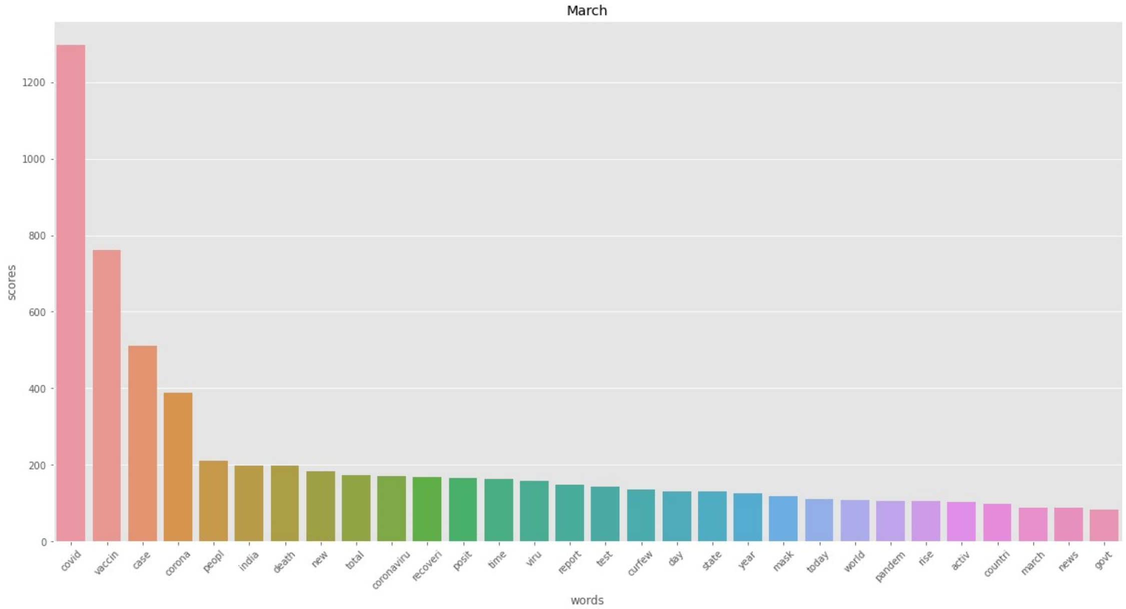


*Figure 3:Buzzwords of March*

- K-means algorithm is applied on the buzzwords to find the clusters of March.
For clusters no. 2, the average silhouette score is: 0.782552832108916
For clusters no. 3, the average silhouette score is: 0.5907462063591088
For clusters no. 4, the average silhouette score is: 0.6065066731272062
The number of clusters here is 2 as the silhouette score of the 2nd cluster is maximum. We also get the tweet density of two clusters and also the top features of both the clusters.

