Twitter Sentiment Analysis using IBM Cloud

Submitted in fulfillment of IBM Hack Challenge 2020

Team

CodePlay

By

Shreyas Udupa

Kunal Kotkar

Sanket Jangale

Manoj Ayyappan

TABLE OF CONTENTS

INTRODUCTION

OVERVIEW

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and personal information. We can apply Sentiment analysis to the voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Sentiment analysis provides some answers into what the most critical issues are, from the perspective of customers, at least. Because sentiment analysis can be automated, decisions can be made based on a significant amount of data rather than plain intuition, that isn't always right.

Sentiment analysis is instrumental in social media monitoring as it allows us to gain an overview of the public opinion behind specific topics. We can find Customer sentiment in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the domain of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace.

One of the most essential and convenient platform to express one's feelings about an issue or an event is Twitter. Twitter is an important platform where people can talk or "tweet" about the current trending events going on in the world. People can talk about existing events or start a new trending topic altogether.

PURPOSE

Sentiment analysis is one of the most innovative and productive ways used by companies and the government alike to analyze the sentiments of the people based on the events/decisions taken by them. The proper analysis of the tweets will help them conduct their affairs in a particular way.

One of the most well-documented uses of sentiment analysis is to get a full 360 view of how your customers and stakeholders view your brand, product, or company.

Widely available media, like product reviews and social, can reveal critical insights about what your business is doing right or wrong. Companies can also use sentiment analysis to measure the impact of a new product, ad campaign, or consumer's response to recent company news on social media. Private companies like Unamo offer this as a service.

Sentiment analysis is used in business intelligence to understand the subjective reasons why consumers are or are not responding to something (e.x. why are consumers buying a product? What do they think of the user experience? Did customer service support meet their expectations?). We can implement Sentiment analysis in the areas of political science, sociology, and psychology to analyze trends, ideological bias, opinions, and gauge reactions.

We can make use of Sentiment analysis to analyze the tweets made by people during times like these, i.e., Corona pandemic. People tweet about the daily impact and the effect of Covid-19 on the world. Whenever the number of positive cases in a day increases above a specific limit or if it doesn't increase, it elicits a different response each time. More positive cases means that people are getting scared and confused. No positive cases or less positive cases means people are getting happy and relieved. Companies and the government can perform sentiment analysis over these tweets and determine the mood of the people as it fluctuates. They can use this data to implement decisions/ introduce new products in the market.

LITERATURE SURVEY

EXISTING PROBLEM

On average, there are around 500 million tweets made every day. This vast amount of tweets is an untapped data source for any developer/company. Currently, this magnitude of data is lying waste and not being utilized properly. Stats show that 40% of buyers form an opinion of a business after reading 1-3 online reviews, and that 64% software buyers read at least 6 online reviews before making a purchase, giving us a clue of how important it is for companies to track the conversation around them and uncover the feelings behind what's being said.

Efficient and methodical analysis of the tweets related to some topic/hashtag using various algorithms at our disposal can bring out the actual emotion of the people. Using this result, companies and the government can alter their existing products and decisions.

For example, let's take a company named ABC. ABC recently launched an ad campaign for their new upcoming product. Based on the sales of the new product, they found that the sales rate was just 50% of that of their previous product. This situation is not favorable for a company in any way. They had to find a reason for the poor performance of the product. It is in these kinds of situations that sentiment analysis plays an enormous and vital role.

PROPOSED SOLUTION

In the example mentioned above, the company uses sentiment analysis on the reviews of the product to find the shortcomings in the product. The sentiment analysis divides the reviews into three sentiments- positive, negative, and neutral. So by examining the negative reviews, they found that most of the viewers found the visuals to be disturbing and offensive, which forced them to overlook this product. So acting on this result, the company decided to revamp their ad campaign and relaunch their product. When the sales of the revamped product was compared with the previous launch of the

product, they found out that the sales outperformed the previous sales by 50%. This is a significant increase compared to the previous product.

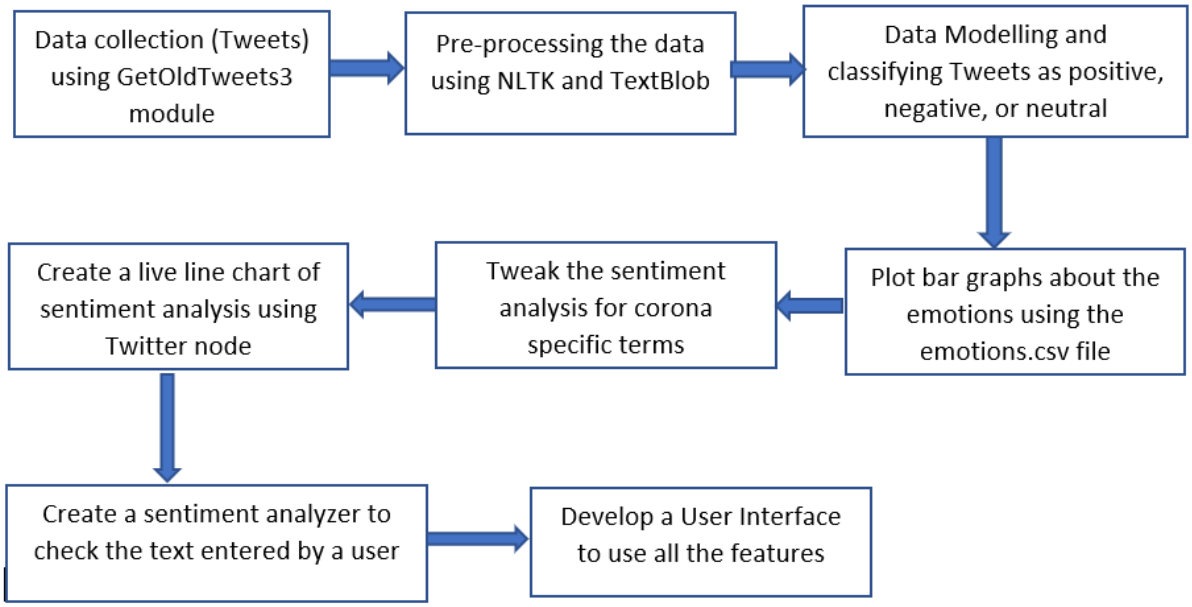
The increase in sales was possible due to the sentiment analysis performed on the reviews.

