Sentiment Analysis of COVID-19 Tweets – Visualization Dashboard

Submitted by:

Application ID: SPS_CH_APL_20200005849

Team Name: Zerobugz

Name: Manasa

SBID: SB20200047593

Email: manasa7919@gmail.com

Name: Narendra

SBID: SB20200068820

Email: narendranairy19@gmail.com

Name: Nikhil V

SBID: SB20200047588

Email: nikhilshettigar269@gmail.com

1. INTRODUCTION

1.1 Overview

Sentiment analysis nowadays can be considered as one of the most popular research topics in the field of natural language processing. The uses of sentiment analysis are covered by some interesting scientific and commercial areas, such as opinion mining, recommender systems and event detection. The events occurring in normal daily life are discussed on social media and any individuals are free to discuss and express their opinion about these events. Coronavirus known as COVID-19 has been one of the most discussed and one of the most spreading diseases worldwide. We have developed a model TO analyze the sentiments of people in this pandemic and also estimated what would their state of mind if lockdown is extended.

1.2 Purpose

As governments and organizations continue to work towards COVID-19 and stem the growing humanitarian toll it is exacting, the economic effects are also beginning to be felt. We can track sentiment to gauge how people's expectations, incomes, spending, and behaviors change throughout the crisis across the country over time and also predict their state if the lockdown is extended. The project mainly focus on people's sentiment towards the pandemic, understand the sentiments of people on government's decisions to extend the lockdown and possibility to predict riots against the government

2.LITERATURE SURVEY

2.1 Existing Problem

The sentiment analysis of Indians after the extension of lockdown announcements to be analyzed with the relevant #tags on twitter and build a predictive analytics model to understand the behavior of people if the lockdown is further extended. A machine learning model much be done which leverages historical data to predict insights into the future. This problem statement is aimed at analyzing sentiments of people about the epidemic and also drawing insights about how people might react if government extends the lockdown.

2.2 Proposed Solution

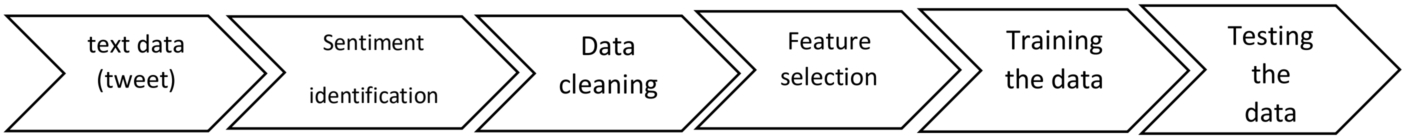
The solution to the problem can be developing a machine learning model using IBM cloud. The dataset containing the twitter hashtags are trained in the model. The dataset contains all the covid-19 tweets. The keywords such as “corona”, "coronavirus", "covid", "pandemic", "lockdown", "quarantine", "hand sanitizer", "ppe", "n95", different possible variants of "sarscov2", "nCov", "covid-19", "ncov2019", "2019ncov", "flatten(ing) the curve", "social distancing", "work(ing) from home" are used as a reference to extract only the related tweets which are used with #. In the obtained data, only the tweets from India is taken by eliminating the unnecessary rows. Firstly, this dataset is imported in the Jupiter notebook. The front end displays a pie chart which shows sentimental analysis of the people who have tweeted between March 19 and June 23. This graph will have 3 parts negative, positive and neutral reactions.

Later the future prediction is to be calculated based on the dataset we have. The dataset used for this is dated from March 19. We predict the future date using LSTM (Long Short Term Memory networks).We plot a graph of the outcome by considering sentiment values in y axis and dates in x axis. By looking at this we can estimate what the people feel if the lockdown is extended. The front end is done using react JS and Node red.

3.THEORETICAL ANALYSIS

3.1 Block diagram

Sentiment Analysis has a certain procedure that begins with grabbing the collected data, then identifying the data. Later on, the required features will be extracted to the next step which is sentiment classification. Then the dataset is cleaned by keeping only the required data. After this the feature selection is done. Finally, the data is trained and tested to complete the model. The flow of the model is as below



According to the tweets and sentiments we plot a graph which shows the present sentiments of the people. This can be all three parts negative, positive and neutral sentiments. This shows what people had gone thought during the lockdown.

For the future prediction we have used Long Short-Term Memory (LSTM) networks as layers in the sequential algorithms. LSTM are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Sequence prediction problems have been around for a long time. These include a wide range of problems; from predicting sales to finding patterns in stock markets' data, from understanding movie plots to recognizing your way of speech, from language translations to predicting your next word on your iPhone's keyboard. Using this we can predict the future date and assign the sentiment value accordingly.

