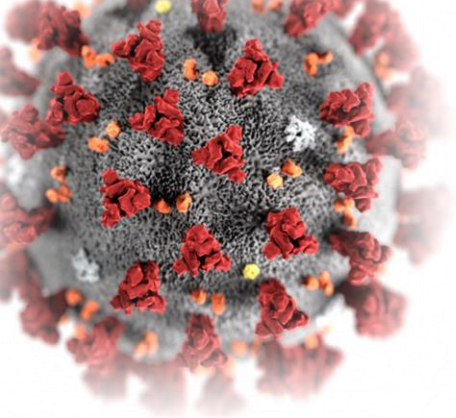


PROJECT REPORT

Covid 19 Twitter Data Analysis

By : Abhishek Garg

PROBLEM STATEMENT :

To build a Twitter trend analyzer that will analyze a set of tweets using NLP and text-processing techniques. The trend analyzer will work on a given set of tweets, seeded on COVID19 / CORONA.

PRIMARY TASKS :

1) A tag cloud depicting what topics / Word was being talked about on Twitter .

2) Which hashtag trended (Hashtags are words or phrases beginning with # eg #COVID) .

3) Which Twitter Handler dominated the conversation on Twitter .

ADDITIONAL TASKS PERFORMED :

4) Exploring which tweet is most retweeted.

5) Analysing how the number of tweets changes as months are passed.

6) Exploring weekday based trend of tweets.

7) SENTIMENT ANALYSIS :Performing sentiment analysis after text cleaning using **VADER Sentiment Analysis.**

8) Analyzing how the sentiment of tweets changed as months passed .

9) Analyzing how the sentiment of tweets changed with weekdays .

10) Visualzing positive and negative tweets by separate wordclouds for each category respectively .

ABSTRACT:

The novel coronavirus took the entire world by a storm . People from every part of the world were severely affected by the impact of this virus . There was a situation of panic everywhere . Globally it claimed around 2.5 million lives and more than 100 million people were infected by it .Businesses were shut down and people were locked down in their houses so naturally their enagagement on social media grew drastically.People were using social media platforms to communicate with each other , share their thoughts and opinions and even to help those who were severely affected by the virus . Twitter is also one such very famous platform which is used by people to express their views on various topics.

The pandemic also trended widely on twitter and large number of tweets were made each day . This project deals with analysis of these tweets and extract useful trends and insights from them.We will be deep diving into vaious details about the data ranging from popular hashtags which trended to exploring the sentiments of these tweets and will be coming up with various interesting and concise visualizations for better interpretation of the data and then finally making our observations.

TASK 1 - MAKING A WORDCLOUD SHOWING WHICH WORDS WERE BEING TALKED ABOUT .