We are going to use the sentiment analysis to analyze the emotions of the people during this pandemic -Covid-19 and lockdown implementation of the government. We are taking 2 sets of tweets, one having Covid-19 as the hashtag and the other one having lockdown as the hashtag.

We are going to perform sentiment analysis on these tweets and plot the sentiments on a pie chart and the emotions of the tweets on a bar graph. We are going to create a UI where the consumer can visualize the results in an interactive and graphical platform. Based on this result, the consumer can make appropriate decisions to alter their working.

THEORETICAL ANALYSIS

BLOCK DIAGRAM:



HARDWARE/SOFTWARE DESIGNING:

As there was a limitation on the retrieving the tweets older than a week through the Twitter API, we used the GetOldTweets3 library in python. We have also specified the location as required. We collected a total of 8000 Tweets, out of which 4000 were related to Covid-19 and rest, to government lockdown in India. The spelling mistakes in these tweets were then corrected and saved as a .csv file.

This dataset was then cleaned and assigned a polarity number using the TextBlob module. Higher the polarity, the more positive is the tweet. We removed the Stopwords and used the NLTK library to clean the data further.

We then imported a custom made file called emotions.csv, which has all the words mapped to specific emotions. So we classify all the tweets in our dataset into the different emotions and plot bar graphs for the same. We then create pie charts of the data according to the polarity number of the tweets.

Now we implement the whole model on the IBM Cloud platform using Node-RED and thus create an interactive website. This website contains 5 tabs, through which we can access all the data.

There is also a live line chart that shows the live sentiment analysis of the Tweets in real-time. We also tweaked the flows on Node-RED using switch nodes so that they show the correct sentiment for corona related tweets. There is an additional section where users can check the sentiment of any text that they enter. Along with this, we have displayed the live coronavirus count in India, the number of active cases, and the number of people recovered. We have displayed the same using a pie chart. We used the website https://documenter.getpostman.com/view/10808728/SzS8rjbc?version=latest#c34162be-7c20-418e-9866-a24dca632b3c to get this statistic which is updated daily.

EXPERIMENTAL INVESTIGATION

In our first implementation of the model, we have taken 2000 tweets, each having the keywords 'Corona' and 'lockdown.' On running this dataset in our model, we achieved an accuracy of 73%. Our sentiment analysis of the same gave us 45% positive in tweets containing the keyword 'lockdown' and 50% positive in tweets containing the keyword 'Corona.'

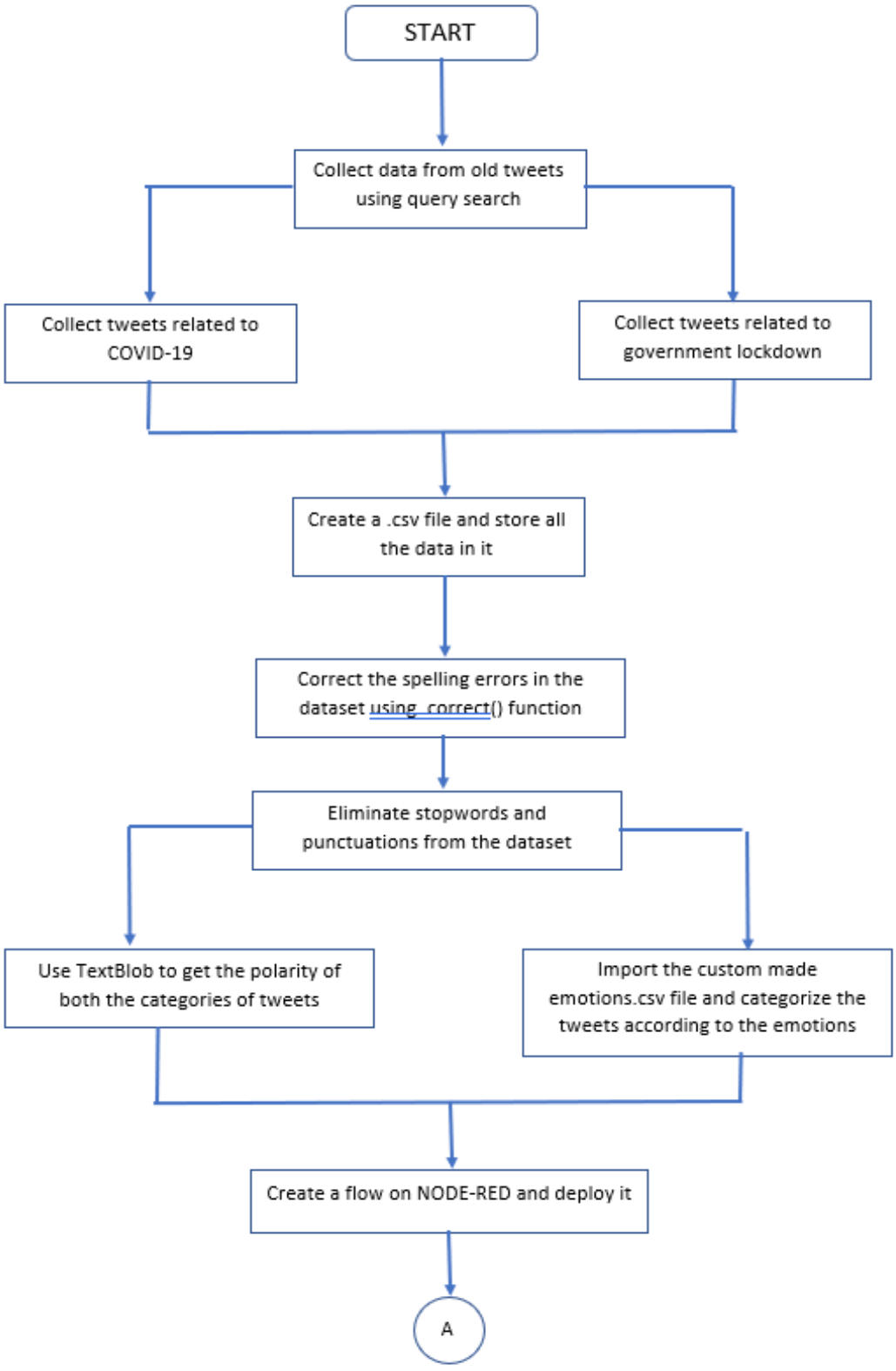
In our second implementation of the model, we have taken 4000 tweets, each having the keywords 'Corona' and 'lockdown.' On running this dataset in our model, we achieved an accuracy of 76%. Our sentiment analysis of the same gave us 49% positive in tweets containing the keyword 'lockdown' and 39% positive in tweets containing the keyword 'Corona.'

