Sentiment Analysis of COVID-19 Tweets – Visualization Dashboard

Project Report

INDEX

INTRODUCTION

1.1 Overview

The Covid-19 has been endangering the public safety and the government has been imposing lockdowns to tackle it. But, the lockdown has its own effects on the public and they have their own reactions and thoughts on the same. Social media platforms such as Twitter act as a host to honest reviews by the public, hence it can be used for a better collection of the public opinion and come up with a better solution.
Hence, our aim is to develop a Twitter Sentiment Analysis Visualization dashboard, depicting various public sentiments (on twitter) in India during the lockdown.
The sentiments of the Indian public after the announcement of lockdown to be analyzed with the relevant #tags on twitter and a predictive analytics model to be built to understand the public behavior if the lockdown is further extended.

1.2 Purpose

Fighting the coronavirus not only needs the guidance from the government but also a positive attitude from the public. Our analysis provides a potential approach to reveal the public's sentiment status and help institutions respond timely to it.The purpose of this project is to analyze tweets for the lockdown due to covid-19 and understand the sentiments of people to know how are they taking this lockdown personally and how would they react if this continues.

This study would help organizations to understand one's mental state during the initial stage of lockdown to the state they currently possess, also better understand the situation and its impact on people more accurately for which they need not have any prior knowledge of coding. This would enable organizations to adjust accordingly before occurrence of any serious damage and help them come up with better ideas to enhance and promote their business or make better policies.

The above solution can be further applied to analyse racism, hatred, bullying like sentiments present in the tweets, so that they can be removed and prevented from spreading such sentiments further and cause harm to the feelings of a community or basically other humans.These visualizations would help authorities to make appropriate decisions based on current situations for different domains.

With more granular data such as geographic data, demographic information, and so on, further insights can be generated, such as public sentiment monitoring the hardest-hit areas. With a more specific target, the analysis would be more valuable for institutions or governments to take action.

LITERATURE SURVEY

2.1 Existing problem

As the Covid-19 cases keep increasing asymptotically, the need for the extension of the lockdown may arise as well.

For the same, the public opinion and consequences of such a drastic decision need to be analyzed. Hence, arises the need for a smart solution that analyses the public opinion on the current scenario and helps predict further consequences if the lockdown is extended.

Many different approaches can be used to implement the same but IBM Watson provides the smartest and the fastest approach for the analysis.

2.2 Proposed solution

I. Data Collection

Used kaggle to extract the dataset containing tweets https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april for March 29, 2020. It contains 22 columns namely 'status_id', 'user_id', 'created_at', 'screen_name', 'text', 'source', 'reply_to_status_id', 'reply_to_user_id', 'reply_to_screen_name', 'is_quote', 'is_retweet', favourites_count', 'retweet_count, 'country_code', 'place_full_name', 'place_type', 'followers_count', 'friends_count', 'account_lang', 'account_created_at', 'verified', & 'lang' and 564141 rows.
Below is shown the snapsht of the dataset.

	status_id	user_id	created_at	screen_name	text	source	reply_to_status_id	reply_to_user_id	reply_to_screen_name	is_quote	...	retweet_count	country_co
0	1244051646071611394	860252856829587457	2020-03-29T00:00:00Z	IMSS_SanLuis	Ante cualquier enfermedad respiratoria, no te ...	TweetDeck	NaN	NaN	NaN	False	...	0	Na
1	1244051645039706112	1125933654943895553	2020-03-29T00:00:00Z	intrac_ccs	#ATENCIÓN En el Terminal Nuevo Circo se implem...	TweetDeck	NaN	NaN	NaN	False	...	1	Na
2	1244051645975191557	80943559	2020-03-29T00:00:00Z	rlieving	“People are just storing up. They are staying ...	TweetDeck	NaN	NaN	NaN	False	...	0	Na
3	1244051646750928897	817072420947247104	2020-03-29T00:00:00Z	Tu_IMSS_Coah	Si empezaste a trabajar, necesitas dar de alta...	TweetDeck	NaN	NaN	NaN	False	...	0	Na
4	1244051647032102914	788863557349670913	2020-03-29T00:00:00Z	Tabasco_IMSS	Una sociedad informada está mejor preparada an...	TweetDeck	NaN	NaN	NaN	False	...	0	Na
5	1244051645710897155	132225222	2020-03-29T00:00:00Z	SSalud_mx	¡#Infórmate! #ConferenciaDePrensa sobre el #Co...	TweetDeck	NaN	NaN	NaN	False	...	49	Na

In order to extract **latest tweets**, the twitter developer account is linked with notebook by

providing concerned credentials. After authentication, latest tweets are extracted for that instance using hashtag '#lockdownIndia'.

II. Data Pre-processing and cleaning

The tweets specific to India are extracted and the following attributes are dropped from the dataset, namely, 'status_id', 'country_code', 'user_id', 'screen_name', 'source', 'reply_to_status_id', 'reply_to_user_id', 'is_retweet', 'place_full_name', 'place_type', 'reply_to_screen_name', 'is_quote', 'followers_count', 'friends_count', 'account_lang', 'account_created_at', 'lang', and 'verified'.

The resultant dataset contains 4 columns:

- created_at
- text
- favourites_count
- retweet_count

and 1574 rows.

Below is shown the snapshot of the same.

```
(1574, 4)
```

	created_at	text	favourites_count	retweet_count
0	2020-03-29T00:04:06Z	"#Covid19: #SocialDistancing the Indian way" ...	10374	0
1	2020-03-29T00:04:55Z	Possible case of human to animal transmission ...	22880	0
2	2020-03-29T00:07:20Z	#Coronavirus: #US eclipses 120,000 confirmed c...	12968	6
3	2020-03-29T00:09:14Z	The bitter truth 😔\n#covid_19 #covid19india #l...	6	0
4	2020-03-29T00:13:20Z	#CoronavirusOutbreak All Churches shd realize...	38	0

Next, text attribute which contains tweets is being cleaned by removing all stopwords, symbols, numbers and special character. Now, the output of pre-processed tweets in the text attribute becomes more meaningful and readable when compared to the collected tweets.

```
0              covid19 socialdistancing the indian way
1    possible case of human to animal transmission ...
2    coronavirus us eclipses 120 000 confirmed case...
3    the bitter truth covid 19 covid19india lockdow...
4    coronavirusoutbreak all churches shd realize t...
Name: text, dtype: object
```

Analyzed the dataset to understand its structure more deeply and come up with the patterns which were found to be repeating with time.

For **Latest Tweets**, only the required atributes are selected, namely, id, created_at, source, favourite_count, retweet_count. These attributes are changed to ID, Date, Source, Likes and RTs respectively and new attribute named *len* to store the length of each tweet is added in database.