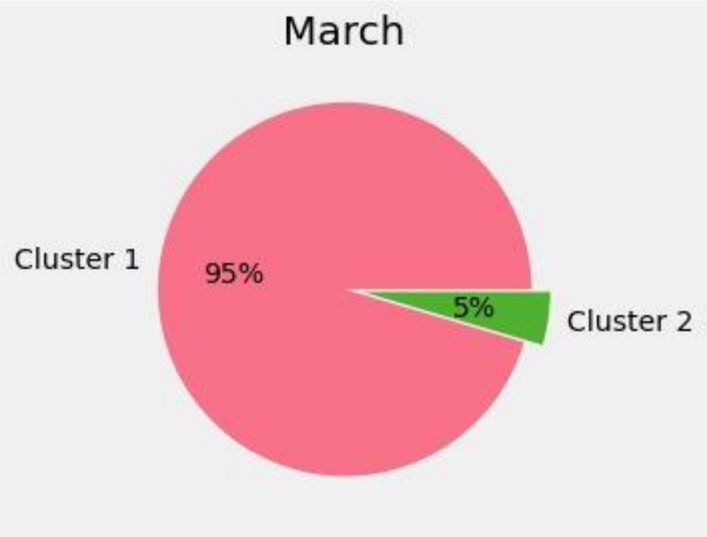


*Figure 4:Tweet Density of Clusters*

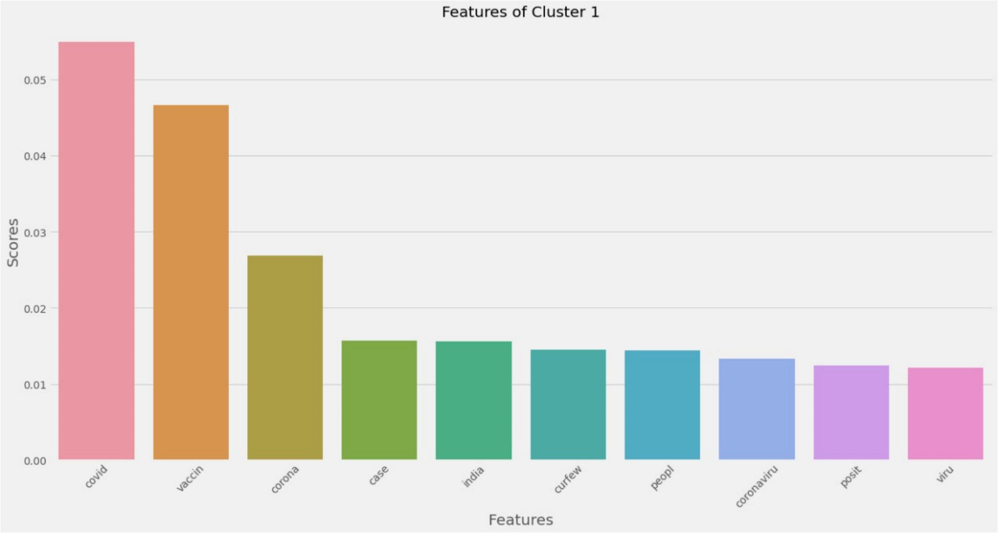


*Figure 5:Top Features of Cluster 1*

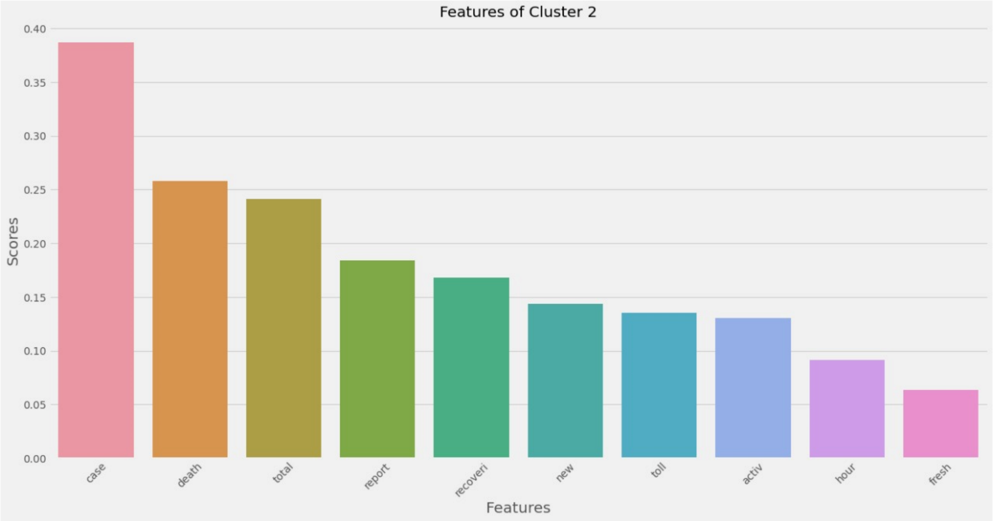


*Figure 6:Top Features of Cluster 2*

## 5. Results and Discussion:

- Clusters for each month from March 2021- August 2021

❖ ***March:***

For clusters no. 2, the average silhouette score is: 0.782552832108916
For clusters no. 3, the average silhouette score is: 0.5907462063591088
For clusters no. 4, the average silhouette score is: 0.6065066731272062

The number of clusters here is 2 as the silhouette score of the 2nd cluster is maximum.

Top 3 Features of 1st cluster: covid, vaccine, corona.
Top 3 Features of 2nd cluster**:** case, death, total.

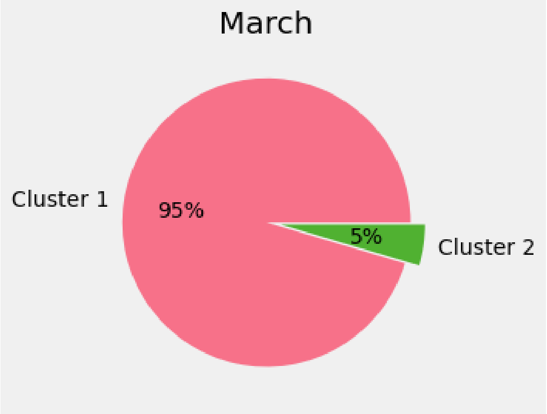


*Figure 7:Density of Tweets in cluster in March*

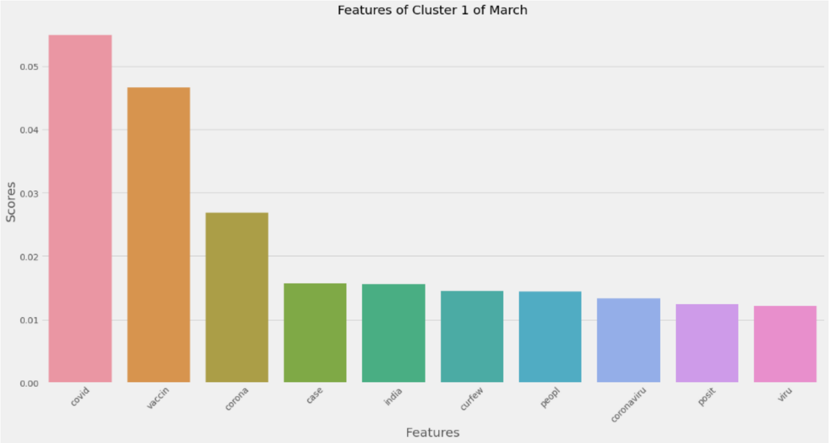