3.2 Hardware / Software Designing

The project is implemented in Google Colab notebook which can be opened in any browser in your PC. We have also made use of IBM cloud services. It provides various services such as IBM Watson studio, IBM machine learning service, IBM auto AI feature, Databases and Node-red. It also provides virtual systems to run our model. In this project we have used DB2 which is a database used to store all the data. The dataset is filtered using python to get the tweets related to India. This is done using spider. Hydrator is used to extract the tweets from corresponding tweet ID. We have used Jupiter notebooks to train the date and build the model. The visual studio is used in the project as an editor to code the UI. The IBM service node-red is used for UI.

4. EXPERIMENTAL INVESTIGATION

Dataset is downloaded from https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset. Import the database and draw a pie chart by classifying the sentiments as positive negative and neutral. According to that we have noticed that positive sentiments occupy a large portion (51.5%) which tells that people are enjoying this period staying home. It is one period where the entire family will gather together and spend quality time. 51.5% according to the dataset is making good use of this period. The neutral sentiments occupy 32% of the total data which makes it second highest. This type indicates that 32% of the people have mixed feeling regarding the lockdown.16.5% of people have negative opinions this may tell us that people are facing difficulty in this time. Which may include financial problems, away from home, loss of job boredom etc.

The line graph predicts sentiments of people in future dates if the lockdown is extended, where X axis represents the dates starting from 19 March until 31st July, and Y axis represents the sentimental values. We have taken the average of everyday tweets sentiments and multiplied it with 1000 in order to get maximum difference between corresponding points. The values till 23rd June are trained and are tested on data from June 23 to June 30 and the prediction has been extended till July 31st.

From the pie chart it is observable that Positive Sentiments > Neutral Sentiments > Negative Sentiments and in the graph maximum of average sentiment values are between 50 and 175 and also the future predicted values are within the same range.

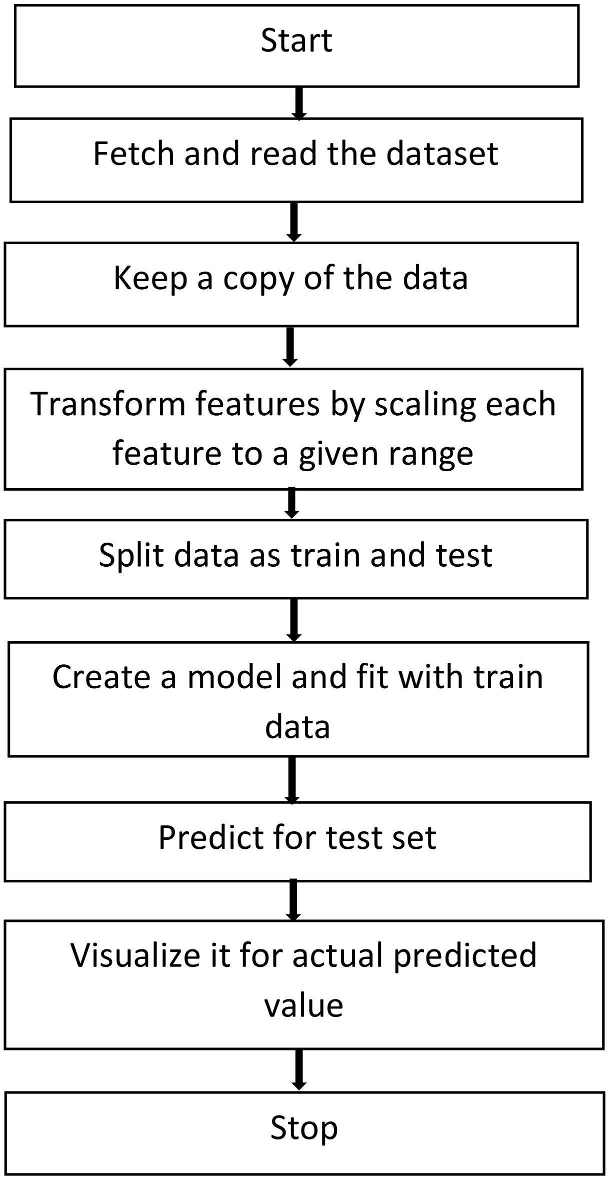
This gives a conclusion that if government decides to extend the lockdown there will be more positive sentiments and neutral sentiments due to the uncontrolled increase of number of patients effected due to corona, people think of protecting themselves so support the lockdown.