	Tweets	len	ID	Date	Source	Likes	RTs
0	When will this lockdown end \n#COVID19 #lockdo...	86	1283076566767988736	2020-07-14 16:31:06	Twitter for Android	0	0
1	If #Deewar was filmed in 2020 🙂 \n#lockdown #l...	81	1283071767586594816	2020-07-14 16:12:02	Twitter for Android	1	0
2	#कडवासच #Bittertruth \nWith due credits to car...	111	1283068511263186944	2020-07-14 15:59:06	Twitter for Android	0	0
3	𝙋𝙤𝙩𝙖𝙩𝙤 𝙁𝙚𝙣𝙪𝙜𝙧𝙚𝙚𝙠 𝙇𝙚𝙖𝙫𝙚𝙨 𝘽𝙪𝙣𝙨\n\n#lockdownindia...	137	1283060781047668736	2020-07-14 15:28:22	Twitter for Android	1	1
4	#thekingofkhansra #changez #changezandme #lock...	136	1283057041184100354	2020-07-14 15:13:31	Instagram	0	0
5	RT @LavekarBharati: Our Volunteers Distributin...	143	1283055271204605953	2020-07-14 15:06:29	Twitter for iPhone	0	9
6	RT @ShreyaR03426090: please stay at home and a...	140	1283053826774269959	2020-07-14 15:00:44	Twitter for Android	0	1
7	RT @NagarajuGujjeti: मंगलवार, 14 जुलाई 2020 ऑन...	140	1283052817389953028	2020-07-14 14:56:44	Twitter Web App	0	2
8	RT @kakoligdastidar: If Europe south Korea New...	140	1283052037681434625	2020-07-14 14:53:38	Twitter Web App	0	45
9	RT @Scimitar_SS: He was so distraught that he ...	139	1283045258918531075	2020-07-14 14:26:42	Twitter for Android	0	1

III. Sentiment Analysis

i. Before Second lockdown tweets

Sentiment analysis is being done using one of the NLP library 'TextBlob' under IBM Watson Notebook. We created a new dataset which stores sentiments (positive, negative or neutral) and polarity corresponding to all tweets as shown below.

	created_at	text	favourites_count	retweet_count	sentiment	polarity
0	0.066667	covid19 socialdistancing the indian way	10374	0	neutral	0
1	0.066667	possible case of human to animal transmission ...	22880	0	negative	-0.23125
2	0.116667	coronavirus us eclipses 120 000 confirmed case...	12968	6	positive	0.4
3	0.150000	the bitter truth covid 19 covid19india lockdow...	6	0	negative	-0.1
4	0.216667	coronavirusoutbreak all churches shd realize t...	38	0	positive	0.1375

ii. Latest tweets

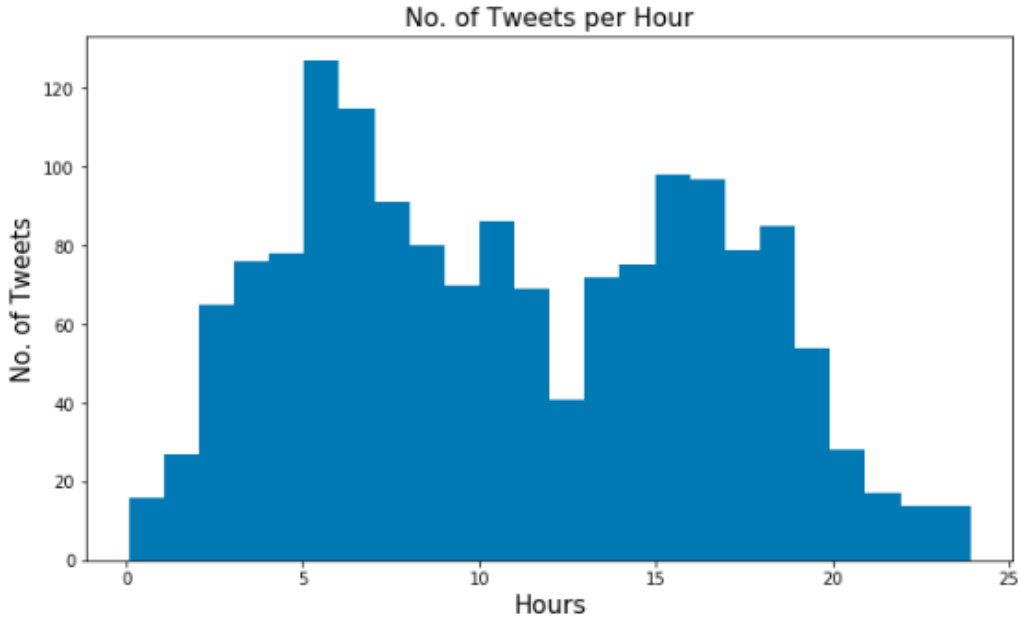
Using 'Tweepy', we have extracted the latest tweets based on hashtag #lockdownIndia and applied NLP library 'TextBlob' under IBM Watson Notebook to get sentiments corresponding to fetched tweets.

	Tweets	len	ID	Date	Source	Likes	RTs	sentiment	polarity
0	He was so distraught that he didn't even accep...	140	1283022106439987201	2020-07-14 12:54:42	Twitter for Android	0	0	negative	-0.2
1	RT @GauravJain_9: #Againstcorona #lockdownindi...	140	1283018315963576325	2020-07-14 12:39:38	Twitter for Android	0	2	neutral	0
2	RT @GauravJain_9: #LockdownIndia #Mingrantwork...	117	1283018289849856002	2020-07-14 12:39:32	Twitter for Android	0	3	neutral	0
3	RT @HSnewsLive: #coronavirus के कारण बिहार में...	140	1283017858860019713	2020-07-14 12:37:49	Twitter for Android	0	2	neutral	0
4	RT @kakoligdastidar: If Europe south Korea New...	140	1283017789234573312	2020-07-14 12:37:32	Twitter for Android	0	44	positive	0.136364