*Figure 8:Top 10 features of Cluster1 of March*

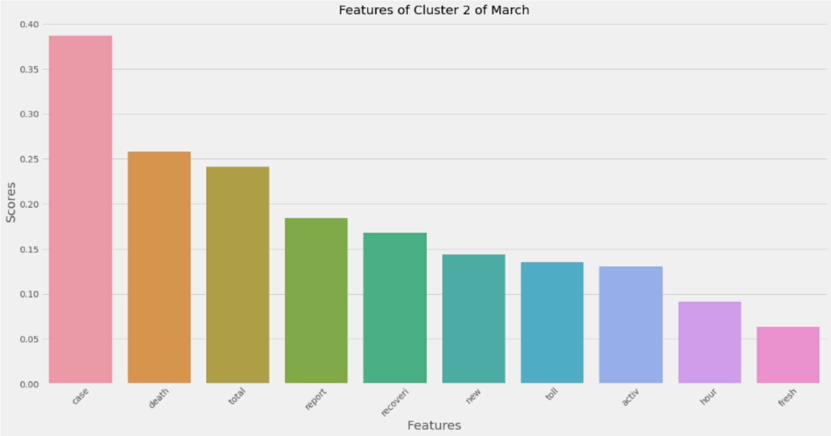


*Figure 9:Top 10 features of Cluster2 of March*

When we observe both the clusters, both of them have 'case' but in 1st cluster it is dependent on 'covid' as it is the most important feature in the 1st cluster, in 2nd cluster 'case' and 'death' related tweets are mostly present.

As of Report of WHO on 29th March 2021, India reported a total of 12,039,644 confirmed cases. Cases per million is 8,594 and a total of 161,843 deaths have been reported. As the number of death and covid cases were increasing in huge rate, 'covid', 'case', 'death' became the buzzwords.

In 1st cluster, top 28 out of 30 buzzwords matches with the 1st cluster, whereas, in 2nd cluster only 20 of 30 matches. So, we can say that the 1st cluster is the better collection of the buzzwords.