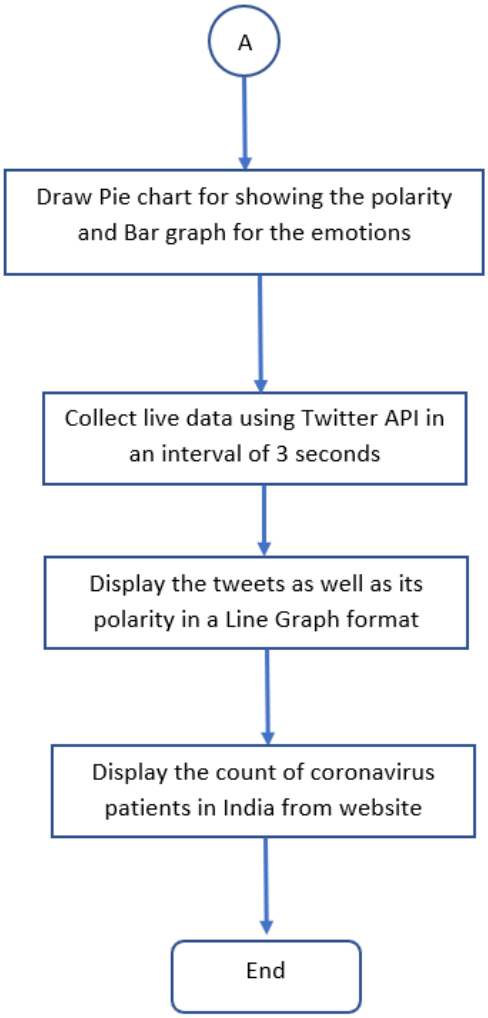
We can concur from the above experiment that as the number of tweets in our dataset increases, the accuracy of our model increases, and the percentage of the sentiments also becomes more accurate.

In our sentiment analyzer, we have used the Sentiment analyzer resource of IBM Cloud. When we enter the phrase "I tested positive for Covid-19 today," it was classified as a positive tweet even though we know it should be qualified as a negative tweet. This happens because of the word "positive" in the phrase, as it gets classified as a positive tweet.

So to increase the accuracy of the model, we use the switch feature of Node-RED, where we have used two sets of switch. One contains the words which, when used in a sentence, gives a negative sentiment and another which contains the words that give a positive sentiment.

FLOWCHART


START
Collect data from old tweets using query search
Collect tweets related to COVID-19
Collect tweets related to government lockdown
Create a .csv file and store all the data in it
Correct the spelling errors in the dataset using _correct() function
Eliminate stopwords and punctuations from the dataset
Use TextBlob to get the polarity of both the categories of tweets
Import the custom made emotions.csv file and categorize the tweets according to the emotions
Create a flow on NODE-RED and deploy it
A


A
Draw Pie chart for showing the polarity and Bar graph for the emotions
Collect live data using Twitter API in an interval of 3 seconds
Display the tweets as well as its polarity in a Line Graph format
Display the count of coronavirus patients in India from website
End

RESULT

We have developed a User Interface to interact with all the features and data that we have collected and processed on a website hosted using the Node-RED platform on the IBM Cloud. We have divided The website into five sections(tabs) :

1. Home
2. COVID Tweets Analysis
3. Lockdown Tweets Analysis
4. Live Analysis
5. Check your sentiment & COVID India tracker

We have observed that the general sentiment of the people regarding the Covid-19 pandemic is mostly positive, which shows people are hopeful despite the daily increase in the number of cases around the world. Regarding the lockdown implementation by the Government of India, people have a positive outlook towards it, hoping that imposing of the lockdown will curb the spread of the coronavirus. On the other hand, people who own businesses are somewhat disappointed by these decisions as they are unable to earn their livelihood and are suffering the most.

We achieved an accuracy of 76% when the dataset containing the tweets related to COVID-19 was passed to our model, while implementing the SVM algorithm. We can increase the accuracy upon further research into Machine Learning models and by using a vast dataset.