There might be some number of negative sentiments also with regards to increase in number of deaths and unemployment and lack of money due to no work indicating such situations.

There is possibility that this model might not be accurate but the past responses of people towards corona and lockdown shows that they will support the extension of lockdown.

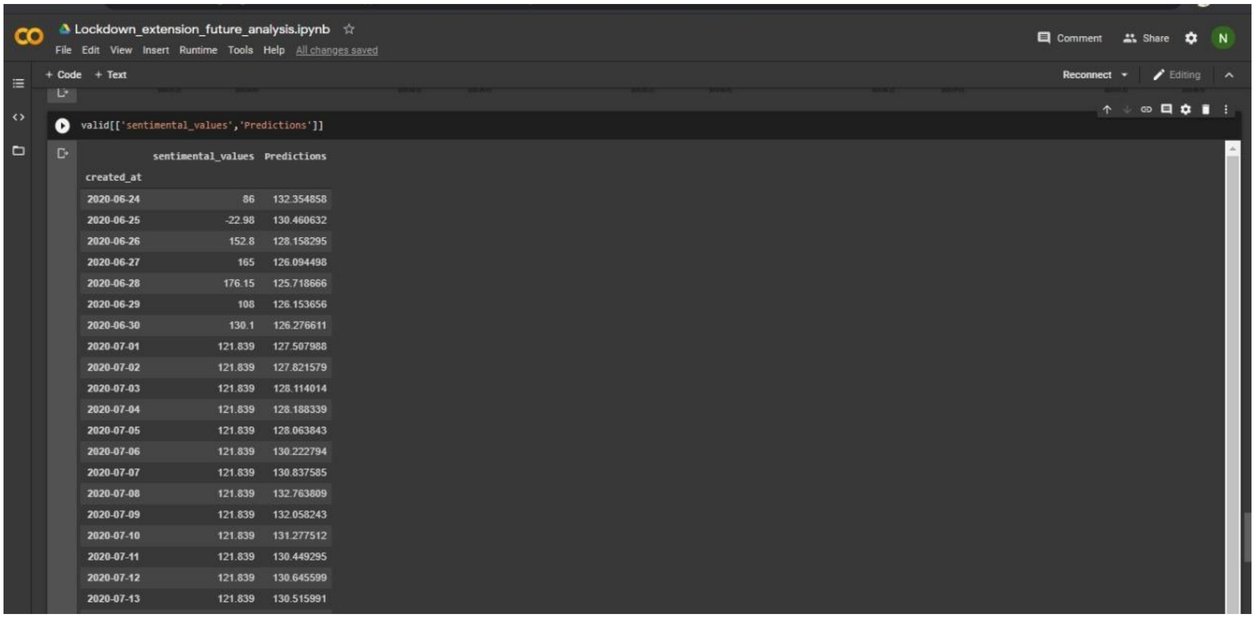
5. FLOW CHART

Model flow chart for predicting sentiments when lockdown is extended


Start
Fetch and read the dataset
Keep a copy of the data
Transform features by scaling each feature to a given range
Split data as train and test
Create a model and fit with train data
Predict for test set
Visualize it for actual predicted value
Stop

6. RESULT

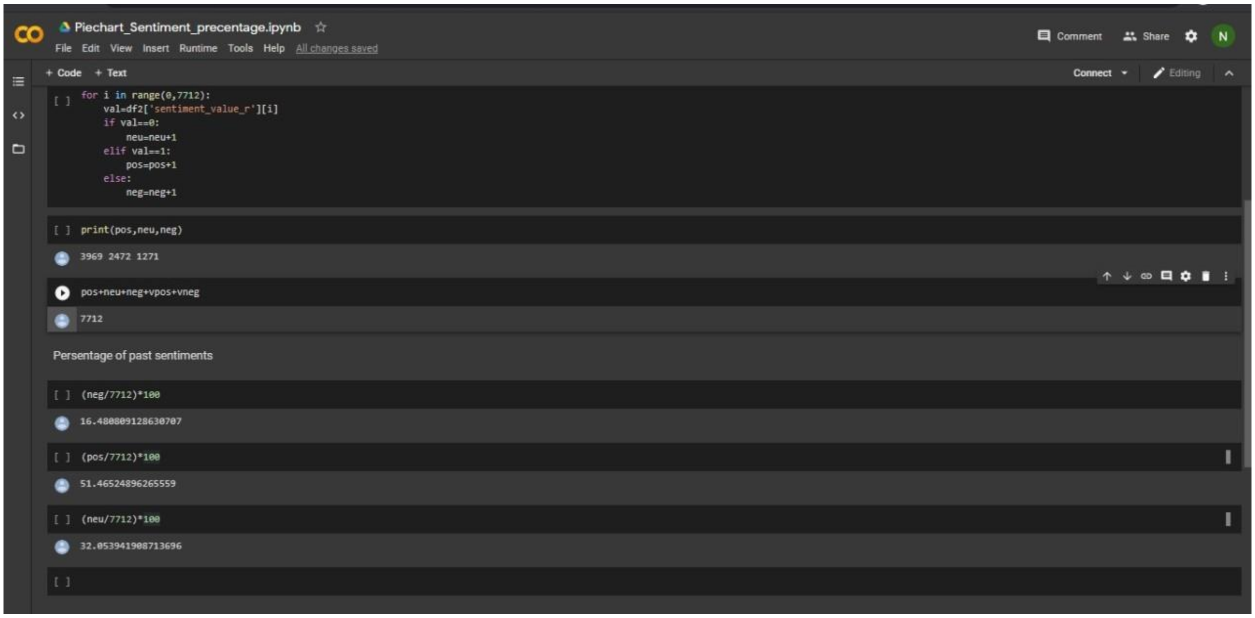
Actual and Predicted values from our LSTM model:


Lockdown_extension_future_analysis.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Reconnect
Editing

```
valid[['sentimental_values','Predictions']]
```

```
            sentimental_values  Predictions
created_at
2020-06-24                  86   132.354858
2020-06-25              -22.98   130.460632
2020-06-26               152.8   128.158295
2020-06-27                 165   126.094498
2020-06-28              176.15   125.718666
2020-06-29                 108   126.153656
2020-06-30               130.1   126.276611
2020-07-01             121.839   127.507988
2020-07-02             121.839   127.821579
2020-07-03             121.839   128.114014
2020-07-04             121.839   128.188339
2020-07-05             121.839   128.063843
2020-07-06             121.839   130.222794
2020-07-07             121.839   130.837585
2020-07-08             121.839   132.763809
2020-07-09             121.839   132.058243
2020-07-10             121.839   131.277512
2020-07-11             121.839   130.449295
2020-07-12             121.839   130.645599
2020-07-13             121.839   130.515991
```

Calculating the percentage to display on pie chart:


Piechart_Sentiment_precentage.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Connect
Editing

```
for i in range(0,7712):
    val=df2['sentiment_value_r'][i]
    if val==0:
        neu=neu+1
    elif val==1:
        pos=pos+1
    else:
        neg=neg+1
```

```
print(pos,neu,neg)
```

```
3969 2472 1271
```

```
pos+neu+neg+vpos+vneg
```

```
7712
```

Persentage of past sentiments

```
(neg/7712)*100
```

```
16.480809128630707
```

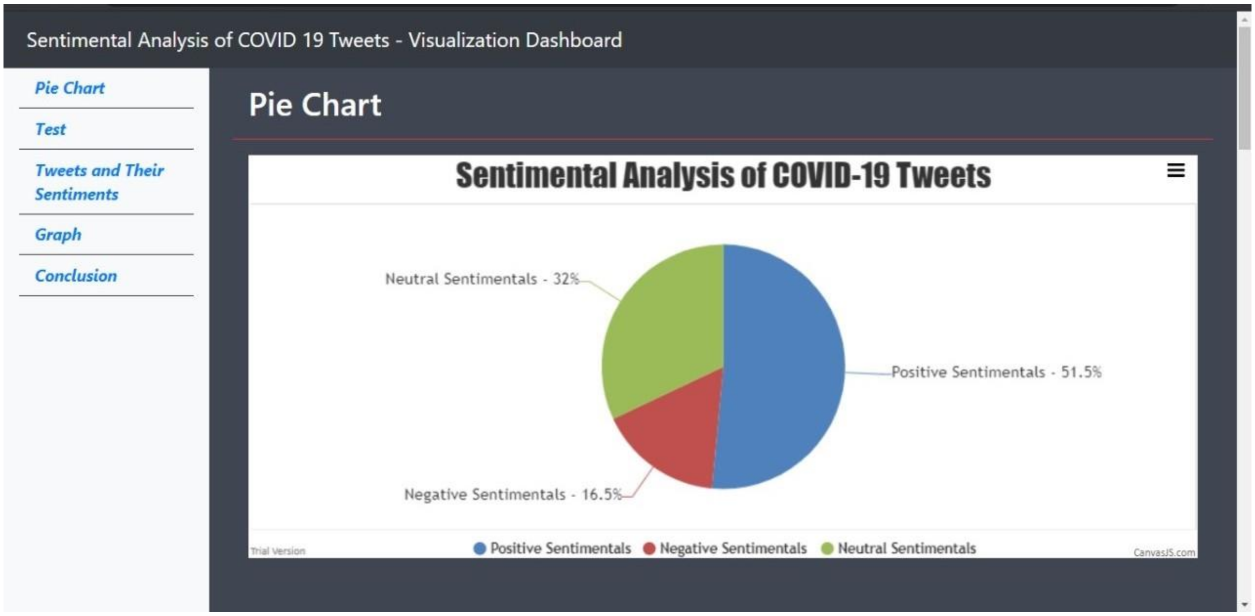
```
(pos/7712)*100
```

```
51.46524896265559
```