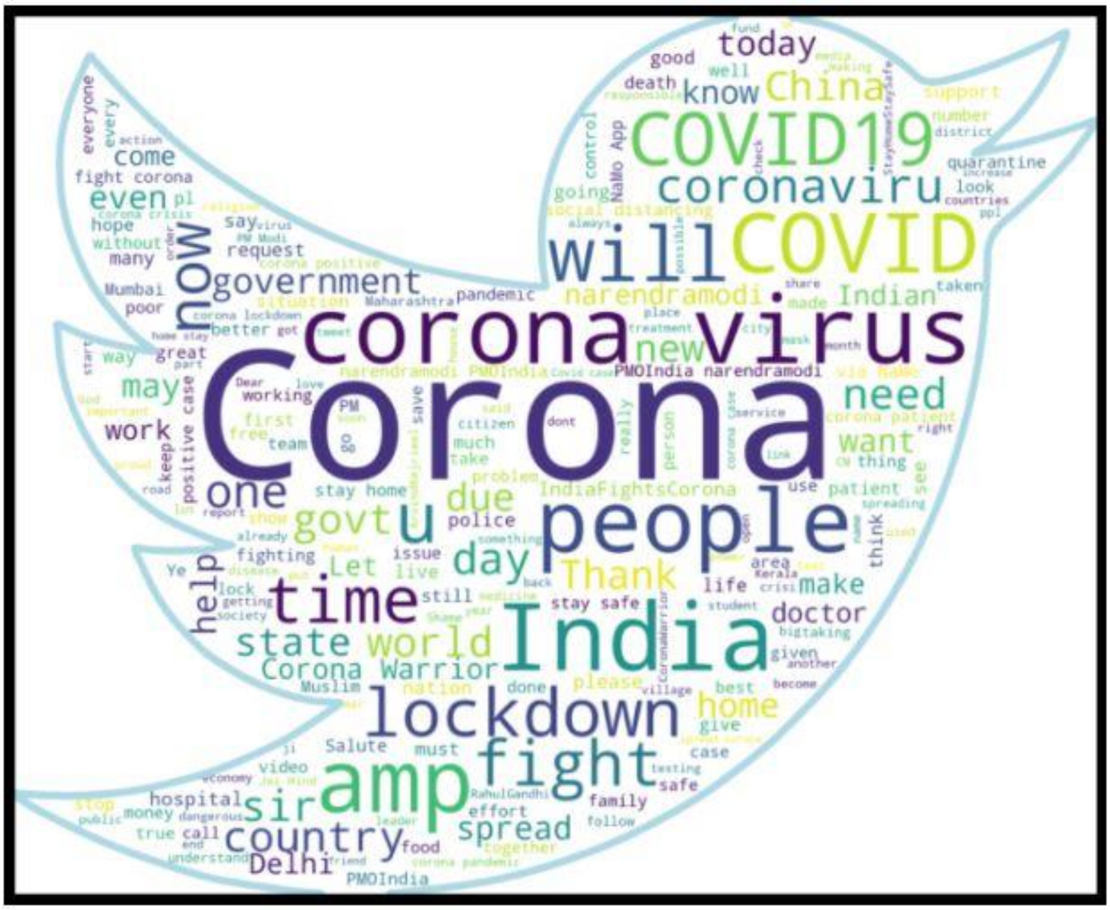
STEP 1.a) : IMPORTING LIBRARIES

The first thing we do is import some necessary libararies like pandas , wordcloud , maplotlib , seaborn , nltk , re , numpy and PIL .Matplotlib for data visualization, nltk for NLP, padas and numpy for data exploration, PIL for image processing and wordcloud for word clouds.

STEP 1.b) : IMPORTING THE DATASET

We import the dataset which has 44179 rows 19 columns contaning various information of many tweets that revolves around the covid 19 pandemic.

STEP 1.c) : CREATING A WORDCLOUD WITH ADDED STOPWORDS AND A CUSTOM MASK



Since these tweets are made about the coronavirus pandemic therefore we can see that words like Corona , virus , Lockdown , spread , Covid , China etc pops up in the wordcloud .