ADVANTAGES & DISADVANTAGES

Advantages

- The results from sentiment analysis help businesses understand the conversations and discussions taking place about them, and helps them react and take action accordingly.
    - They can quickly identify any negative sentiments being expressed and turn poor customer experiences into desirable ones.
    - They can create better products and services, and they can formulate the marketing messages they send out according to the sentiments being expressed by their target audience or customers.
    - They can identify where they may be excelling, or identify where there's room for improvement compared to the competition.
- By listening to and analyzing comments on Facebook and Twitter, local government departments can gauge public sentiment towards their department and the services they provide, and use the results to improve services such as parking and leisure facilities, local policing, and the condition of roads.
- Universities can use sentiment analysis to analyze student feedback and comments garnered either from their surveys, or from online sources such as social media.
    - They can then use the results to identify and address any areas of student dissatisfaction, as well as identify and build on those areas where students are expressing positive sentiments.
- And by analyzing the sentiment behind customer reviews on sites like TripAdvisor and Yelp, hotels and restaurants can not only manage their reputations by improving the services offered, but can also gauge the general customer attitude to their business or brand.

Disadvantages

- Computer programs have problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations - failing to recognize these can skew the results.
- 'Disappointed' may be classified as a negative word for sentiment analysis, but within the phrase "I wasn't disappointed," it should be classified as positive.
- We would find it easy to recognize as sarcasm the statement "I'm really loving the enormous pool at my hotel!" if this statement is accompanied by a photo of a tiny

swimming pool. In contrast, an automated sentiment analysis tool probably would not, and would most likely classify it as an example of positive sentiment.

- With short sentences and pieces of text, for example, like those you find on Twitter, especially, and sometimes on Facebook, there might not be enough context for a reliable sentiment analysis.
- So, automated sentiment analysis tools do a great job of analyzing text for opinion and attitude, but they're not perfect .

APPLICATIONS

The sentiment analysis of the tweets has various applications.

- It can be used by companies to alter their business model by taking the feedback on their products
- Govts can alter their decision regarding the implementation of specific rules and regulations
- Brands can use it to improve their products.
- Social media personalities and celebrities can use to check the emotion and sentiments of their tweets before tweeting it officially.

CONCLUSION

Sentiment analysis is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. Sentiment analysis has a vital role in ensuring people satisfaction in various sectors of society. Tweets made on Covid-19 and Lockdown from Twitter.com were selected as data used for this study. A pie chart of sentiments and a Bar graph of emotions associated with the tweets are displayed. A live chart of the analysis of the tweets is shown as an additional feature. A Tweet analyzing model is provided to the user where the user can check the sentiment of their tweets. A live COVID-19 India tracker displaying the number of active, recovered, and total cases is displayed, and a pie chart depicting the same is displayed for better user experience.

FUTURE SCOPE

The Sentiment Analysis model, which we have developed, focuses on the Tweets between specific dates (approximately 2 months) and limited data. The user cannot alter this dataset as in; the user cannot check the sentiment analysis over a particular period of time according to his/her will. So in future developments of this model, we can expect to vary the dates between which data about Tweets is collected and display the results accordingly.

In the "Check Your Sentiment" feature, we did not have a significant enough dataset to train a Machine Learning(ML) model for specific Corona related Tweets. We covered a significant part of this specialized data with the use of switch nodes, but it would work better with a proper ML model. So in future developments, we would want to create/get access to a big enough dataset so that it can be incorporated into an ML model for better and more accurate results.

With the ever-increasing vocabulary, it is almost impossible to incorporate all the words which show emotions in the "emotions.csv" file that we made. But it is certainly possible to continually update all the new words that come up. So over a period of time, this dataset should grow and be reasonably accurate in classifying all the emotions in a Tweet.

Sentiment analysis has a very bright future ahead. We can virtually use it in all the sectors in the society. From corporates, to the government, to schools and colleges, etc., it has an extensive range of applications.

Analyzing the feedback of the sentiment of students related to their academics can help the schools and colleges to alter their decisions to ensure that the students are satisfied with their learning process.

Public Transportation system can use the sentiment analysis to introduce new travel routes, more buses on more frequented routes, improve the quality of buses, take appropriate actions against the employees, and many more.

BIBLIOGRAPHY

- The Bootcamp organized by SmartInterns on Youtube :
  https://www.youtube.com/watch?v=x_5kH26xics

- Node-RED documentation :
  https://nodered.org/docs/

- Twitter API documentation :
  https://developer.twitter.com/en/docs

- Get old Tweets from Twitter :
  https://pypi.org/project/GetOldTweets3/

- Data cleaning and Pre-processing :
  https://www.youtube.com/channel/UCirPbvoHzD78Lnyll6YYUpg

- COVID Statistics API :
  https://covid19api.com/

APPENDIX

Main code

Data Collection

```
pip install GetOldTweets3
import GetOldTweets3 as got
import csv
from textblob import TextBlob
!pip install vaderSentiment
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
sid_obj = SentimentIntensityAnalyzer()
tweetCriteria = got.manager.TweetCriteria().setQuerySearch('Covid-19 corona')\
                                .setSince("2020-05-01")\
                                .setUntil("2020-07-02")\
                                .setLang("en")\
                                .setMaxTweets(4000)
tweets = got.manager.TweetManager.getTweets(tweetCriteria)
text_tweet = [[tweet.text] for tweet in tweets]

tweetCriteria_1 = got.manager.TweetCriteria().setQuerySearch('government lockdown')\
                                .setSince("2020-05-01")\
                                .setUntil("2020-07-02")\
                                .setNear("Nagpur")\
                                .setWithin("1100mi")\
                                .setLang('en')\
                                .setMaxTweets(4000)
tweets_1 = got.manager.TweetManager.getTweets(tweetCriteria_1)
text_tweet_1 = [[tweet.text] for tweet in tweets_1]

# opening the csv file in 'w' mode
flatten_list = [j for sub in text_tweet for j in sub]
flatten_list_1=[j for sub in text_tweet_1 for j in sub]
print(len(flatten_list))
print(len(flatten_list_1))
file = open('tweet_dataset.csv', 'w', newline ='',encoding="utf-8") #We have included utf-8 to
represent unicode present in the text
```

```
with file:
    # identifying header
    header = ['Tweets','Polarity','Tweets1','Polarity1']
    x,y=0,0
    writer = csv.DictWriter(file, fieldnames = header)
    # writing data row-wise into the csv file
    writer.writeheader()
    for i in range(0,4000):
        flatten_list_1[i]=str(TextBlob(flatten_list_1[i]).correct())
        review_1=sid_obj.polarity_scores(flatten_list[i])
        review_2=sid_obj.polarity_scores(flatten_list_1[i])
        x=review_1['compound']
        y=review_2['compound']
        writer.writerow({'Tweets':flatten_list[i],'Polarity':x,'Tweets1':flatten_list_1[i],'Polarity1':y})
#Polarity is used to check whether the given text is positive or negative
```