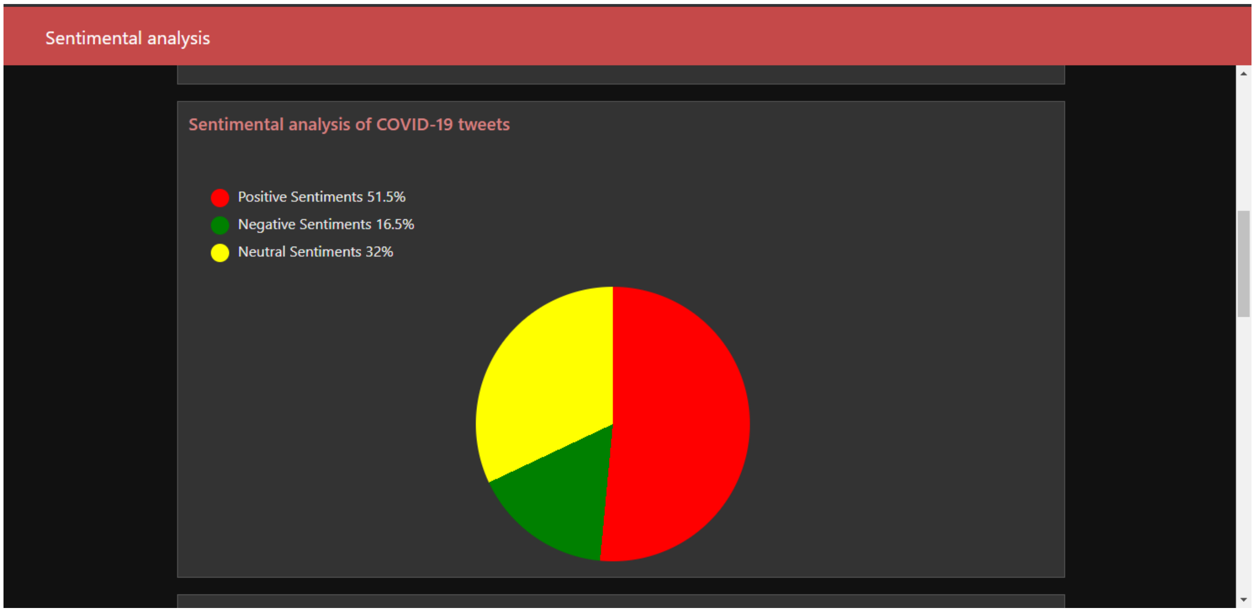
```
(neu/7712)*100
```

```
32.053941908713696
```