Here, the words 'vaccine', 'covid' and 'corona' are still interrelated and 'death', 'total', 'report', 'recovery' is related to case.

- ***April:***

For clusters no. 2, the average silhouette score is: 0.44015121662831413
For clusters no. 3, the average silhouette score is: 0.47669285247560417
For clusters no. 4, the average silhouette score is: 0.38995805240582765

The number of clusters here is 3 as the silhouette score of the 3rd cluster is maximum.

Top 3 Features of 1st cluster: covid, vaccine, people.
Top 3 Features of 2nd cluster: corona, curfew, night.
Top 3 Features of 3rd cluster: case, death, new.

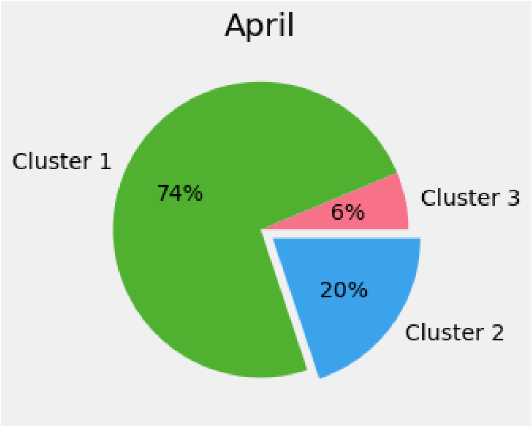


*Figure 10:Density of Tweets in Cluster in April*

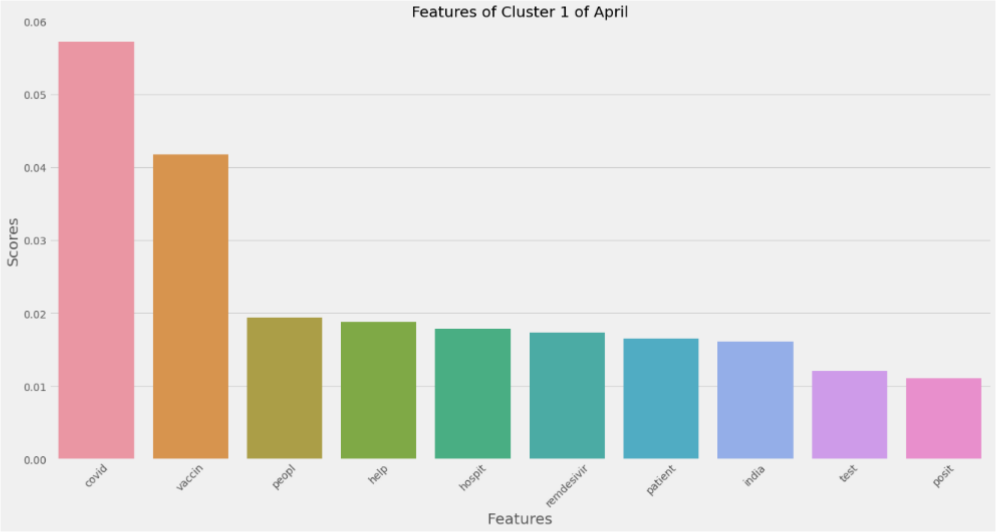


*Figure 11:Top 10 features of Cluster1 of April*

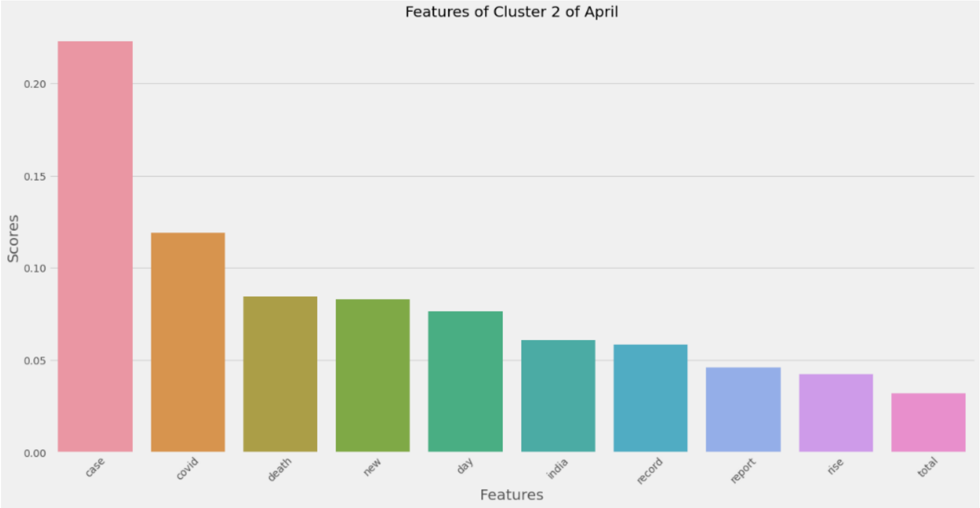


*Figure 12:Top 10 features of Cluster2 of April*

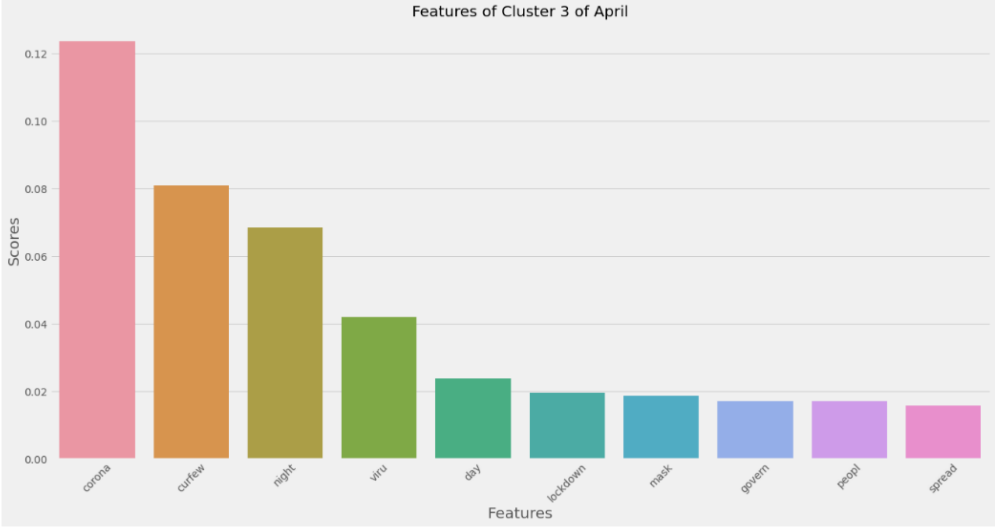


*Figure 13:Top 10 features of Cluster3 of April*

We can see 'vaccine' (though it's frequency here is less than march) is still related to 'Covid' and 'corona' is in different cluster. The frequency of corona is more than that of march. The frequency of 'case' here is less than march but still it's forming a different cluster.

According to India Situation Report by WHO on 28th April 2021, India was reporting highest number of daily cases in the world, almost 50% of new cases reported in the world. As, the increasing case rate became highest, people started asking about Vaccines. That's how vaccine became a buzzword.

In 1st cluster, 26 out of 30 most frequently used buzzwords are present. So, we can say the 1st cluster is the better collection of buzzwords.

- ***May***:

For clusters no. 2, the average silhouette score is: 0.5579674094073745
For clusters no. 3, the average silhouette score is: 0.5454836790759061
For clusters no. 4, the average silhouette score is: 0.4455334459062161

The number of clusters here is 3 as the silhouette score of the 3rd cluster is maximum.

Top 3 Features of 1st cluster: covid, india, help.
Top 3 Features of 2nd cluster: vaccine, covid, dose.
Top 3 Features of 3rd cluster: corona, virus, curfew.