DATA CLEANING

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import types
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.
client_abd1760a68c5414bb43cc67e72f9da4f = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='KZ0LylVxLMUJLuqKeFuJoNWI0RYHSGA_1A2cOLQ6IZ25',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3.eu-geo.objectstorage.service.networklayer.com')
```

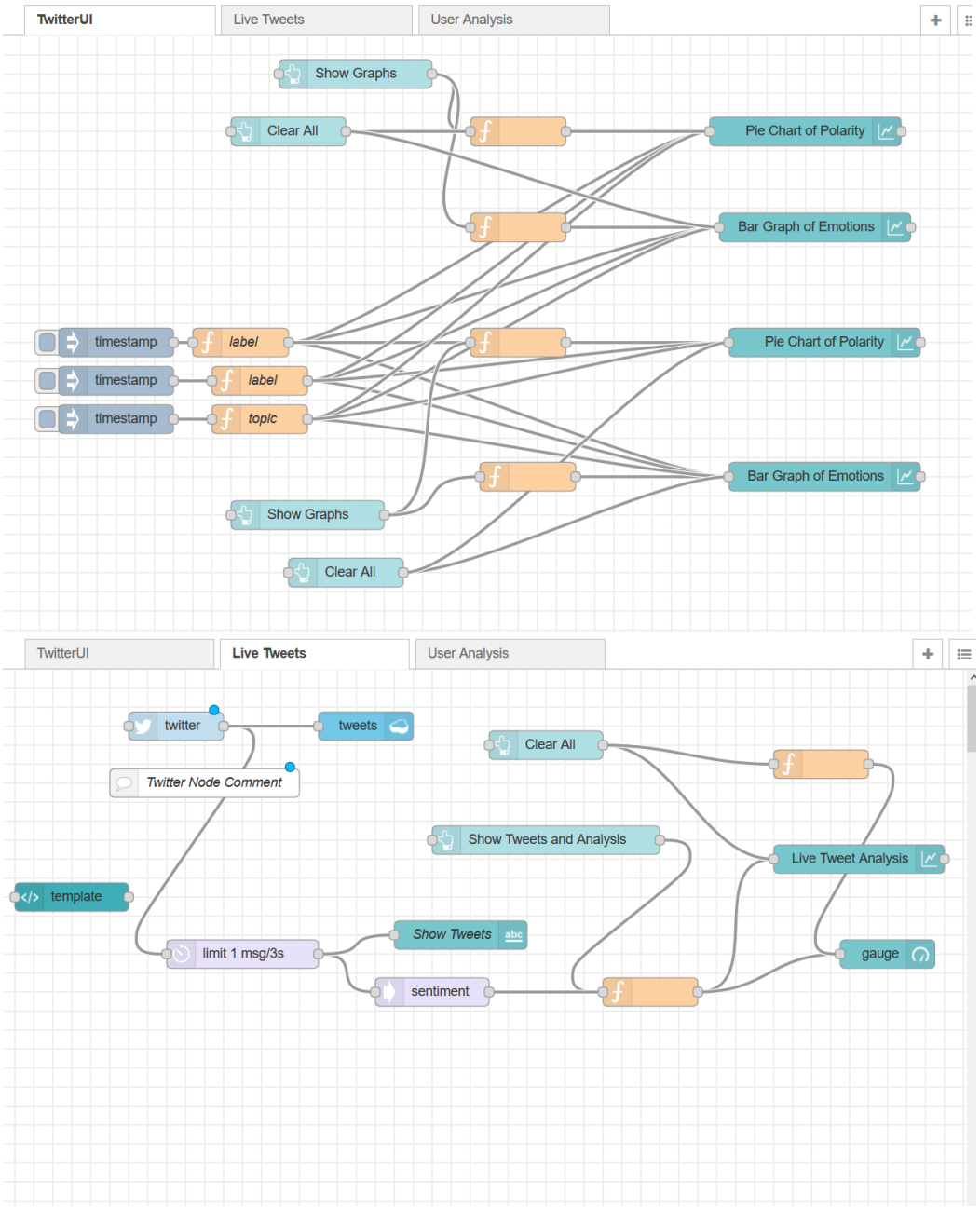
```
body = client_abd1760a68c5414bb43cc67e72f9da4f.get_object(Bucket='ibm-donotdelete-pr-
be1hyl5syosczh',Key='tweet_dataset (1).csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df_data_3 = pd.read_csv(body)
df_data_3.head()

import nltk
import re
!pip install textblob
from textblob import TextBlob
nltk.download('stopwords')
from nltk.corpus import stopwords
corpus = []
all_stopwords=stopwords.words('english')
all_stopwords.remove('not')
for i in range(0,4000):
    review = re.sub('[^a-zA-Z]',' ',df_data_3['Tweets'][i])#removing all punctuations
    review = review.lower() #lower case conversion
    review = review.split() #converting statement into list of words
    review = [word for word in review if not word in set(all_stopwords)]
    corpus.append(review)