IV. Visualization

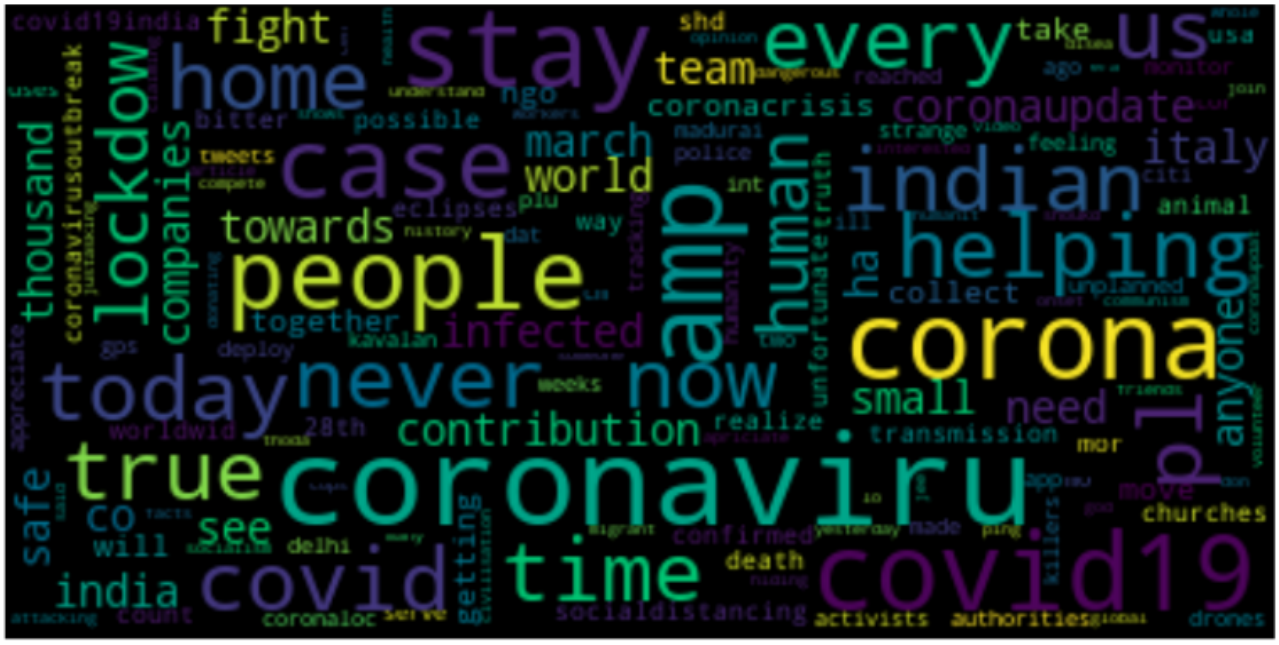
Libraries like matplotlib and seaborn are used to visualize the results. Following are the results to better understand the dataset and analysis to find the patterns in tweets.

1. Number of Tweets per Hour



The graph shows no. of tweets per hour and depicts that people are found active on twitter for an average of 5 to 7 hours. With addition to this, maximum no. of tweets are found to be above 120 in number.

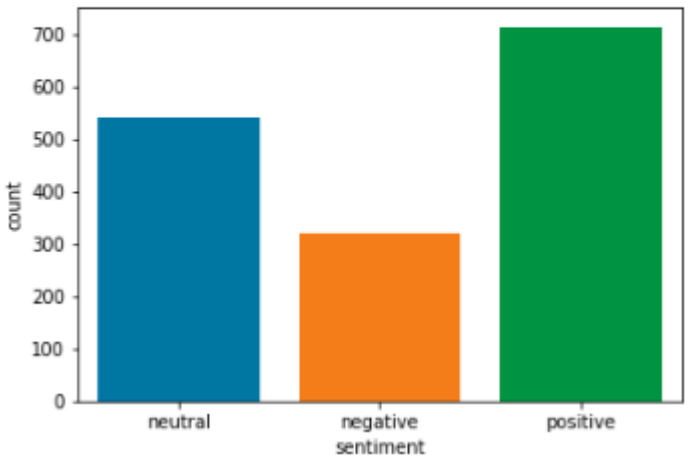
2. Wordcloud



This shows frequent occuring words in our tweets. Some of the words are corona,India,stay, home, lockdown , etc.

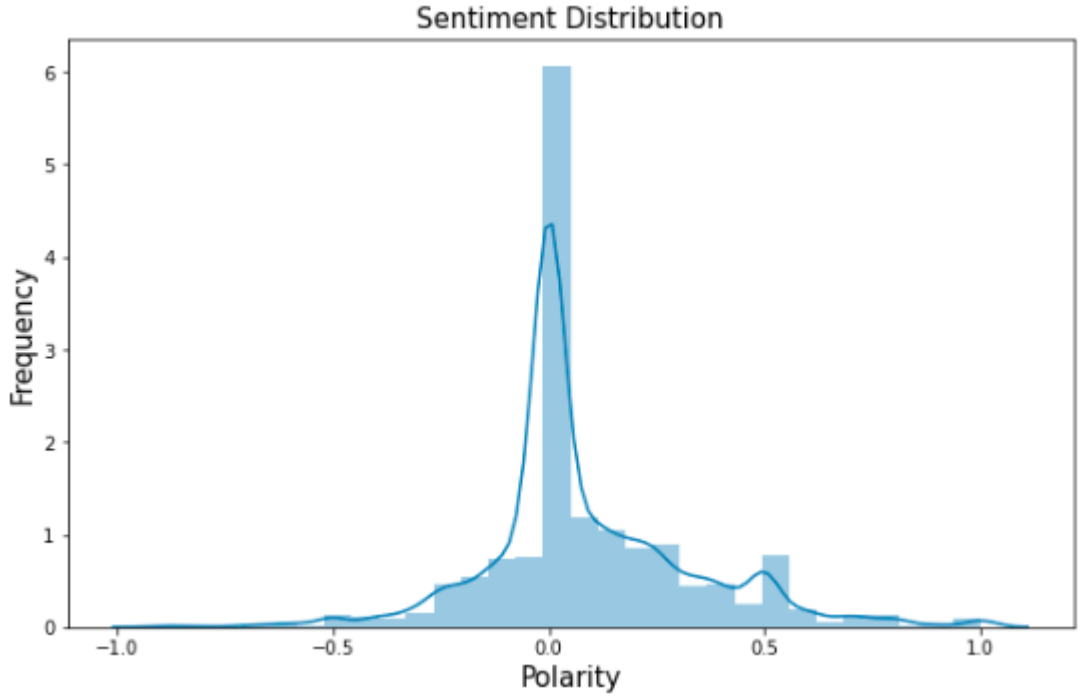
3. Sentiment Analysis

```
positive    715
neutral     540
negative    319
Name: sentiment, dtype: int64
```