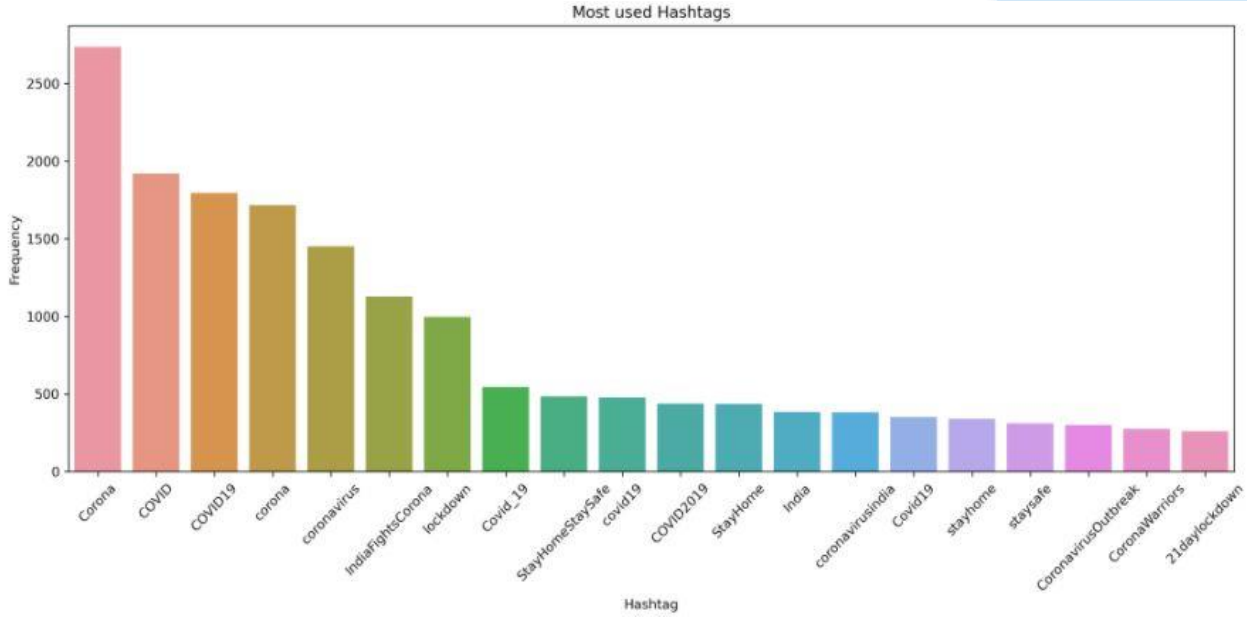
TASK 2 - SHOWING THE HASHTAGS WHICH TRENDED

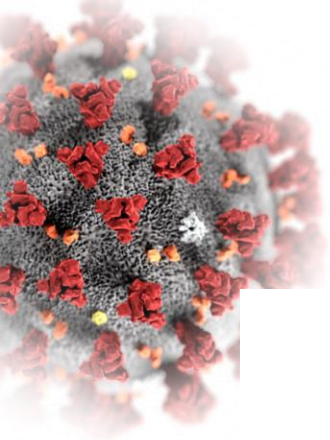
STEP 2.a) : CREATING A DATAFRAME CONTAINING TOP 20 HASHTAGS AND THEIR COUNTS

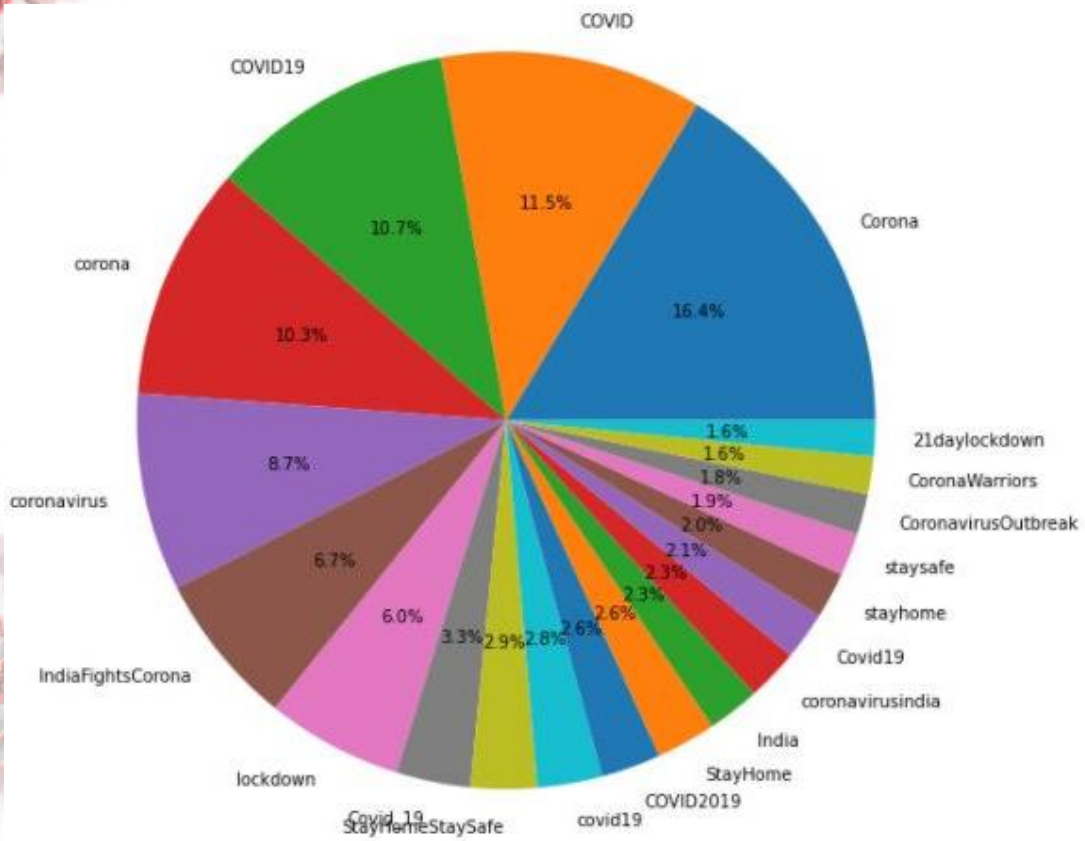
	Hashtag	Frequency
0	#Corona	2736
1	#COVID	1921
2	#COVID19	1797
3	#corona	1715
4	#coronavirus	1452
5	#IndiaFightsCorona	1128
6	#lockdown	996
7	#Covid_19	545
8	#StayHomeStaySafe	484
9	#covid19	476
10	#COVID2019	438
11	#StayHome	437
12	#India	385
13	#coronavirusindia	381
14	#Covid19	352
15	#stayhome	339
16	#staysafe	313
17	#CoronavirusOutbreak	299
18	#CoronaWarriors	275
19	#21daylockdown	261

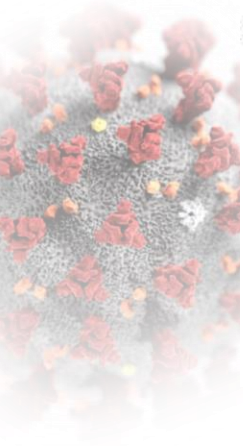
#Corona is the top trending hashtag with **2736** mentions followed by others as shown in above table.Here we know that the tweets are about the coronavirus pandemic so the hashtags which trended also contains other related tags like COVID , lockdown , Stay home , Stay safe and so on .

STEP 2.b) : VISUALIZING THE POPULARITY OF ABOVE HASHTAGS WITH A BARGRAPH AND PIE CHART FOR BETTER INTERPRETATION






COVID
COVID19
11.5%
10.7%
Corona
corona
16.4%
10.3%
1.6%
21daylockdown
8.7%
1.6%
CoronaWarriors
1.8%
coronavirus
1.9%
CoronavirusOutbreak
2.0%
2.1%
staysafe
6.7%
2.3%
2.3%
stayhome
6.0%
2.6%
3.3%
2.9%
2.8%
2.6%
Covid19
IndiaFightsCorona
coronavirusindia
India
StayHome
lockdown
COVID2019
Covid_19
covid19
StayHomeStaySafe



TASK 3 : ANALYZING WHICH TWITTER HANDLERS WERE THE CENTRE OF ATTENTION .