corpus_1=[]
for i in range(0,4000):
    review_1 = re.sub('[^a-zA-Z]',' ',df_data_3['Tweets1'][i])#removing all punctuations
    review_1 = review_1.lower() #lower case conversion
    review_1 = review_1.split() #converting statement into list of words
    review_1 = [word for word in review_1 if not word in set(all_stopwords)]
    corpus_1.append(review_1)
```

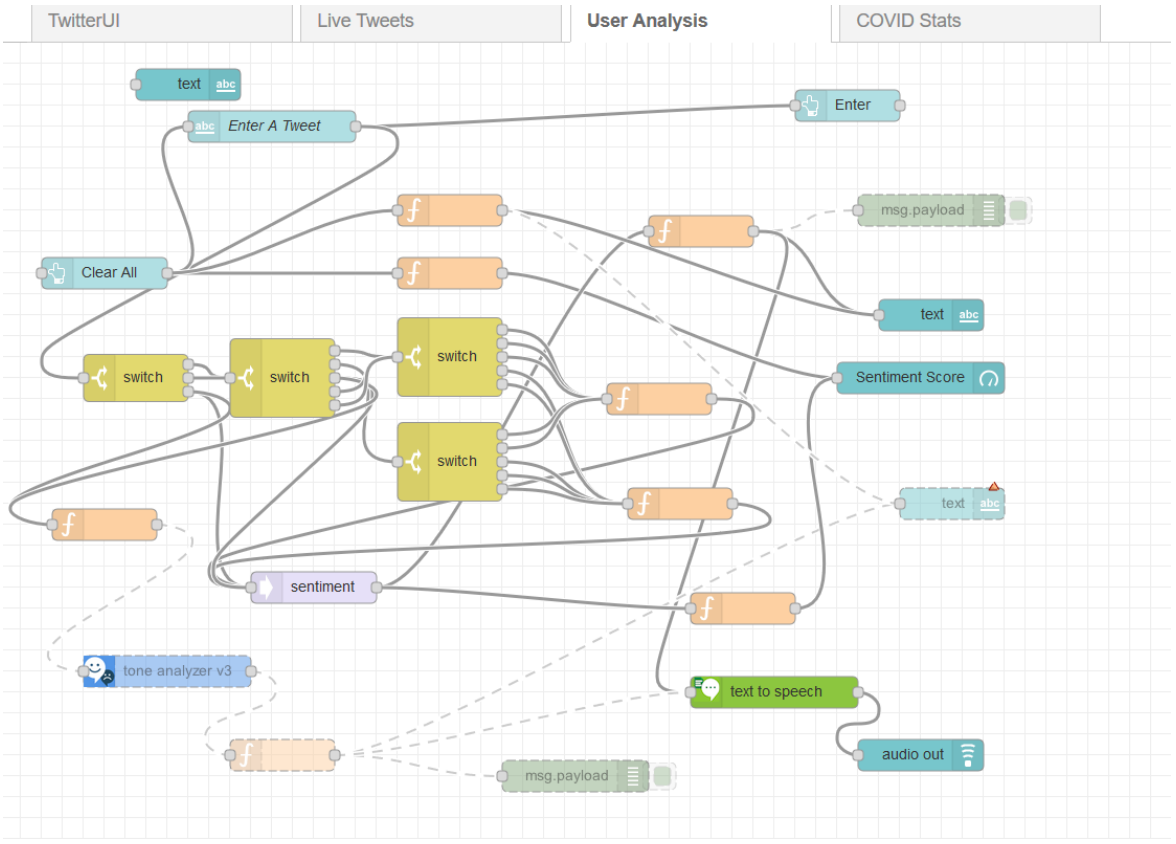
Data Modelling

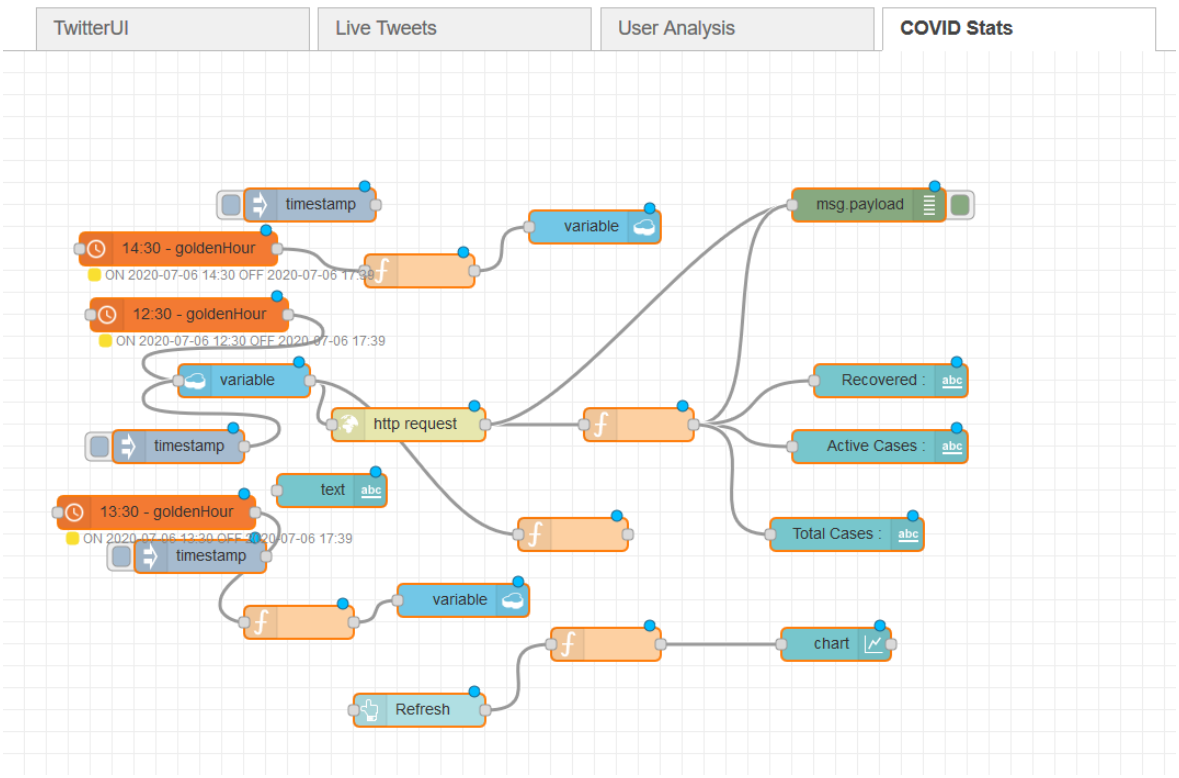
```
x=df_data_3.iloc[:,1].values
x1=df_data_3.iloc[:,3].values