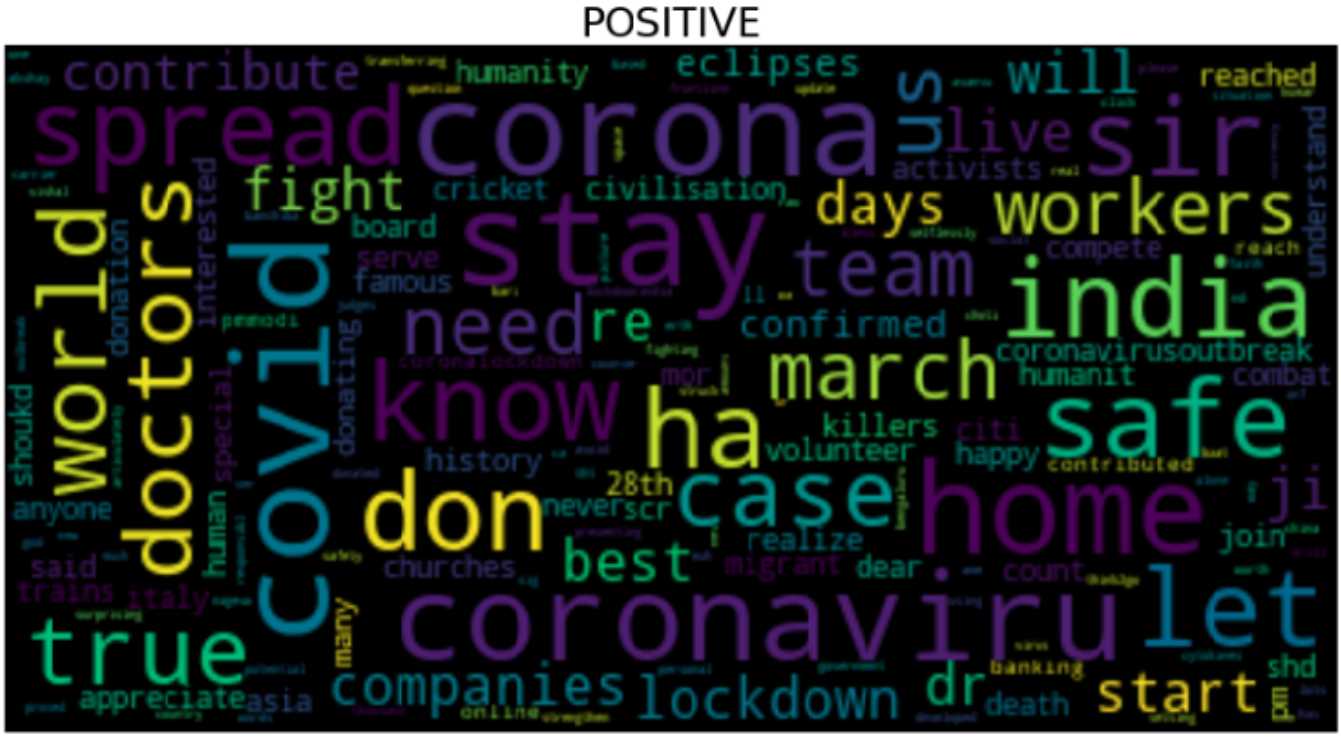
Out of 1,574 total tweets, 715 tweets are found to be in favour of lockdown whereas 319 tweets are found to be against the lockdown.

4. Sentiment Distribution



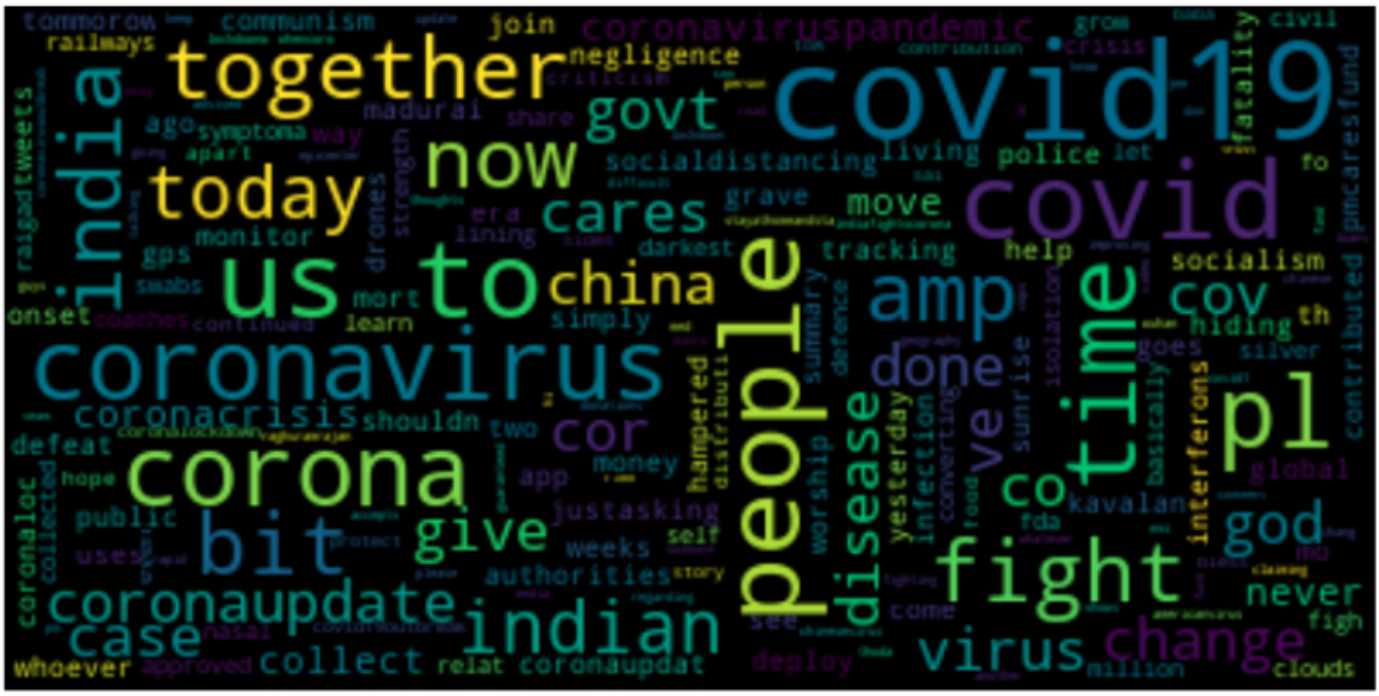
This depicts the polarity frequency of nature of tweets. For polarity greater than 0, nature is positive, polarity less than 0, negative and equal to 0, neutral. So, frequency found to be mostly neutral and positive in context of covid19 lockdown.