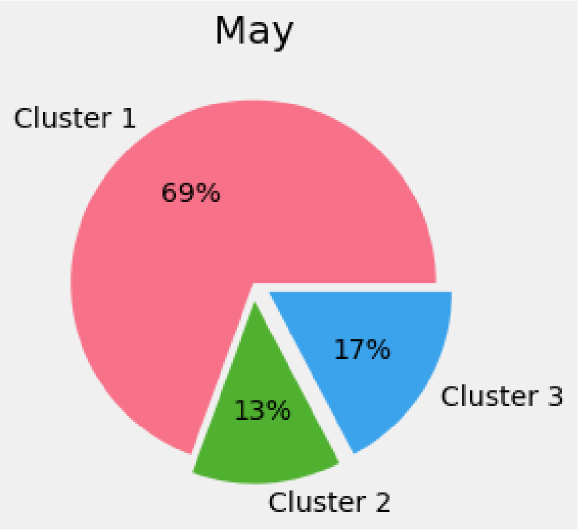


*Figure 14:Density of Tweets in Cluster in May*

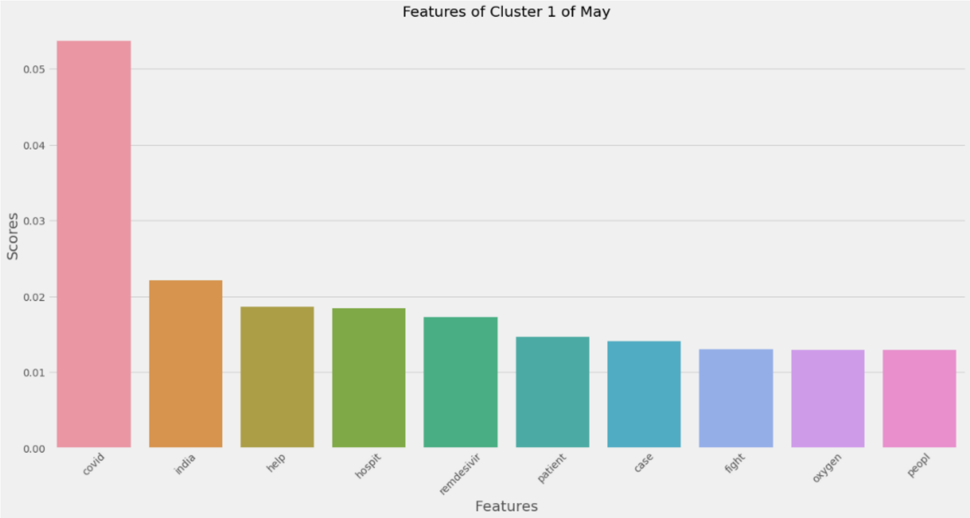


*Figure 15:Top 10 features of Cluster1 of May*

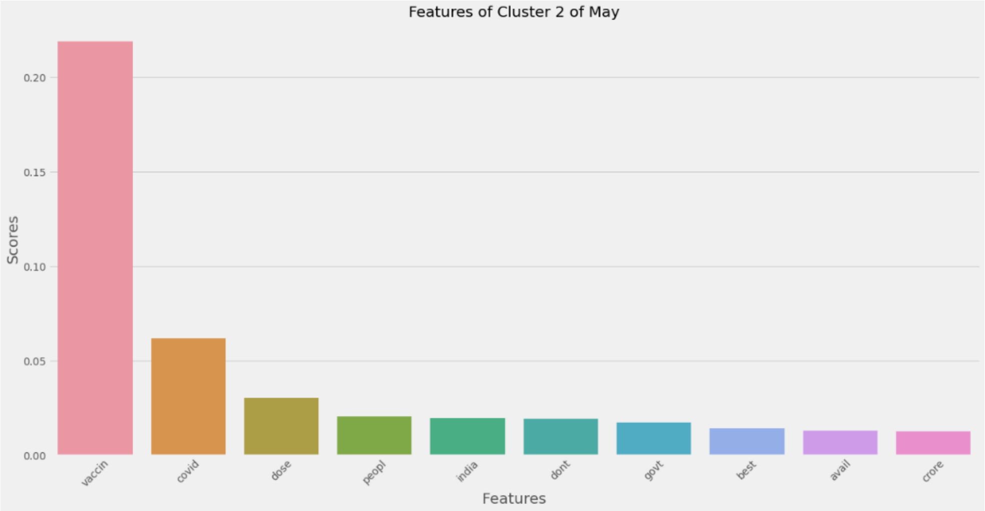


*Figure 16:Top 10 features of Cluster2 of May*

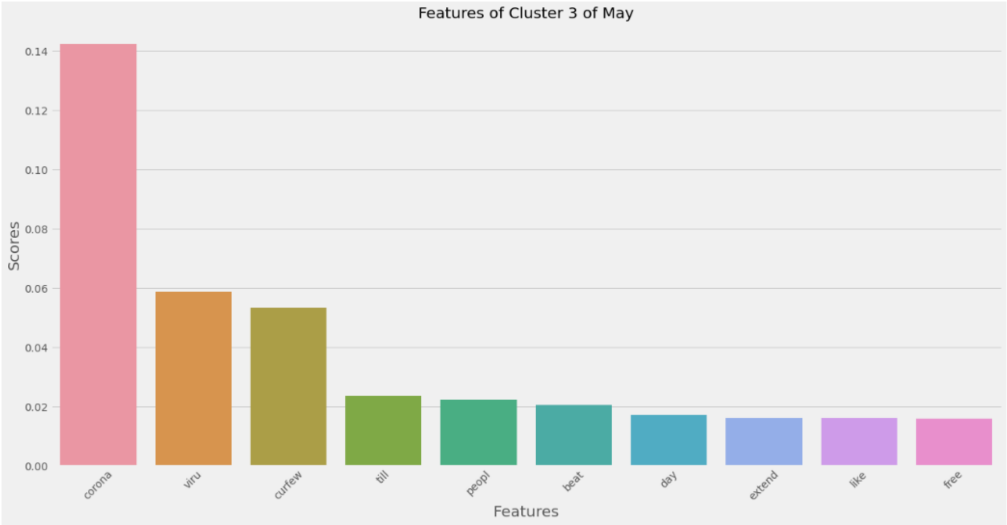


*Figure 17:Top 10 features of Cluster3 of May*

We can see vaccine is forming different cluster. In 1st cluster ‘covid’ is the main feature but in 2nd cluster ‘vaccine’ is the main but ‘covid’ is also there as the 2nd important feature. But, the score of ‘vaccine’ is 0.218 and of ‘covid’ is 0.061, so, we can say ‘vaccine’ is the most important feature. And 3rd cluster is almost similar to that of previous one.

In May, lack of essential supplies like Oxygen, Hospital Bed, Medicines is observed. People started asking for help through Twitter. That’s how ‘covid’, ‘corona’, ‘help’ became frequently used words. According to India Situation Report by WHO on 26th May 2021, there was continuous decline in daily cases after reporting the highest number of cases (4,14,188) on 7th May. Cases per million was 19,235. As, vaccines were introduced to the public. So, ‘vaccine’ was also a buzzword in May.

In 1st cluster, the frequency of the buzzwords is more (28 out of 30 buzzwords).

❖ ***June***:
For clusters no. 2, the average silhouette score is: 0.5960750984332851

For clusters no. 3, the average silhouette score is: 0.6093093451211761
For clusters no. 4, the average silhouette score is: 0.543175356042765

The number of clusters here is 3 as the silhouette score of the 3rd cluster is maximum.

Top 3 Features of 1st cluster: vaccine, covid, dose.
Top 3 Features of 2nd cluster: covid, update, life.
Top 3 Features of 3rd cluster: covid, corona, curfew.