positive,positive_1=0,0
negative,negative_1=0,0
neutral,neutral_1=0,0
s,s_1=0,0
avg,avg_1=0,0
for i in range(0,4000):
   if x[i]>0.05:
      positive+=1
   if x[i]<=0.05 and x[i]>=-0.05:
      neutral+=1
   if x[i]<-0.05:
      negative+=1
   if x1[i]>0.05:
      positive_1+=1
   if x1[i]<=0.05 and x1[i]>=-0.05:
      neutral_1+=1
   if x1[i]<-0.05:
      negative_1+=1
   s+=x[i]
   s_1+=x1[i]
avg= s/4000
avg_1 = s/4000

emotion_list=[]
emotion_list_1=[]
word=df_data_2.iloc[:,0].values
emotion=df_data_2.iloc[:,1].values
for i in range(0,4000):
   for j in range(0,np.size(word)):
      if word[j] in corpus[i]:
         emotion_list.append(emotion[j])
      if word[j] in corpus_1[i]:
         emotion_list_1.append(emotion[j])

from collections import Counter
w=Counter(emotion_list)
```

```
w1=Counter(emotion_list_1)
print(w)
print(w1)

plt.rcParams['figure.figsize']=(15,9)
data_1=[positive,negative,neutral]
label_1='Positive','Negative','Neutral'
plt.pie(data_1,labels=label_1,autopct='%1.1f%%')
plt.title('Sentiment Analysis of #corona')
plt.axis('equal')
plt.show()

fig1,ax1=plt.subplots()
ax1.bar(w.keys(),w.values())
fig1.autofmt_xdate()
plt.title('Emotion analysis of #corona tweets')
plt.xlabel('Emotions')
plt.ylabel('Counts')

data_2=[positive_1,negative_1,neutral_1]
label_2='Positive','Negative','Neutral'
plt.pie(data_2,labels=label_2,autopct='%1.1f%%')
plt.title('Sentiment Analysis of #government lockdown')
plt.axis('equal')
plt.show()