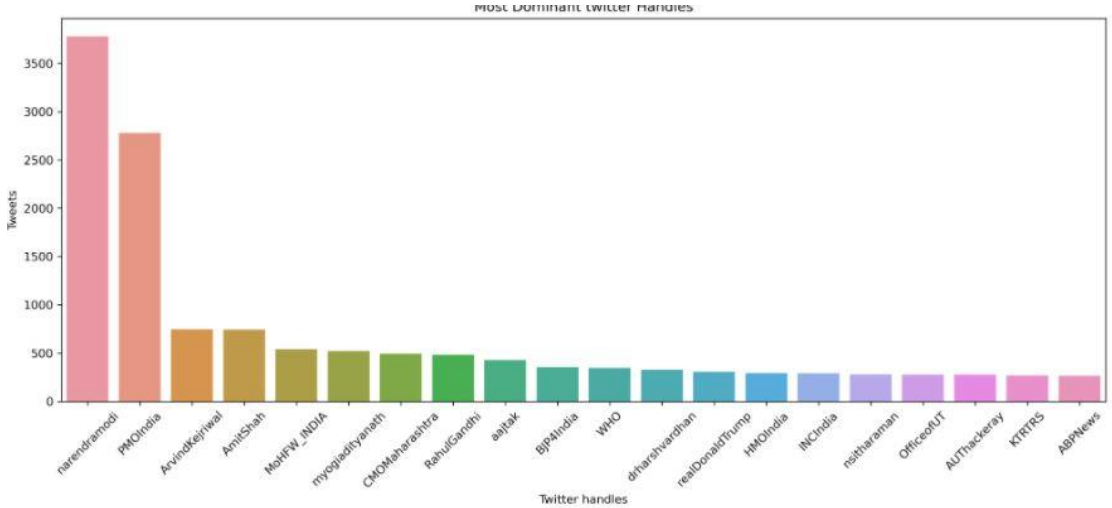
People on twitter use the **'@'** symbol to tag or mention in their tweet so we'll explore which all users were mentioned the highest number of times.

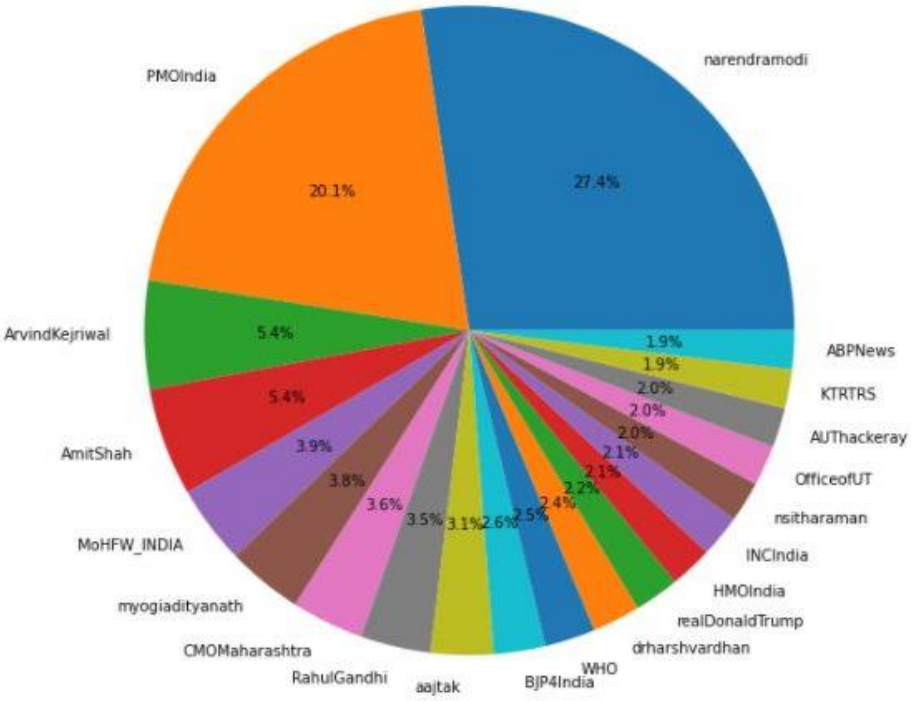
STEP 3.a) CREATING A DATAFRAME CONTAINING TOP 20 USERS WHO WERE MENTIONED AND THEIR COUNTS.