5. Wordcloud for each sentiment



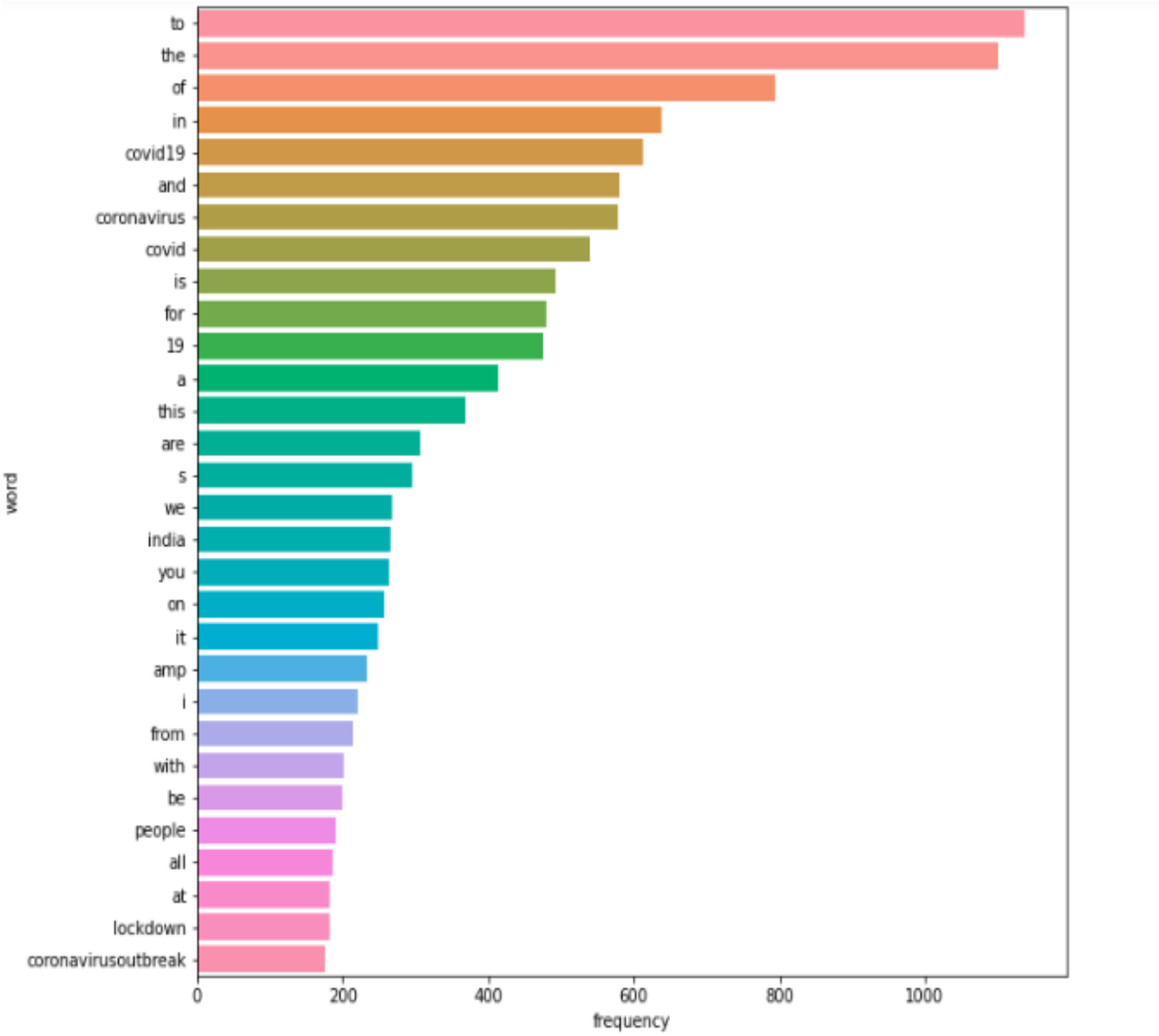
NEGATIVE

NEUTRAL



These wordclouds shows frequently occuring words in tweets for each sentiments.
For positive sentiment wordcloud, words like corona, india, workers, safe, covid, doctors, etc. are observed. In case of negative sentiment wordcloud, words like covid, unforutnate, wish, waste, unplanned, humanity, fight, donation, etc. are observed and for neutral sentiment wordcloud, corona, people, time, today, cares, give, etc. words are observed.

6. word vs frequency

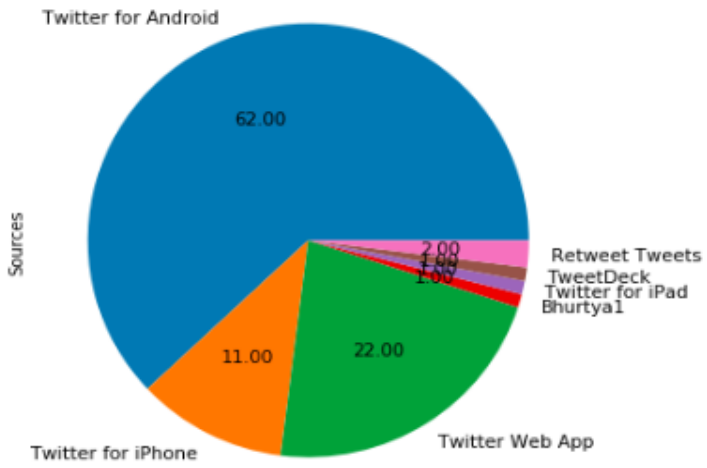


This graph depicts the frequency of most frequent word used in tweets during lockdown and it is found that frequency of 'covid19' is above 600, which is greater than words like coronavirus, lockdown, etc.

7. For Latest Tweets

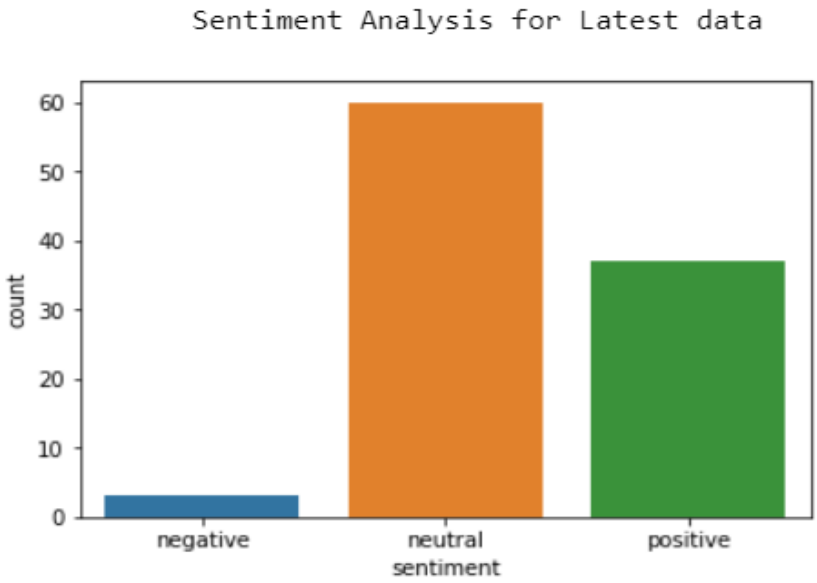
i. Sources frequency

Mostly tweets are being done by android devices as compared to other sources.



ii. Sentiment Analysis

```
neutral     60
positive    37
negative     3
Name: sentiment, dtype: int64
```



Out of 100 tweets, 37 are found to be in favour of lockdown whereas 3 are found to be against lockdown.

For data visualization of late march tweets, a dashboard with different graphs is created using IBM Watson Dashboard with the resultant dataset formed after performing sentiment analysis on the original data. Representation is done using a bar graph, pie chart,word cloud of tweets, etc. to show various comparisons.
For data visualization of latest tweets, used tweepy to extract the tweets and plot its graph using matplotlib and seaborn.

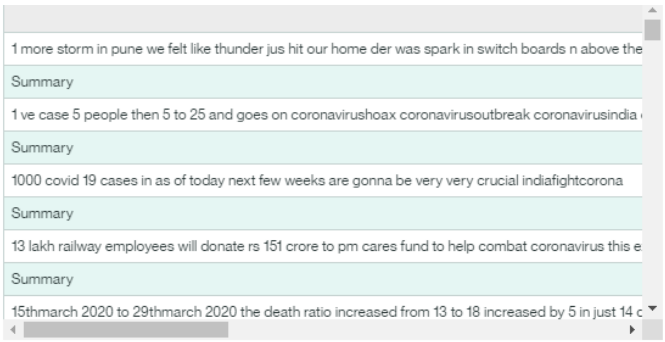
V. Dashboard Creation

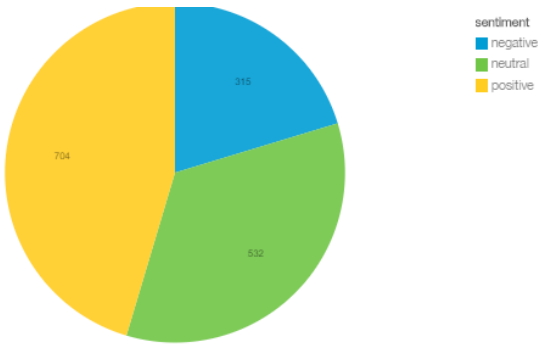
Comparisons were also made on the recent public opinion and the opinion during the initial phases of lockdown (late March).
Dashboard is imported in project using IBM Watson Studio.