fig1,ax1=plt.subplots()
ax1.bar(w1.keys(),w1.values())
fig1.autofmt_xdate()
plt.title('Emotion analysis of #government lockdown tweets')
plt.xlabel('Emotions')
plt.ylabel('Counts')
```

UI


TwitterUI
Live Tweets
User Analysis
Show Graphs
Clear All
Pie Chart of Polarity
Bar Graph of Emotions
timestamp
label
timestamp
label
timestamp
topic
Pie Chart of Polarity
Bar Graph of Emotions
Show Graphs
Clear All
TwitterUI
Live Tweets
User Analysis
twitter
tweets
Twitter Node Comment
Clear All
Show Tweets and Analysis
Live Tweet Analysis
template
limit 1 msg/3s
Show Tweets
sentiment
gauge

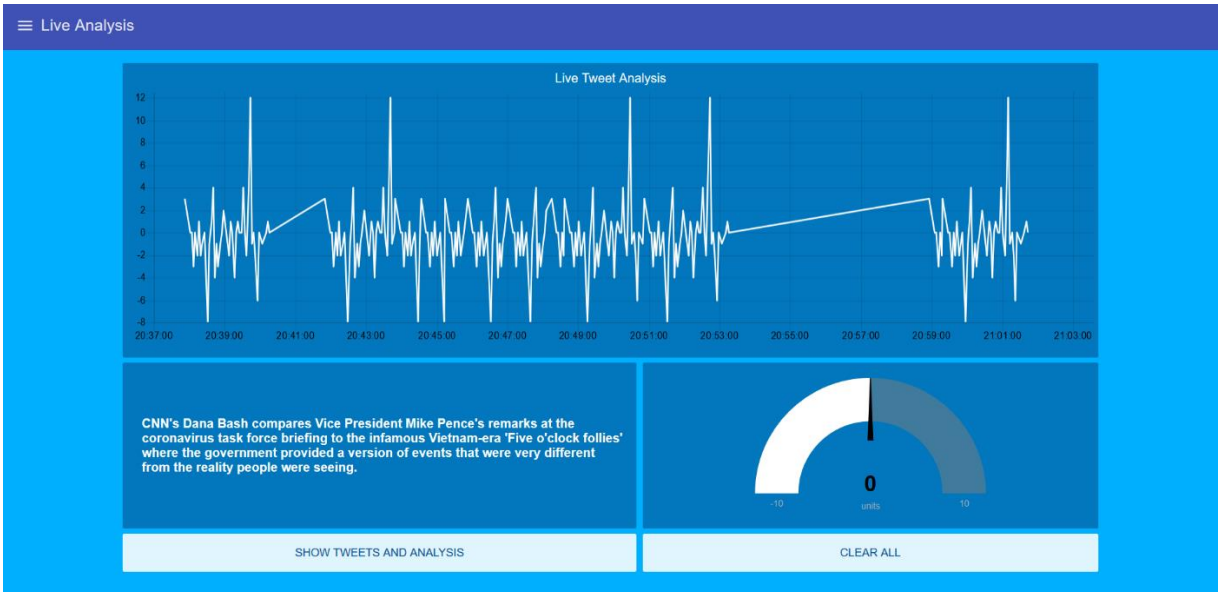

TwitterUI
Live Tweets
User Analysis
COVID Stats
text
Enter A Tweet
Enter
msg.payload
Clear All
text
switch
switch
switch
Sentiment Score
switch
text
sentiment
tone analyzer v3
text to speech
audio out
msg.payload

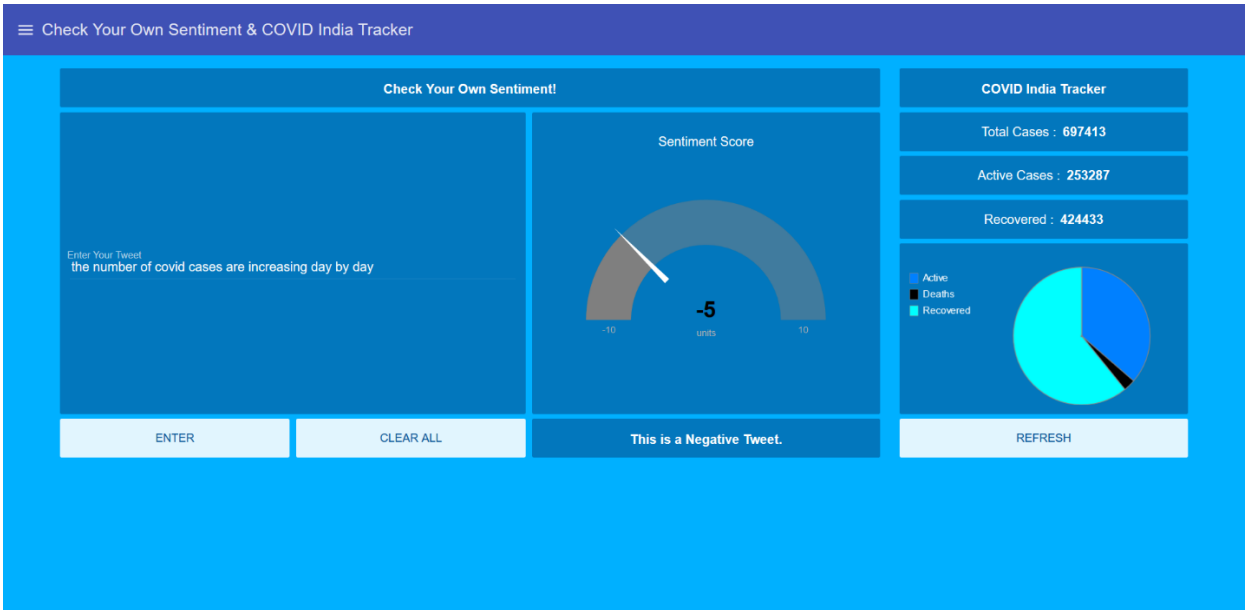

TwitterUI
Live Tweets
User Analysis
COVID Stats
timestamp
variable
msg.payload
14:30 - goldenHour
ON 2020-07-06 14:30 OFF 2020-07-06 17:39
12:30 - goldenHour
ON 2020-07-06 12:30 OFF 2020-07-06 17:39
variable
Recovered :
http request
timestamp
Active Cases :
text
13:30 - goldenHour
Total Cases :
timestamp
variable
chart
Refresh


Home
Twitter Sentiment Analysis
IBM HACK CHALLENGE 2020
This is your one stop solution for obtaining the Tweet Analysis in real time!
You have access to the Sentiment analysis regarding COVID-19 and the government lockdown in the next tabs.
Get the real time analysis in a graphical way in the Live Analysis tab.
You will be able to view Pie charts and Bar graphs giving you the best insights on the sentiment analysis.
The live analysis will be shown as a Line graph.
Team name - CodePlay
Sanket Jangale
Kunal Kotkar
Manoj Ayyappan
Shreyas Udupa
College : VESIT, Mumbai


Covid Tweets Analysis
Analysis of #COVID Tweets
SHOW GRAPHS
CLEAR ALL
These graphs represent the data analysis about tweets posted between 05-05-2020 and 01-07-2020.
Sentiment Analysis
Pie Chart of Polarity
Positive
Negative
Neutral
Emotion Analysis
Bar Graph of Emotions
Tweets Count
120
100
80
60
40
20
0
Happy
Angry
Fearful
Sad
Average
Surprise


Lockdown Tweets Analysis
Analysis of #lockdown Tweets
SHOW GRAPHS
CLEAR ALL
These graphs represent the data analysis about tweets posted between 05-05-2020 and 01-07-2020.
Sentiment Analysis
Pie Chart of Polarity
Positive
Negative
Neutral
Emotion Analysis
Bar Graph of Emotions
Tweets Count
120
100
80
60
40
20
0
Happy
Angry
Fearful
Sad
Average
Surprise


Live Analysis
Live Tweet Analysis
12
10
8
6
4
2
0
-2
-4
-6
-8
20:37:00
20:39:00
20:41:00
20:43:00
20:45:00
20:47:00
20:49:00
20:51:00
20:53:00
20:55:00
20:57:00
20:59:00
21:01:00
21:03:00
CNN's Dana Bash compares Vice President Mike Pence's remarks at the coronavirus task force briefing to the infamous Vietnam-era 'Five o'clock follies' where the government provided a version of events that were very different from the reality people were seeing.
0
-10
units
10
SHOW TWEETS AND ANALYSIS
CLEAR ALL


Check Your Own Sentiment & COVID India Tracker
Check Your Own Sentiment!
Enter Your Tweet
the number of covid cases are increasing day by day
ENTER
CLEAR ALL
Sentiment Score
-5
units
-10
10
This is a Negative Tweet.
COVID India Tracker
Total Cases : 697413
Active Cases : 253287
Recovered : 424433
Active
Deaths
Recovered
REFRESH