	Twitter handles	Tweets
0	@narendramodi	3778
1	@PMOIndia	2779
2	@ArvindKejriwal	746
3	@AmitShah	742
4	@MoHFW_INDIA	541
5	@myogiadityanath	521
6	@CMOMaharashtra	496
7	@RahulGandhi	482
8	@aajtak	429
9	@BJP4India	355
10	@WHO	347
11	@drharshvardhan	329
12	@realDonaldTrump	309
13	@HMOIndia	293
14	@INCIndia	290
15	@nsitharaman	279
16	@OfficeofUT	275
17	@AUThackeray	275
18	@KTRTRS	267
19	@ABPNews	265

This table shows the users who were mentioned along with their respective counts. The most mentioned user is **@narendramodi** with **3778** mentions . We can observe that most of the mentions contains leaders like Narendra modi , Arvind kejriwal , Amit shah etc. and that is because people were in a state of panic and were tagging these powerful leaders and ministers in hope of conveying their message to them so that appropriate steps can be taken by administration to contain the Covid-19 crisis.

STEP 3.b) VISUALIZING WHICH TWITTER HANDLE DOMINATED THE CONVERSATIONS WITH BARGRAPH AND PIE CHART FOR BETTER INTERPRETATION.





The above plots shows that most of the tweets made by the people mentioned **Mr.Narendra Modi** , the prime minister of india .

TASK 4 : EXPLORING THE MOST RETWEETED TWEET.

```
max(df['retweet_count'])
```

```
22549
```

```
df.iloc[df['retweet_count'].argmax()]['text']
```

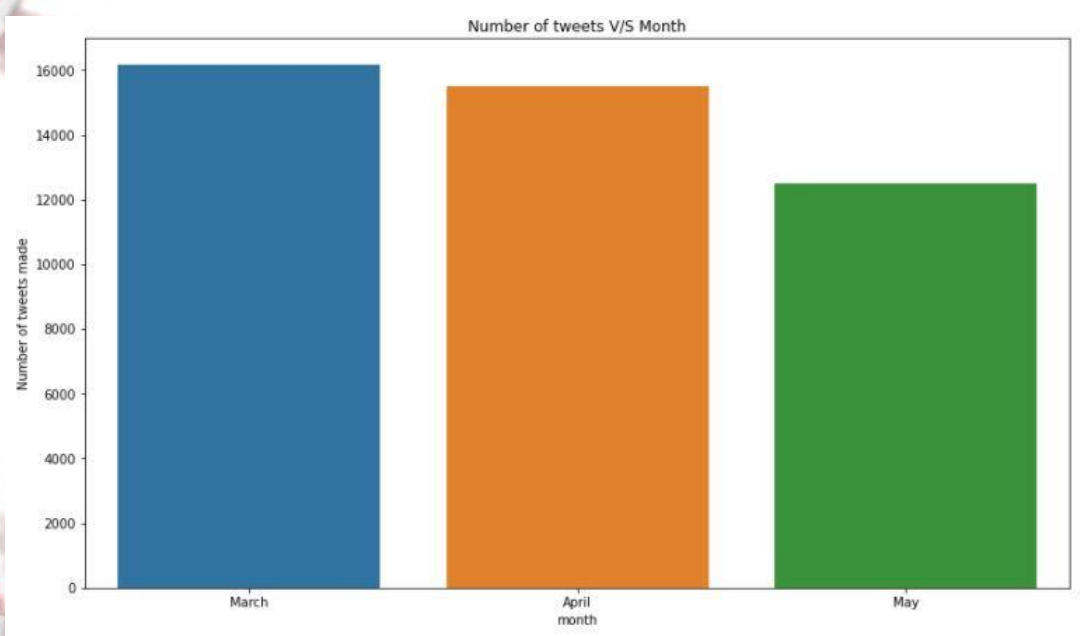
```
'time bit help defeat covid19 pledging 52 lakh fight 31 lakh pm cares fund amp 21 lakh cm disaster relief fund please bit jai h
ind stayhomeindia'
```

```
df.iloc[df['retweet_count'].argmax()]['user_screen_name']
```

```
'ImRaina'
```

The most retweeted tweet is retweeted **22,549** times and it is by famous indian cricketer **Suresh Raina** where he is talking about making monetary donations to the help funds estabilished by the government fight the coronavirus crisis.

TASK 5 : ANALYZING HOW THE TREND OF TWEETS CHANGES WITH MONTHS.



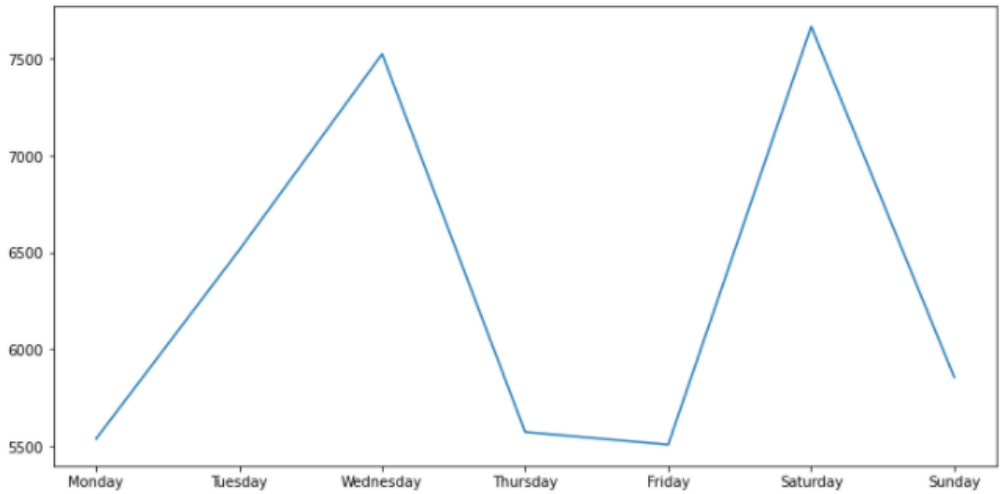
The above plot shows how the number of tweets made changes as months passed by . We can see that the number gradually deacreases.