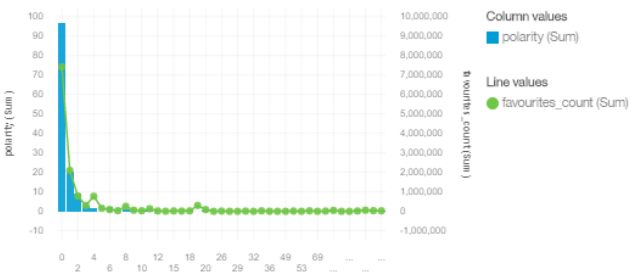
Used dashboard to represent tweets concerned to late March.
link:
https://eu-gb.dataplatform.cloud.ibm.com/dashboards/0ab60232-6212-4fac-a9a0-fbfc291fa87d/view/5811c91811eb03f716b7e6e4079f2a072f322509e4bbd25288807b495e637197a8381498c87b180f89115360f0ef1a5ac8

To dashboard, showed interative UI to know the patterns, sentiments and polarity of tweets belonging to selected variation of sentiment.

Below is how dashboard looks like :



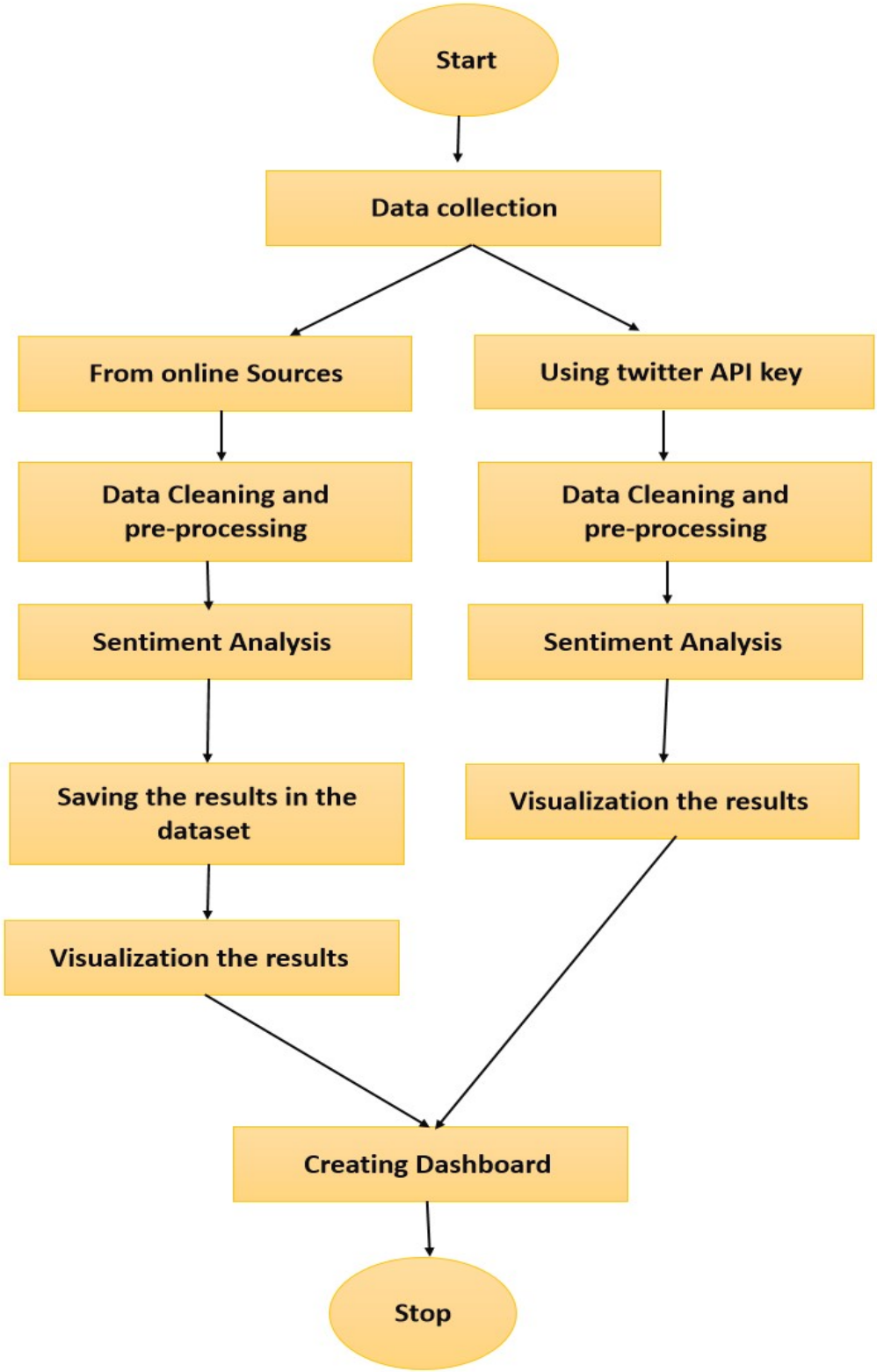




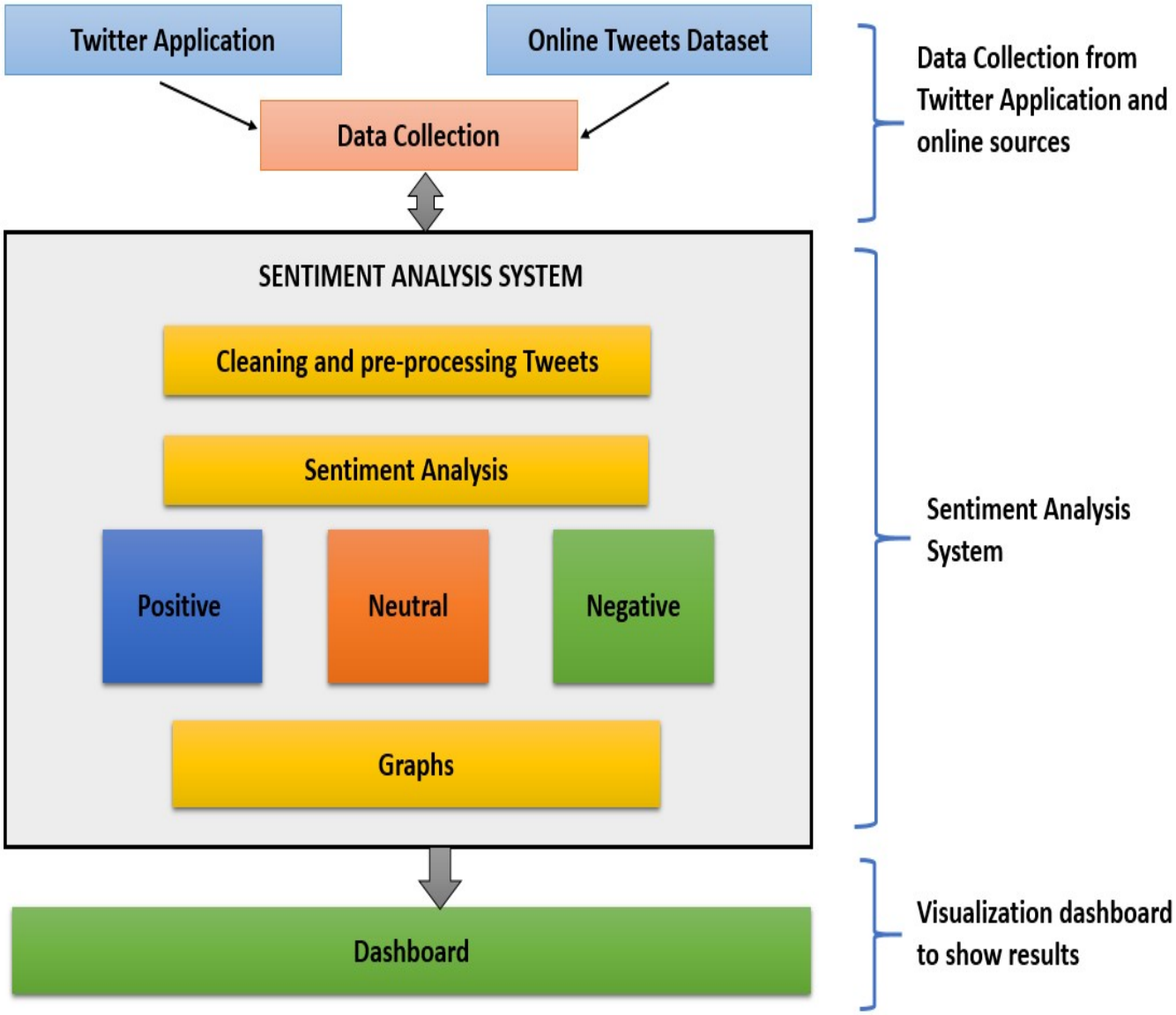
3,654 139.23

THEORITICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

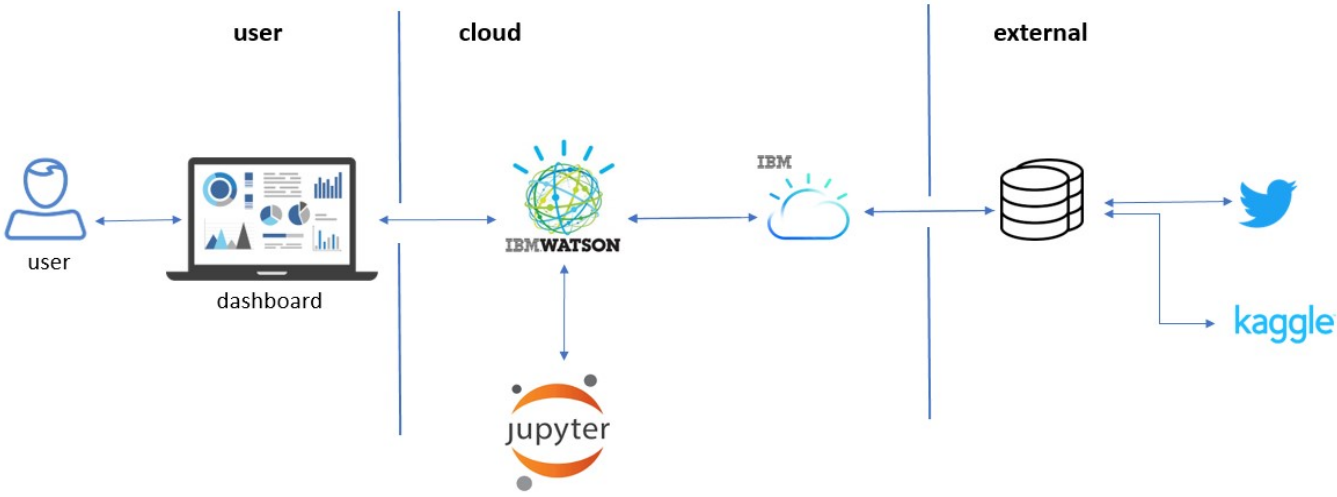


EXPERIMENTAL INVESTIGATIONS

In this process of developing the project I have undergone many investigation processes to learn and understand new concepts so that I can build the Twitter Sentimental Analysis on the Covid 19 lockdown Visualization Dashboard model, successfully. For the same, I had to learn and investigate following:

- IBM Cloud.
- Twitter API.
- IBM Watson Studio.
  - Jupyter Notebook