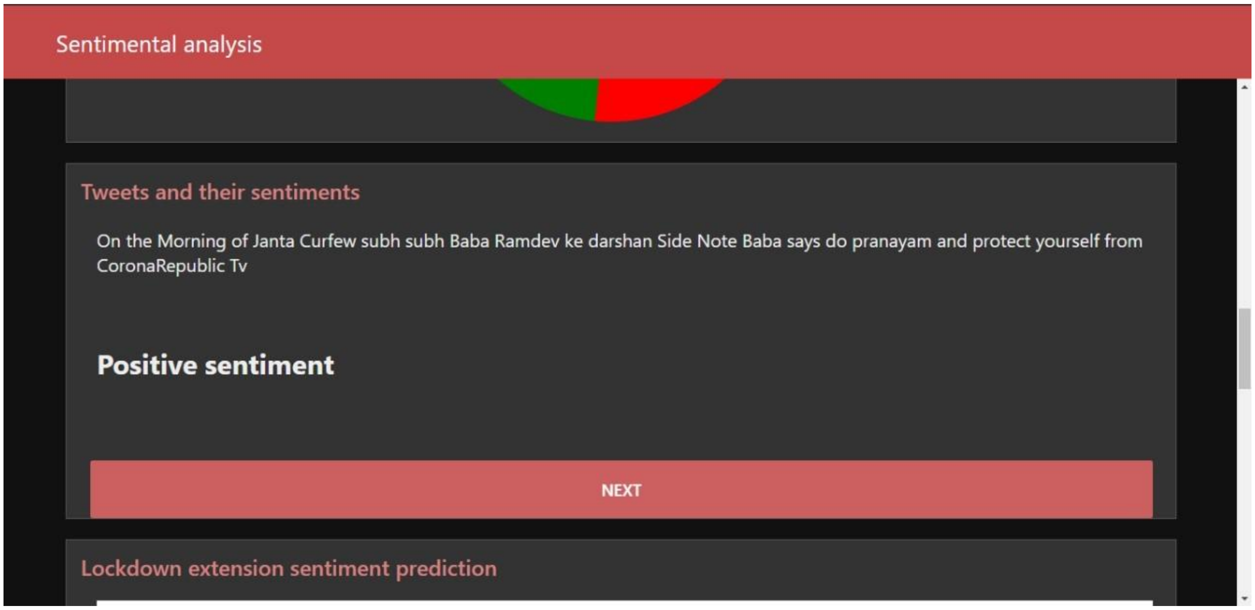
Displaying the percentage of the sentiments of the data in the form of pie chart in React:


Sentimental Analysis of COVID 19 Tweets - Visualization Dashboard
Pie Chart
Test
Tweets and Their Sentiments
Graph
Conclusion
Pie Chart
Sentimental Analysis of COVID-19 Tweets
Neutral Sentimentals - 32%
Positive Sentimentals - 51.5%
Negative Sentimentals - 16.5%
● Positive Sentimentals
● Negative Sentimentals
● Neutral Sentimentals
Trial Version
CanvasJS.com

Displaying the percentage of the sentiments of the data in the form of pie chart in Node red:


Sentimental analysis
Sentimental analysis of COVID-19 tweets
● Positive Sentiments 51.5%
● Negative Sentiments 16.5%
● Neutral Sentiments 32%

Displaying a random tweet from the dataset and its sentimental value:



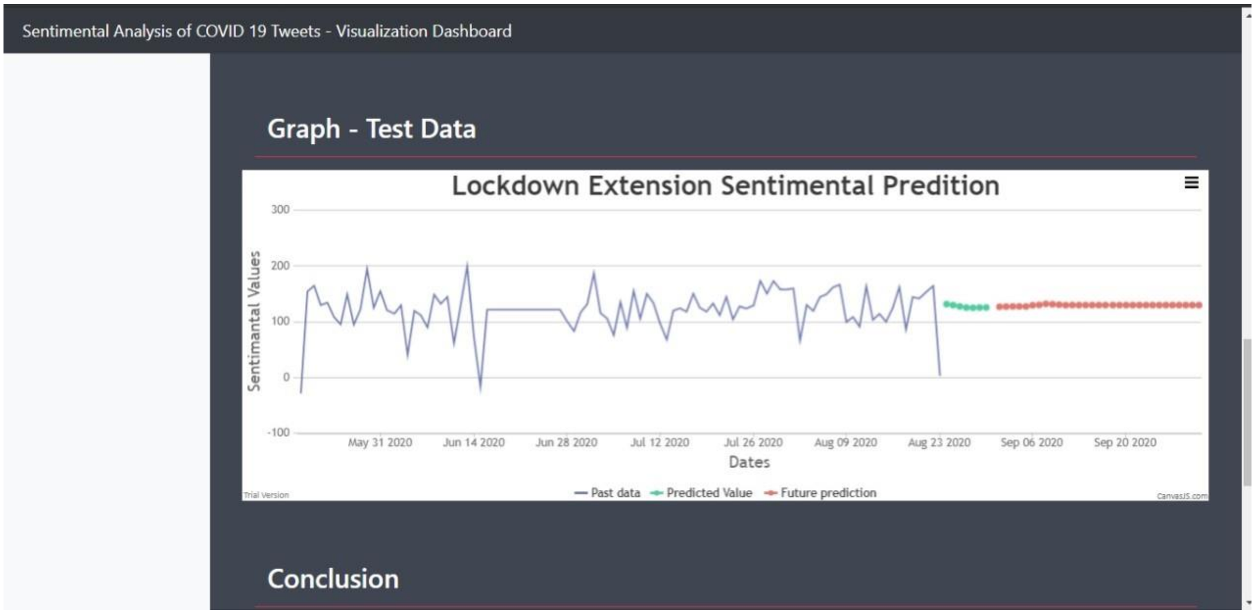
Tweets and their sentiments

On the Morning of Janta Curfew subh subh Baba Ramdev ke darshan Side Note Baba says do pranayam and protect yourself from CoronaRepublic Tv

