

IBMHACKCHALLENGE2020

**Problem Statement:
Sentiment Analysis
Of
COVID-19
Tweets-Visualization Dashboard**

**TEAM NAME:
SUPERSONICS**

**TEAM MEMBERS:
D.BALA KOTESWARA SASTRY
M.BINDU SRI**

CONTENTS

1. INTRODUCTION

1.1. Overview

1.2. Purpose

2. LITERATURE SURVEY

2.1. Existing problem

2.2. Proposed solution

3. THEORITICAL ANALYSIS

3.1. Block diagram

3.2. Hardware / Software designing

4. EXPERIMENTAL INVESTIGATIONS

5. ACTIVITIES

6. FLOWCHART

7. RESULT

8. ADVANTAGES & DISADVANTAGES

9. APPLICATIONS

10. CONCLUSION

11. FUTURE SCOPE

12. APPENDIX

A.Source code

13. BIBILOGRAPHY

INTRODUCTION

Overview

The outbreak of coronavirus disease 2019 (COVID-19) has created a global health crisis that has had a deep impact on the way we perceive our world and our everyday lives.

"Fighting an unknown, unseen enemy, with no definite cure in sight, lots of misinformation and disinformation is being circulated on the internet. The accidental spread of misinformation is a menace and causes fear amongst the people."

Thus, it is crucial to understand public sentiments under COVID-19.



Purpose

The Corona Virus endangers our physical health indeed. The coronavirus (COVID-19) pandemic has spread across 190 countries infecting 4.2 lakhs people and killing 16,500 so far. On the other side, social distancing also poses a threat to our emotional stability. Thus, it is crucial to understand public sentiments under COVID-19.

LITERATURE SURVEY

Existing problem

The major problem we want to solve is that **How are public sentiments changing under Covid-19?**

We try to offer an analytical view to the public that how public sentiments changes with the development of Covid 19 and their effects.

Social media like Twitter best represent public sentiments. To narrow down the condition, we choose Twitter to analyse.

-Do people become more panic as coronavirus spreads?

-Do people become more pessimistic?

-How can we help when sense a status change?

Proposed solution

Our approach to tackle the problem can be divided into 3 steps:

We mainly utilized web scraping and APIs to collect the data we need. For Twitter, we used TwitterScraper API to retrieve 80-day Twitters and web scraped corresponding trending topics.

We applied NLP algorithms to analyze sentiment of Twitter.

Searching keywords in twitter is one of the hardest tasks because of the diversity of the language and the slangs used on the internet.

In the proposed system, the first step involves collection of tweets from twitter and making it as a data set, the second step is preprocessing of the related tweets. In the third step, sentiment analysis is performed using the Natural Language Processing (NLP) algorithm, which is based on numerical statistics . Assigned sentiment value using NLP, is used as a weighting factor in sentiment analysis. The logics that has been used in the proposed system has the following major steps:

- 👉 Collecting tweets

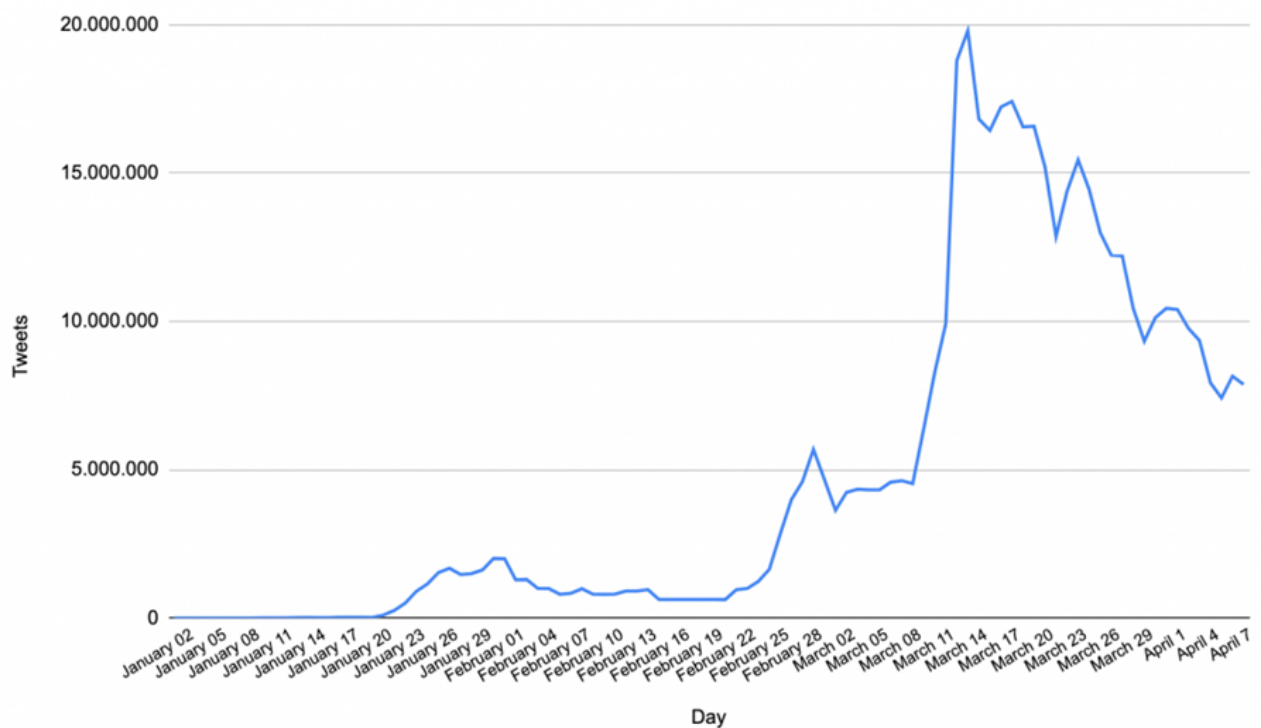
- 👉 Pre-processing tweets

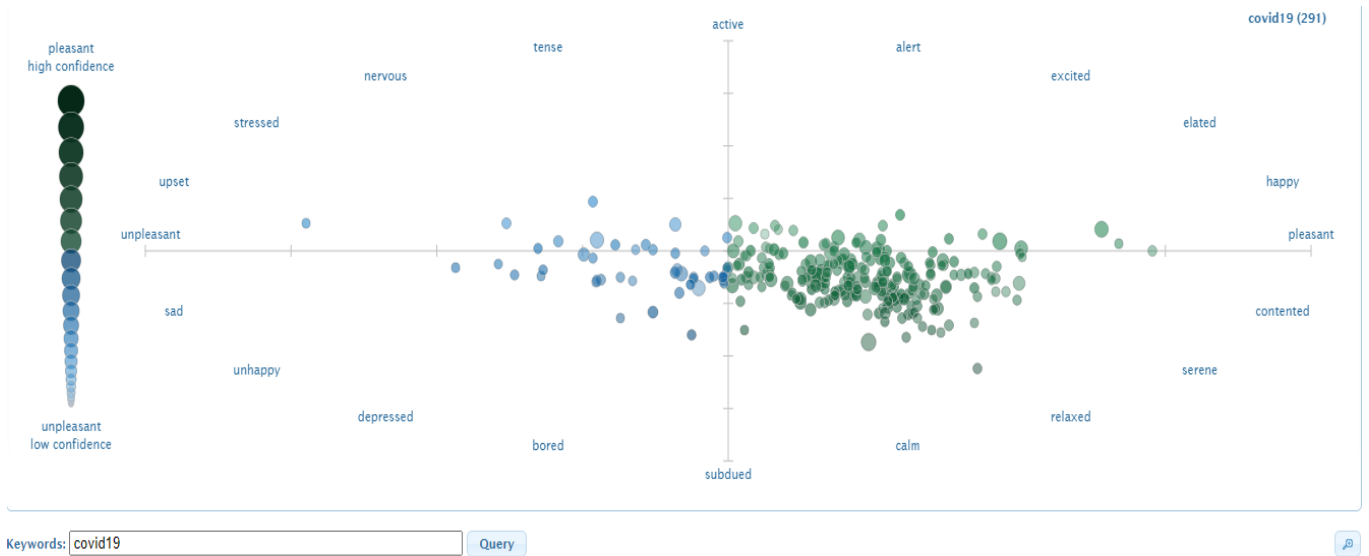
- 👉 Sentiment analysis

And, at last, an application-independent of a dataset only based on Twitter which shows Tweets related to COVID-19 with image and the link to that particular tweet or image or video on Twitter and also planning to do analyze and display Sentiment analysis of COVID-19 tweets.