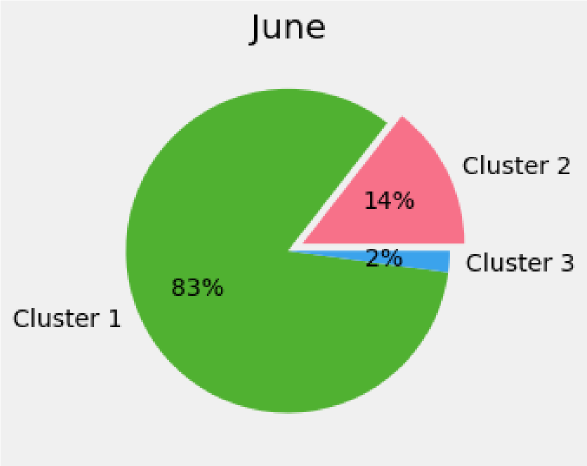


*Figure 18:Density of Tweets in Cluster in June*

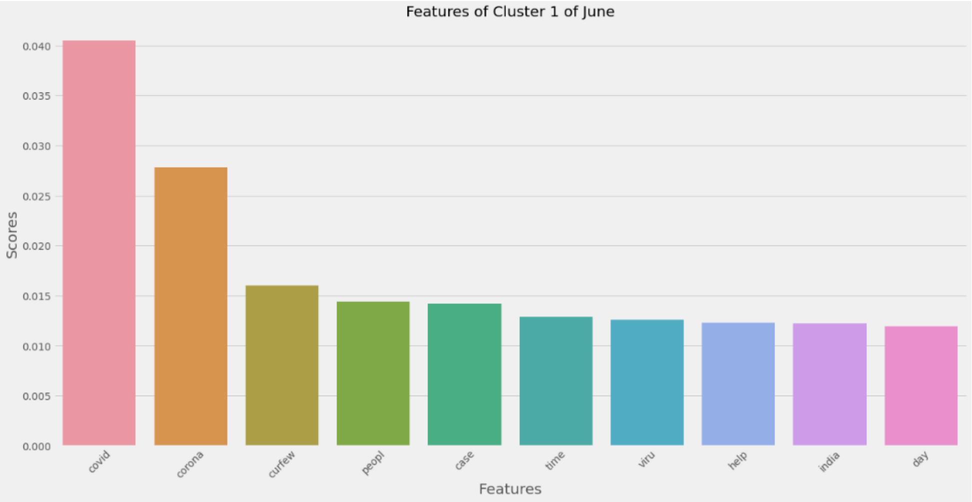


*Figure 19: Top 10 Features of Cluster1 of June*

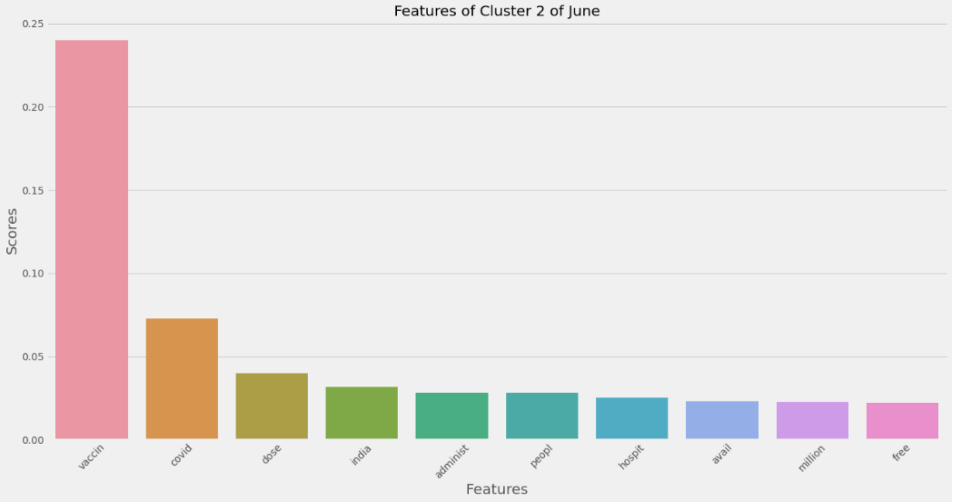


*Figure 20:Top 10 Features of Cluster2 of June*

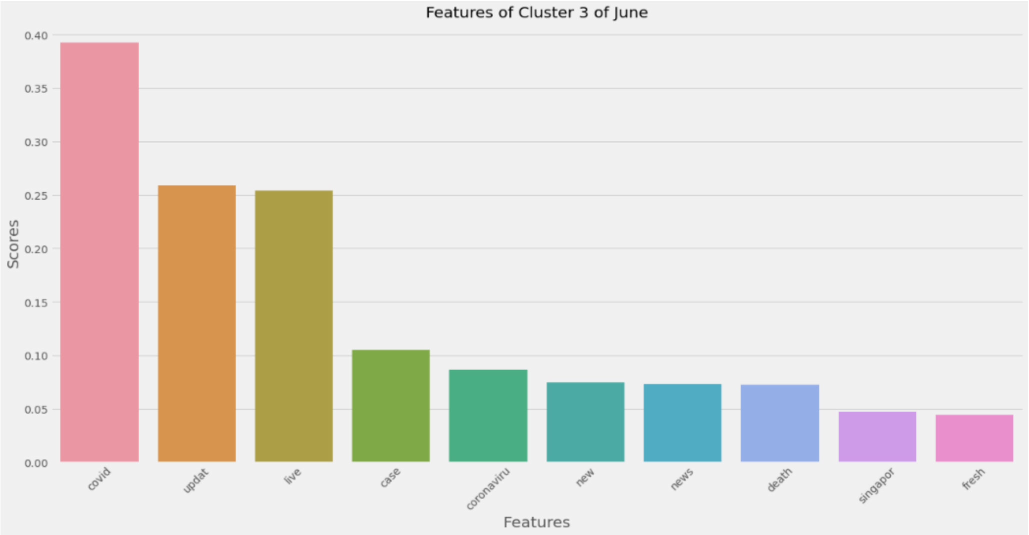


*Figure 21:Top 10 Features of Cluster3 of June*

We can see, ‘covid’ is the most important feature twice and ‘corona’ doesn’t form any different cluster, which is now related to ‘covid’. Also, in 2nd cluster ‘update’ and ‘life’ is related to ‘covid’ which means these tweets are mostly related to various updates of cases. In 3rd cluster. The highest frequency word is ‘covid’ (1229).

According to India Situation Report by WHO on June 30 2021, India reported 13% of new COVID cases reported globally and has recorded the second highest daily number of cases with 48,434 daily cases. Also,total number of dosage of vaccines were 19,85,38,999.This is why, ‘dose’ became the new frequently used word among other words like ‘covid’, ‘vaccine’, ‘case’.

❖ ***July***:
For clusters no. 2, the average silhouette score is: 0.7457518423539595
For clusters no. 3, the average silhouette score is: 0.6575683027605923
For clusters no. 4, the average silhouette score is: 0.5326177732207111

The number of clusters here is 2 as the silhouette score of the 2$^{nd}$ cluster is maximum.

Top 3 Features of 1st cluster: virus, corona, aware.
Top 3 Features of 2nd cluster: covid, vaccine, corona.