The exact number of tweets made in each month are given below :

```
df['month'].value_counts()
```

```
March    16188
April    15505
May      12486
Name: month, dtype: int64
```

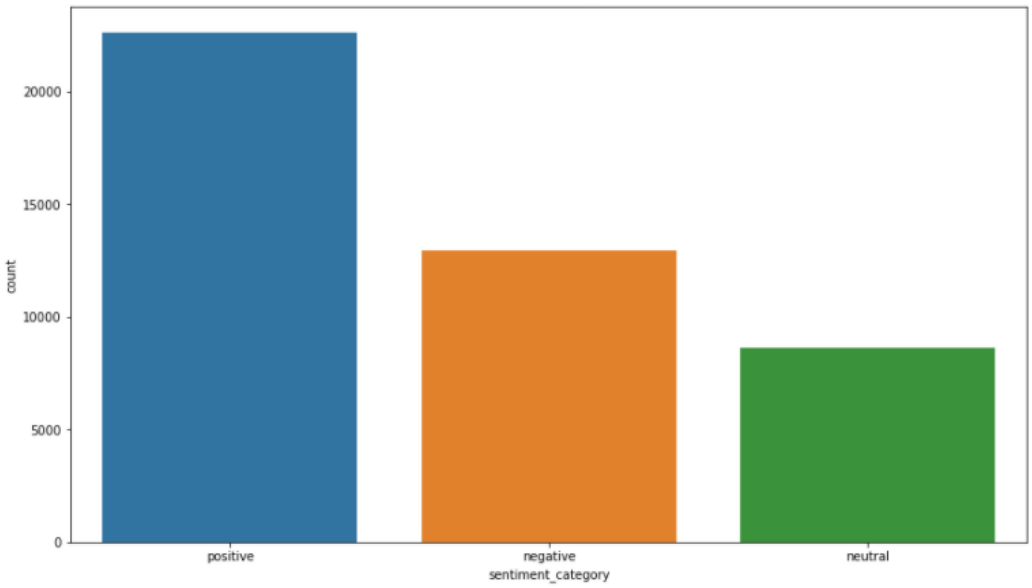
TASK 6 : EXPLORING WEEKDAY BASED TRENDS OF TWEETS .



Most number of tweets were made on **saturdays** followed by **wednesdays .** On the remaining days of the week almost similar number of tweets were made .

TASK 7 : PERFORMING SENTIMENT ANALYSIS ON THE TWEETS .

Sentiment Analysis refers to analyzing a tweet and then determining the possible emotion behind the tweet and finally classifying it as a positive , negative or neutral comment . For our task we have used VADER Sentiment Analyzer .



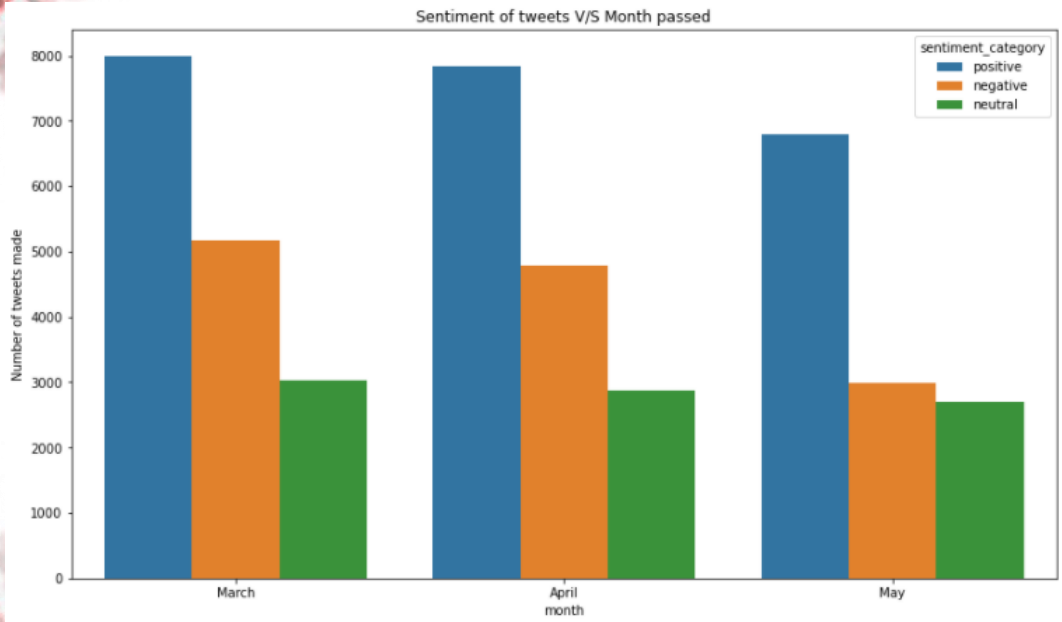
As we can see most number of tweets are **positive** , which is a good thing to see as it shows that even in the times of such acute crisis people were still able to see the bright side of the things . The exact figures are as follows :