Positive sentiment

Lockdown extension sentiment prediction

Displaying sentiments of people when lockdown is extended using React

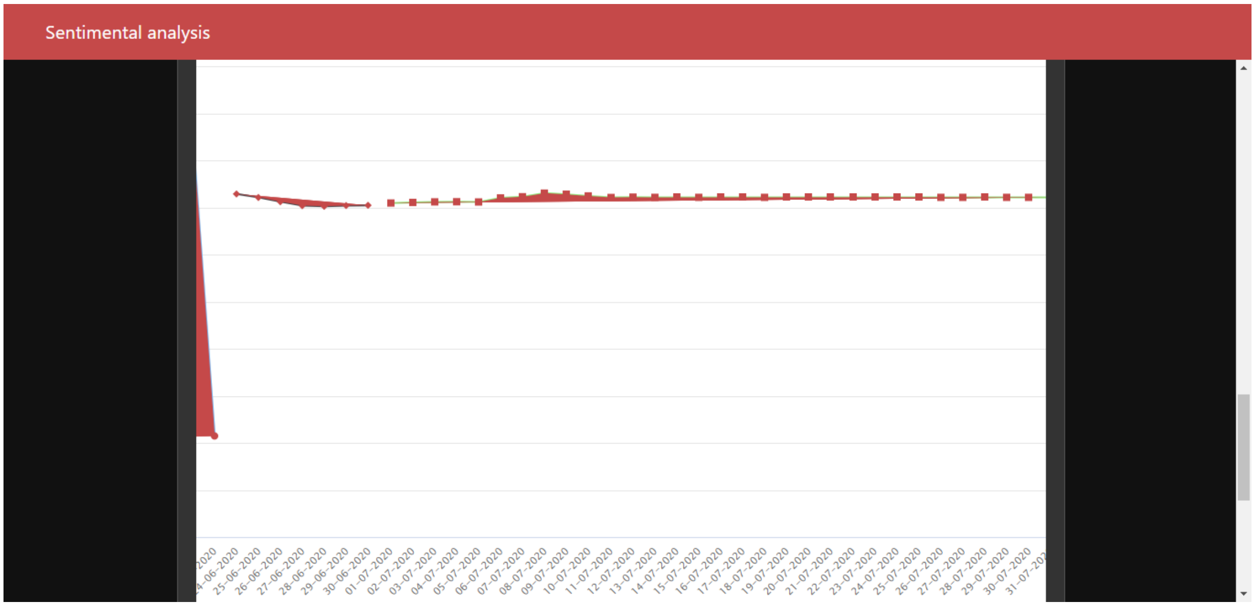


Sentimental Analysis of COVID 19 Tweets - Visualization Dashboard

Graph - Test Data

Conclusion

Displaying sentiments of people when lockdown is extended using Node red



Sentimental analysis

7. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

1. This helps the government to know what their citizens are feeling and what they expect from the government. The government shall put light on these and apply in their new policies.
2. The sentiment analysis will help many organizations, people and even common people to took through a large data in a single website. This saves their time.
3. Helps to know the situations using a single web page. We need not go through thousands of data.

DISADVANTAGES:

1. The current dataset focused on a textual corpus consisting of Tweets filtered by only a few COVID 19 related keywords. There might be some words that are not taken and the projects become less generalized.
2. There might be some of the word in the tweet which are mentioned in their regional language. This model fails the cases when people express their sentiments in their regional languages or languages other than English.
3. The prediction is only limited to a few section of India. We know that many of the people in Indian are least educated and do not know much about the social platform twitter.

8. APPLICATIONS

1. Recommendation mining: The model predicts the sentiments of people regarding the COVID-19 pandemic which may help to recommend the government on what steps must be taken.
2. Opinion mining: Many organisations may want to know the views of their employees and customers from this application it will become easy for the organization. Similarly, government can take opinions of the citizen.

9. CONCLUSION

The project is sentiment Analysis of COVID-19 Tweets where we take a datset of tweets from the month of March to predict the output. From the dataset we classify sentiments as positive, negative or neutral and pie chart accordingly. The tweets from June 3 is trained to predict the sentiments of people when lockdown is extended. This is modelled using sequential algorithm which apply LSTM in the layers. A line graph is plotted based on this. This can help to predict the sentiments of people when lockdown is extended.