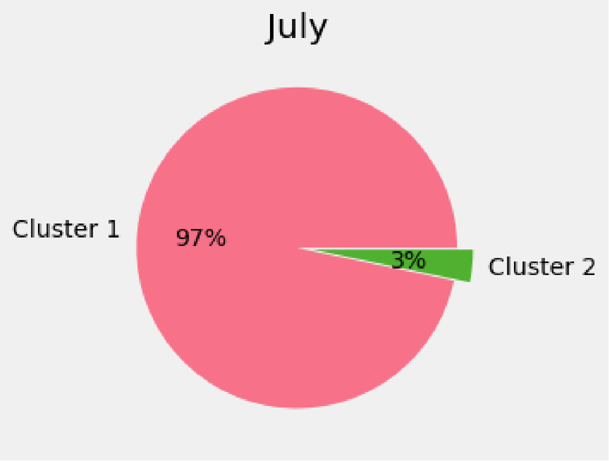


*Figure 22:Density of Tweets in Cluster in July*

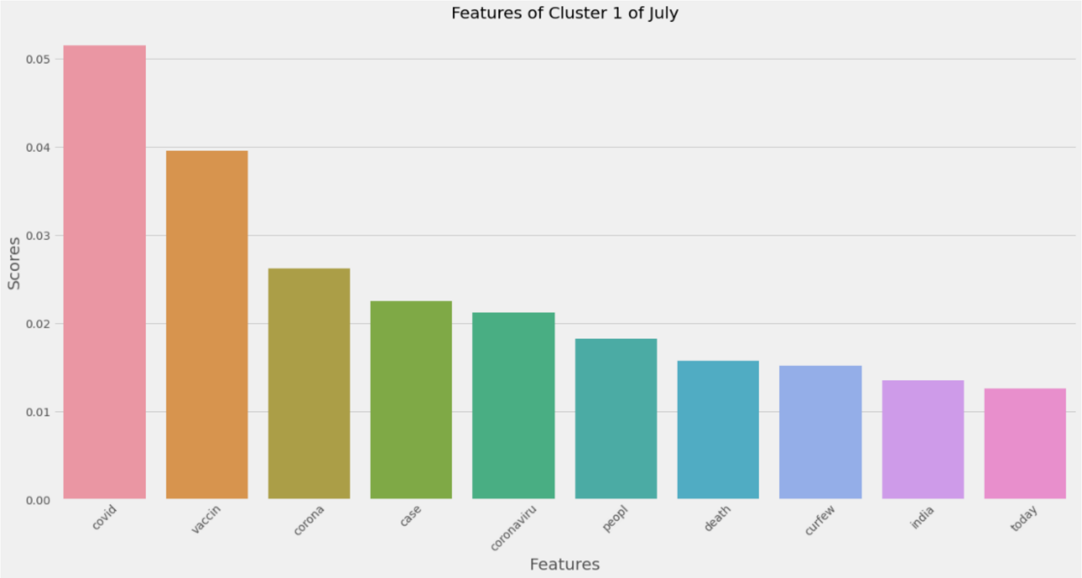


*Figure 23:Top 10 features of Cluster1 of July*

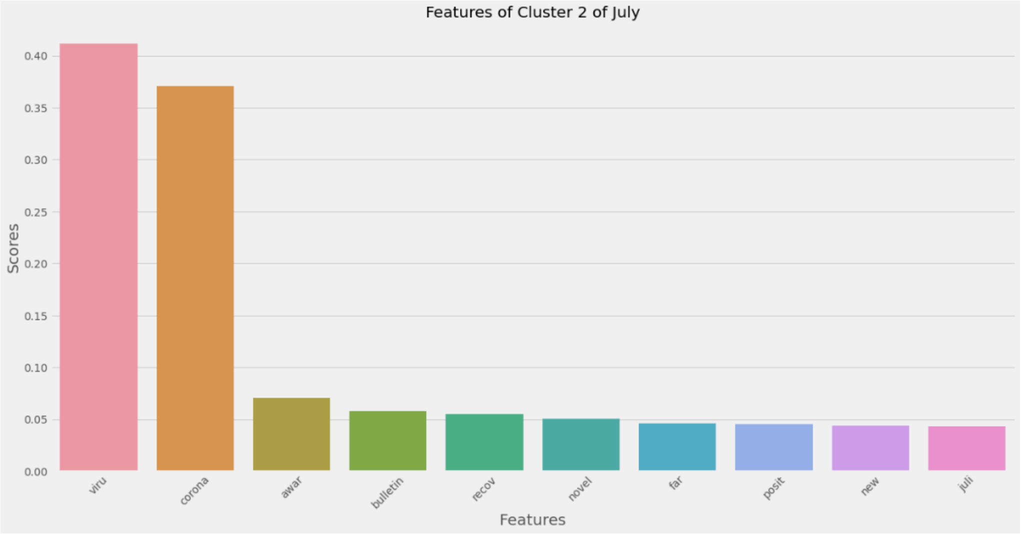


*Figure 24:Top 10 features of Cluster2 of July*

We can see ‘virus’ is forming a different cluster. And in 2nd cluster it is almost similar to that of March. ‘vaccine’ and ‘corona’ isn’t forming a different cluster. In 2nd cluster, 29 out of 30 most used buzzwords are present. So, we can say the clustering of buzzwords is good here.

According to India Situation Report by WHO on 28th July 2021, highest number of cases in a day were 4,14,188; reported on 7th May 2021. Since then, there has been a rapid decline in daily cases however daily cases are plateauing around 30-40,000 for the last several weeks. Cases per million is 22,126. As, cases were decreasing people stopped taking precautions like wearing masks, washing hands etc. So, some people started spreading awareness through twitter. That’s how ‘aware’ became a new buzzword.

❖ ***August***:

For clusters no. 2, the average silhouette score is: 0.76454036655999966
For clusters no. 3, the average silhouette score is: 0.433915537746124446
For clusters no. 4, the average silhouette score is: 0.459644111784331 07

The number of clusters here is 2 as the silhouette score of the 2nd cluster is maximum.

Top 3 Features of 1st cluster: case, report, today.
Top 3 Features of 2nd cluster: covid, vaccine, corona.