POSITIVE TWEETS : **22628**
NEGATIVE TWEETS : **12949**
NEUTRAL TWEETS : **8602**

```
print("Percentage of positive tweets : " , int(22628/len(df)*100) , "%")
print("Percentage of negative tweets : " , int(12949/len(df)*100) , "%")
print("Percentage of neutral tweets : " , int(8602/len(df)*100) , "%")
```

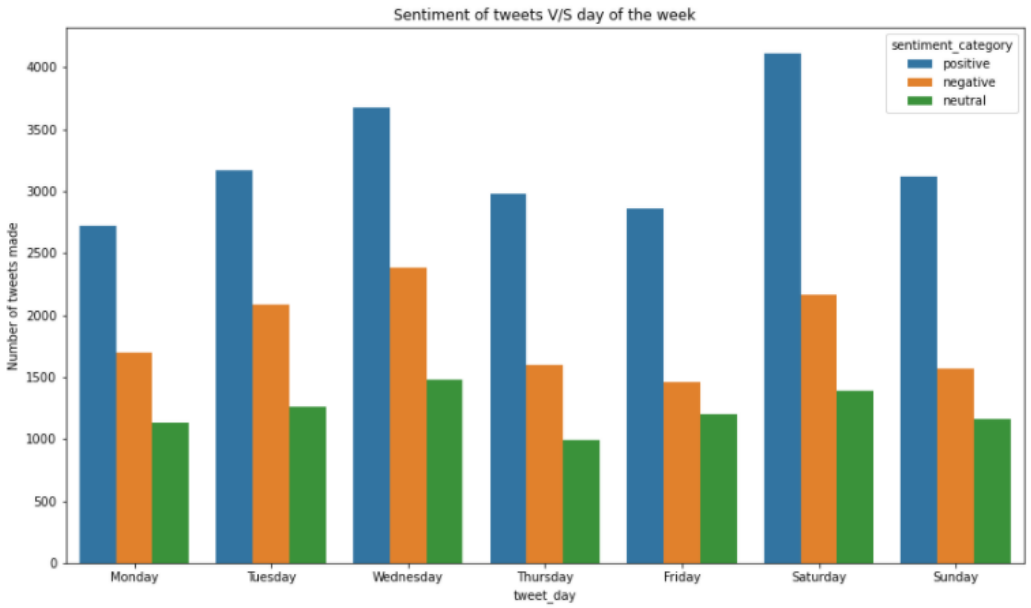
```
Percentage of positive tweets :  51 %
Percentage of negative tweets :  29 %
Percentage of neutral tweets :  19 %
```

TASK 8 : ANALYZING HOW SENTIMENTS OF TWEETS CHANGED WITH PASSING MONTHS .

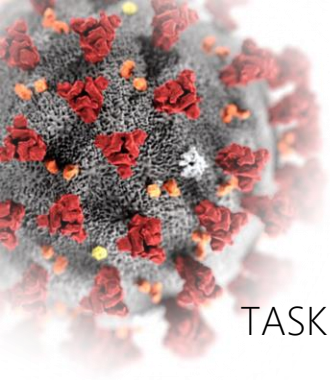


It can be seen that ***as the months passed*** the overall number of tweets also ***decreased*** . The interesting thing to notice is there is a sudden decrease in the number of negative tweets as we go from April to May which again is a good sign . Apart from that there is no such unusual trend followed by positive category tweets and the number of neutral tweets remains almost constant throughout there is a small change only .