    - Natural Language Processing (NLP)
    - Machine Learning Algorithms
    - Text Blob
    - Tweepy
  - Dashboard
- ZOHO Writer.

FLOWCHART



CODE

- Sentiment Analysis Code :

```
df['sentiment'] = ' '
df['polarity'] = None
for i,tweets in enumerate(df.text) :
    blob = TextBlob(tweets)
    df['polarity'][i] = blob.sentiment.polarity
    if blob.sentiment.polarity > 0 :
        df['sentiment'][i] = 'positive'
    elif blob.sentiment.polarity < 0 :
        df['sentiment'][i] = 'negative'
    else :
        df['sentiment'][i] = 'neutral'
df.head()
```

	created_at	text	favourites_count	retweet_count	sentiment	polarity
0	0.066667	covid19 socialdistancing the indian way	10374	0	neutral	0
1	0.066667	possible case of human to animal transmission ...	22880	0	negative	-0.23125
2	0.116667	coronavirus us eclipses 120 000 confirmed case...	12968	6	positive	0.4
3	0.150000	the bitter truth covid 19 covid19india lockdow...	6	0	negative	-0.1
4	0.216667	coronavirusoutbreak all churches shd realize t...	38	0	positive	0.1375

- Sentiment Analysis of Latest tweets Code :