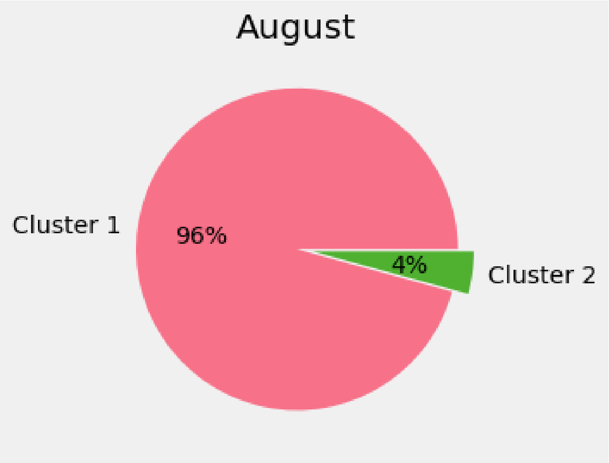


*Figure 25:Density of Tweets in Cluster in August*

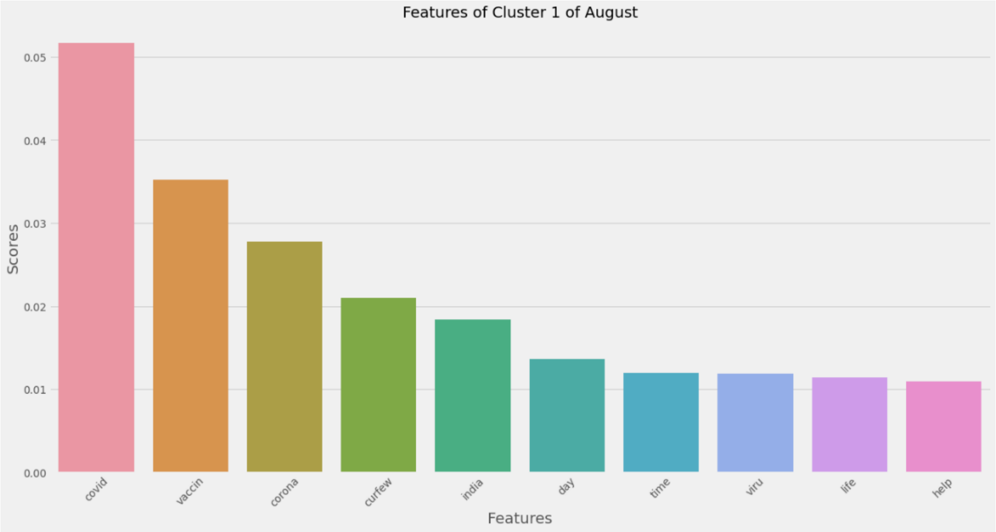


*Figure 26:Top 10 features of Cluster1 of August*

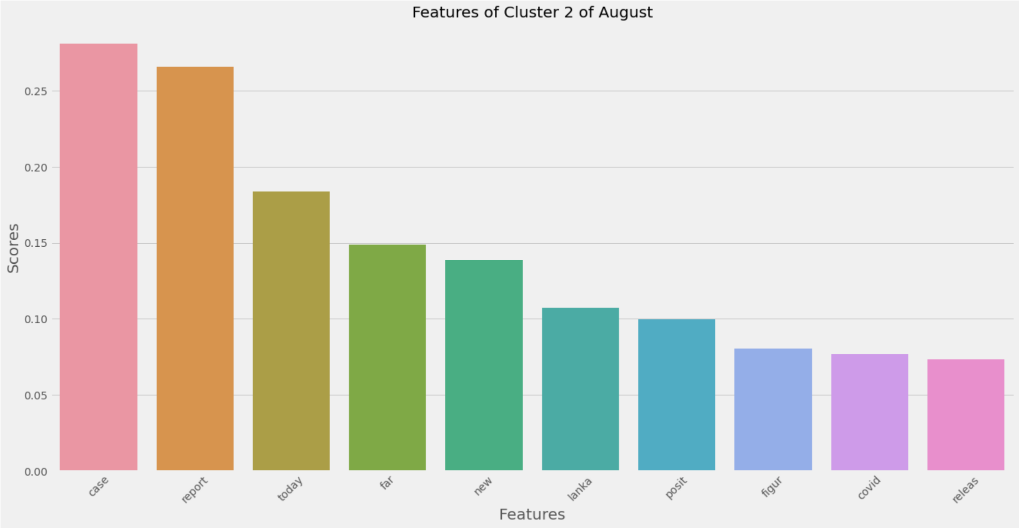


*Figure 27:Top 10 features of Cluster2 of August*

We can see here ‘case’ is forming different cluster and ‘covid’, ‘vaccine’, ‘corona’ has again together formed a cluster. In 2nd cluster, the clustering is good, as 25 out of 30 most frequently used buzzwords are present.

According to India Situation Report by WHO on 28th August 2021, for the last several weeks, daily cases were plateauing around 25- 40,000. Cases per million is 22,854. Various organizations and People sharing daily updates of COVID Cases, availability of resources and so on, made ‘covid’, ‘report’, ‘today’ etc. frequently used words of August.

**TABLE OF CLUSTERS OF MONTHS MARCH’21-AUGUST’21**

| MONTHS | MARCH’ 21 | APRIL’ 21 | MAY’ 21 | JUNE’ 21 | JULY’ 21 | AUGUST’ 21 |
|---|---|---|---|---|---|---|
| NO. OF CLUSTER | 2 | 3 | 2 | 3 | 2 | 2 |
| MOST IMPORTANT FEATURES OF CLUSTER 1 | COVID | CASE | COVID | COVID | COVID | COVID |
| MOST IMPORTANT FEATURES OF CLUSTER 2 | CASE | CORONA | VACCIN | COVID | VIRU | CASE |
| MOST IMPORTANT FEATURES OF CLUSTER 3 | - | COVID | CORONA | VACCIN | - | - |

## 6. Conclusion:

Exploratory Data Analysis is an integral approach towards data analysis in order to drive valid assumptions and data results. It comes in handy whenever it is needed to gain new insights into a massive quantity of data sets. In this aspect, EDA can be beneficial for fields such as research and development, engineering, and data science. Hence, In today’s age, with access to advanced computing power along with the support of modern analytics. EDA can be a stimulating and engaging experience for researchers or data scientists to explore unexpected value in a massive quantity of complex data sets.

In this report, after analysing the datasets, i.e., collections of tweets based on few keywords related to COVID and observing, we can assume from the various relations between the words the various stages of Covid in the given period of time, and can find the number of times various important organisations or peoples who were tagged and the hashtags used for spreading of awareness, seeking help or to update people about cases, recovery rate and so on.

# 7. References:

[1] X. Han, J. Wang, M. Zhang, and X. Wang, “Using social media to mine and analyse public opinion related to COVID-19 in China,” Int., J. Environ. Res. Public Health, vol. 17, no. 8, p. 2788, Apr. 2020.

[2] Usman Naseem, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim, Member, IEEE, “COVID Sentiment: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis”, IEEE Transactions on Computational Social Systems, Vol. 8, No. 4, August 2021.

[3] Tanmay Vijay, Ayan Chawla, Balan Dhanka, Purnendu Karmakar, “Sentiment Analysis on COVID-19 Twitter Data”, IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE 2020 (IEEE Record#51050).