10. FUTURE SCOPE

1. Since there are so many professional and official people on Twitter, you may find more reliable source of information on Twitter than other social medias such as Facebook, Wechat, Instagram. However, it is very essential to explore other social media with regard to sentiment analysis.
2. This research application is not only convenient for Coronavirus health issue but it can also be adopted as model to discover sentiment emotion for the future similar cases.
3. In our project, we chose sequential algorithm but there are other models that may provide interesting results such as lexicon-based algorithms.
4. This project is just limited to the sentiments of Indian citizens. So we can train all covid19 tweets of worldwide and predict the result.

11. BIBLIOGRAPHY

APPENDIX

A. Source code

LSTM

```
#import packages
import pandas as pd
import numpy as np

#to plot within notebook
import matplotlib.pyplot as plt
# %matplotlib inline

#setting figure size
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 20,10

#for normalizing data
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))

#read the file
df = pd.read_csv('/content/date_sentiment3.csv',  encoding='ISO-8859-1')

#print the head
df.head()

df.isnull().sum()

df.shape

df.mean()

df = df.fillna(df.mean())
df.info()

df.isnull().sum()

df.shape

#setting index as date
df['created_at'] = pd.to_datetime(df.created_at,format='%Y-%m-%d')
```

```
df.index = df['created_at']

#plot
plt.figure(figsize=(50,30))
plt.plot(df['sentimental_values'], label='Close Price history')

from sklearn.preprocessing import MinMaxScaler
from keras.models import Sequential
from keras.layers import Dense, Dropout, LSTM

#creating dataframe
data = df.sort_index(ascending=True, axis=0)
new_data = pd.DataFrame(index=range(0,len(df)),columns=['created_at', 'sentimental_values'])
for i in range(0,len(data)):
    new_data['created_at'][i] = data['created_at'][i]
    new_data['sentimental_values'][i] = data['sentimental_values'][i]

new_data.index = new_data.created_at
new_data.drop('created_at', axis=1, inplace=True)

#creating train and test sets
dataset = new_data.values

train = dataset[0:97,:]
valid = dataset[97:,:]

scaler = MinMaxScaler(feature_range=(0, 1))
scaled_data = scaler.fit_transform(dataset)

scaled_data

x_train, y_train = [], []
for i in range(12,len(train)):
    x_train.append(scaled_data[i-12:i,0])
    y_train.append(scaled_data[i,0])
x_train, y_train = np.array(x_train), np.array(y_train)

x_train = np.reshape(x_train, (x_train.shape[0],x_train.shape[1],1))

# create and fit the LSTM network
model = Sequential()
model.add(LSTM(units=50, return_sequences=True, input_shape=(x_train.shape[1],1)))
model.add(LSTM(units=50))
model.add(Dense(1))

model.compile(loss='mean_squared_error', optimizer='adam')
model.fit(x_train, y_train, epochs=15, batch_size=1, verbose=2)

inputs = new_data[len(new_data) - len(valid) - 12:].values
inputs = inputs.reshape(-1,1)
inputs  = scaler.transform(inputs)

inputs.shape

X_test = []
for i in range(12,inputs.shape[0]):
    X_test.append(inputs[i-12:i,0])
X_test = np.array(X_test)

X_test = np.reshape(X_test, (X_test.shape[0],X_test.shape[1],1))
closing_price = model.predict(X_test)
closing_price = scaler.inverse_transform(closing_price)

rms=np.sqrt(np.mean(np.power((valid-closing_price),2)))
rms

#for plotting
train = new_data[:97]
valid = new_data[97:]
valid['Predictions'] = closing_price
plt.figure(figsize=(50,20))
plt.plot(train['sentimental_values'])
plt.plot(valid[['sentimental_values','Predictions']])

valid[['sentimental_values','Predictions']]
```

Calculating percentage of tweets

```
import pandas as pd

df2 = pd.read_csv('D:\\Nikhil\\dcumnts\\Certificates\\others\\IBMhc\\ds\\date_sentiment.csv',
encoding='ISO-8859-1')

df2.shape

pos=0
neu=0
neg=0

for i in range(0,7712):
    val=df2['sentiment_value_r'][i]
    if val==0:
        neu=neu+1
    elif val==1:
        pos=pos+1
    else:
        neg=neg+1

print(pos,neu,neg)

pos+neu+neg+vpos+vneg

(neg/7712)*100

(pos/7712)*100

(neu/7712)*100
```