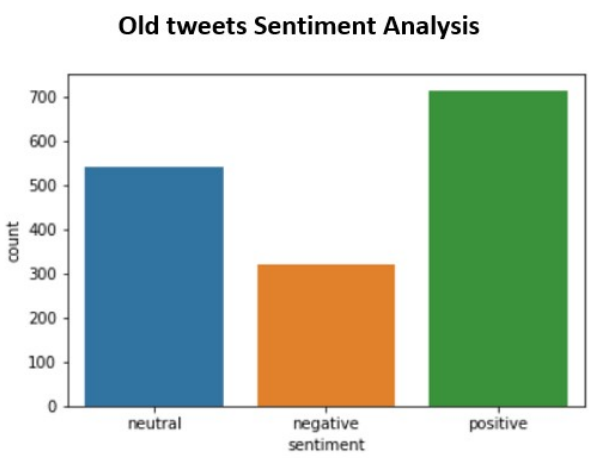
```
from textblob import TextBlob
import re

def clean_tweet(tweet):
    '''
    Utility function to clean the text in a tweet by removing
    links and special characters using regex.
    '''
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z
\t])|(\w+:\/\/\S+)", " ", tweet).split())

data['sentiment'] = ' '
data['polarity'] = None
for i,tweets in enumerate(data.Tweets) :
    blob = TextBlob(tweets)
    data['polarity'][i] = blob.sentiment.polarity
```

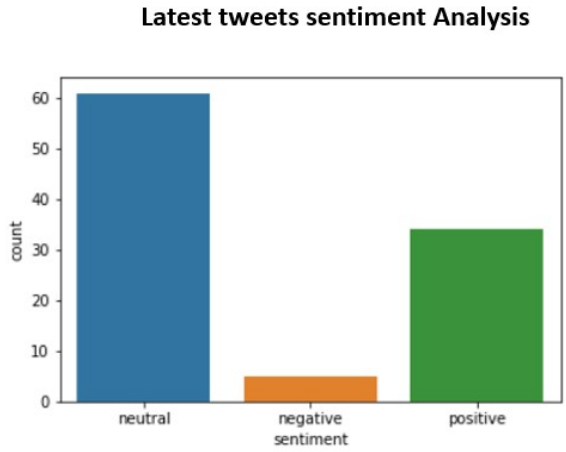
```
            data['sentiment'][i] = 'positive'
        elif blob.sentiment.polarity < 0 :
            data['sentiment'][i] = 'negative'
        else :
            data['sentiment'][i] = 'neutral'
    data.head()
```

	Tweets	len	ID	Date	Source	Likes	RTs	sentiment	polarity
0	He was so distraught that he didn't even accep...	140	1283022106439987201	2020-07-14 12:54:42	Twitter for Android	0	0	negative	-0.2
1	RT @GauravJain_9: #Againstcorona #lockdownindi...	140	1283018315963576325	2020-07-14 12:39:38	Twitter for Android	0	2	neutral	0
2	RT @GauravJain_9: #LockdownIndia #Mingrantwork...	117	1283018289849856002	2020-07-14 12:39:32	Twitter for Android	0	3	neutral	0
3	RT @HSnewsLive: #coronavirus के कारण बिहार में...	140	1283017858860019713	2020-07-14 12:37:49	Twitter for Android	0	2	neutral	0
4	RT @kakoligdastidar: If Europe south Korea New...	140	1283017789234573312	2020-07-14 12:37:32	Twitter for Android	0	44	positive	0.136364

RESULT

EXPERIMENTAL RESULTS :





Comparisons were made on the recent public opinion and the opinion during the initial phases of lockdown (late March)

During the initial phase, the score of positive reactions is more than that of the other sentiments. But, observations from the recent analysis show decrease in the positive sentiment score.

"Covid19" was the trending word and according to the visualizations (frequency analysis graphs).

A dashboard has been created, depicting all the visualizations.

UI INTERFACE :

https://eu-gb.dataplatform.cloud.ibm.com/dashboards/0ab60232-6212-4fac-a9a0-fbfc291fa87d/view/5811c91811eb03f716b7e6e4079f2a072f322509e4bbd25288807b495e637197a8381498c87b180f89115360f0ef1a5ac8

ADVANTAGES

The Watson twitter sentiment analysis model includes following advantages:

- Easy to create and less time required to develop.
- High performance with IBM cloud.
- Provides smart solution with help of Watson inbuilt libraries and resources.
- Easily analyses text and execute functions based on sentiment of message and helps to take actions related to public concern accordingly.
- Easily analyses unstructured data and big masses of data.

DISADVANTAGES:

The Watson twitter sentiment analysis model includes following disadvantages:

- We need to pay for IBM platform service.
- Require storage to store dataset containing tweets.
- Does not analyse and process structured data directly.
- No analysis on the data falling in between the initial lockdown phase and the recent phase.

APPLICATIONS

- It can be used as a platform representing overall public opinion on the covid19 situation, for the people not having a technical background.
- Analyses the sentiments behind the public opinion on the lockdown and helps predict consequences if lockdown is extended.
- Can also be used for other purposes such as for recognizing harmful sentiments such as bullying, racism etc.
- Can be used by the government and organizations to determine public opinion on their policies and products.

CONCLUSION

This project gave us the analytical conclusion on the sentiments expressed in the tweets by people on the lockdown due to the Covid-19 pandemic. This made use of the IBM Watson tools, NLP technologies and Machine learning algorithms and the results were visualized on a dashboard depicting the overall public opinion (positive, negative, neutral) on the current lockdown situation, which will help us further predict their reaction, if the lockdown is extended.

Comparisons were made on the recent public opinion and the opinion during the initial phases of lockdown (late March).

During the initial phase, the score of positive reactions is more than that of the other sentiments. Hence, initially the public was strongly in support of the lockdown. But, observations from the recent analysis show decrease in the positive sentiment score i.e. the public support not against but has decreased a bit due to factors such as economic slowdown, job losses, food shortage etc.

"Covid19" was the trending word and according to the visualizations (frequency analysis graphs) , it shows that the overall reaction represents the opinion of the majority of public.

A dashboard has been created, depicting all the above visualizations.

FUTURE SCOPE

- Not restricting the analysis of the sentiments to just the initial and the recent phase but also considering data from the time phase between the two.
- It can be further expanded to analyze the public opinion throughout the world and not just the India region.
- Improving the accuracy and precision of the sentimental analysis performed and further narrowing it down to determining deeper emotions (like angry, sad, fear, anxiety, lonliness etc) and not just the superficial sentiments (positive, negative and neutral).
- Also, adding more features like determining the worst hit areas etc.

BIBILOGRAPHY

Names: Himanshi Gupta , Abhipriya Sharma, Shruti Mishra

College Name: Indira Gandhi Delhi Technical University for Women

Work Title: Sentiment Analysis of COVID-19 Tweets – Visualization Dashboard

References:

1. IBM Cloud: https://www.ibm.com/cloud/get-started
2. Dashboard: https://www.youtube.com/watch?v=jgl_w05xB9g
3. Sentiment analysis kick-start : https://towardsdatascience.com/twitter-sentiment-analysis-based-on-news-topics-during-covid-19-c3d738005b55

APPENDIX

Link to Dataset Used:

https://drive.google.com/drive/folders/15UXp4DCvVsD__XVr4PA6bvAFIQ_ypJY8

Link to Dashboard UI:

https://eu-gb.dataplatform.cloud.ibm.com/dashboards/0ab60232-6212-4fac-a9a0-fbfc291fa87d/view/5811c91811eb03f716b7e6e4079f2a072f322509e4bbd25288807b495e637197a8381498c87b180f89115360f0ef1a5ac8