TASK 9 : ANALYZING HOW THE SENTIMENT OF TWEETS CHANGES WITH WEEKDAYS



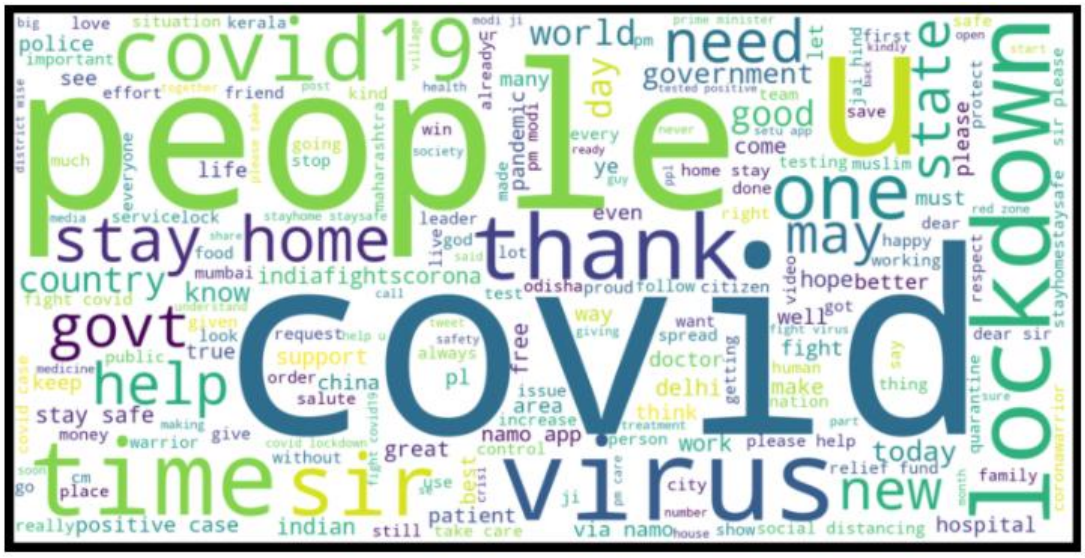
From the graph we can infer that most all the kinds of tweets i.e. positive , negative and neutral follow same pattern . Most number of positive tweets are made on saturdays and this can also be linked to the fact that the most number of tweets are also made on saturdays while the most number of negative tweets are made on wednesdays .



TASK 10 : GENERATING WORDCLOUDS FOR POSITIVE AND NEGATIVE TWEETS .

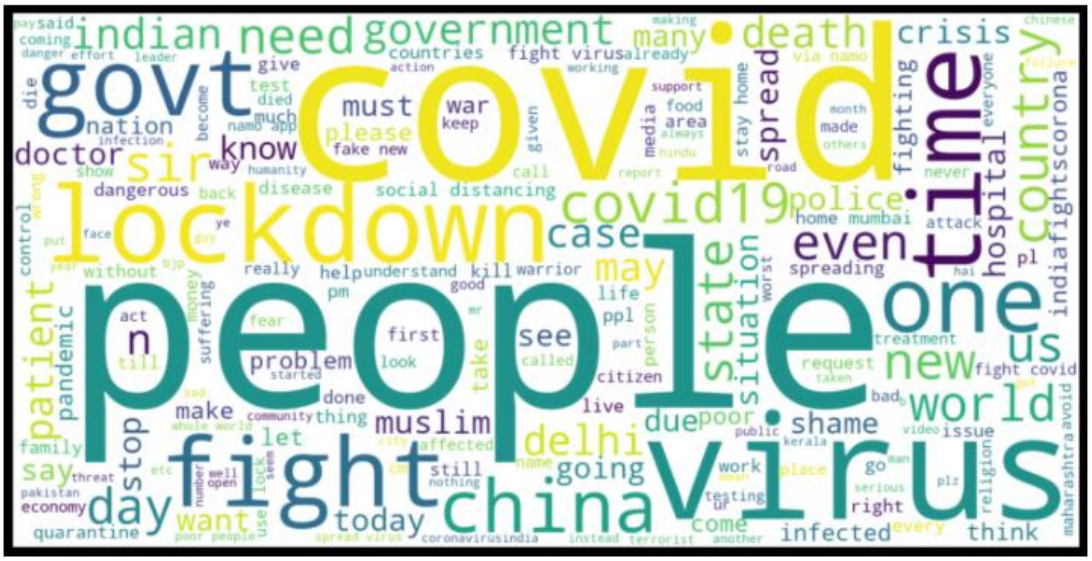
For visualizing the positive and negative tweets in a better way we will separate them into two categories and make separate wordclouds for each category .

STEP 10.a) CREATING POSITIVE TWEETS WORDCLOUD



We can see words like ***thank , safe , hope , better , help*** etc can be seen in the positive tweets wordcloud .

STEP 10.b) CREATING NEGATIVE TWEETS WORDCLOUD



Here words like ***fight , virus , china , kill , problem , poor*** pops up .

OBSERVATIONS

To summarize the analysis of the tweets about the pandemic following observations can be made :

- Most popular hashtag : **#Corona** with **2736** tweets .
- Most popular user : **@narendamodi** who were mentioned **3778** times.
- The most retweeted tweet was by **Suresh Raina** which was retweeted **22,549** times.
- The number of tweets gradually decreases with passing months .This could be due to the fact that as time passed people were getting more used to the coronavirus situation unlike the beginning when there was a feeling of panic among people also we know that in the beginning there was complete lockdown so more people were at their homes and hence number of people using social media was high however slowly businesses were reopening so more people were getting back to their work .
- Maximum number of tweets were made on saturdays followed by wednesdays .
- The sentiment analysis of the tweets shows that **51 %** percent of tweets were of **positive** emotion which shows that the *majority emotion of people towards the entire COVID situation situation was good* .
- The sentiment analysis of tweets with weekdays shows that there is an *increase* in the number of *negative tweets on saturdays* .