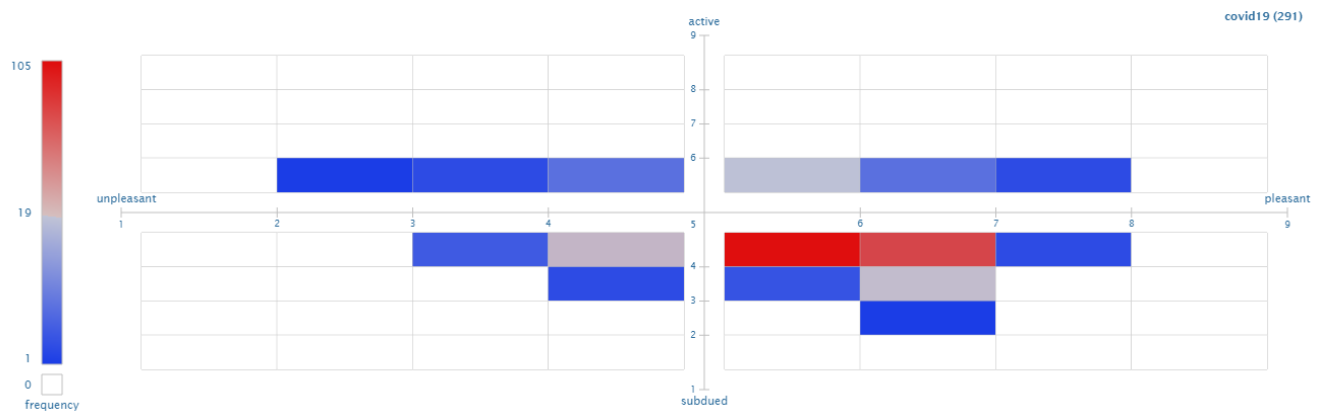
THEORITICAL ANALYSIS

#Covid19 #Coronavirus Tweets by day

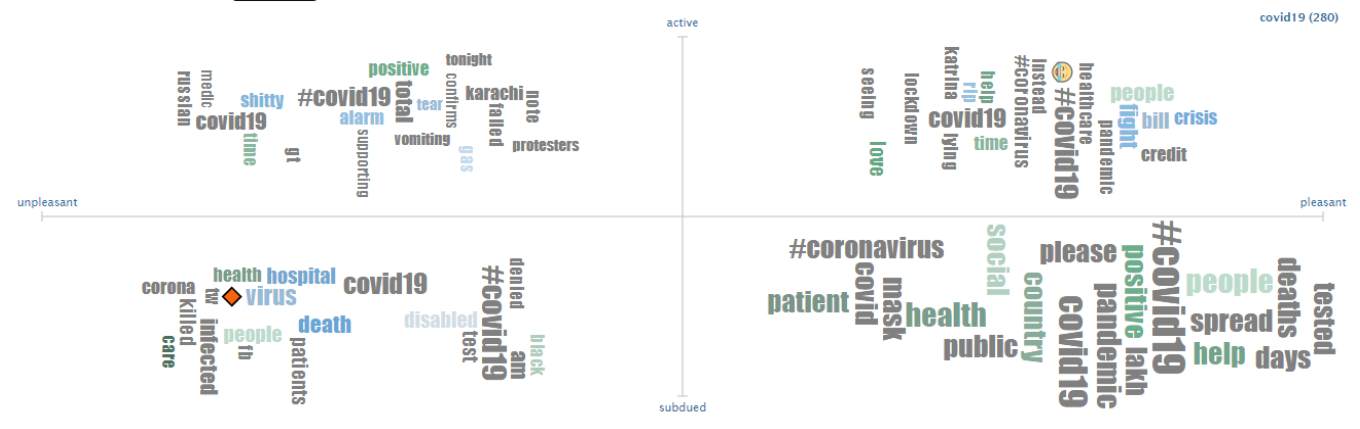




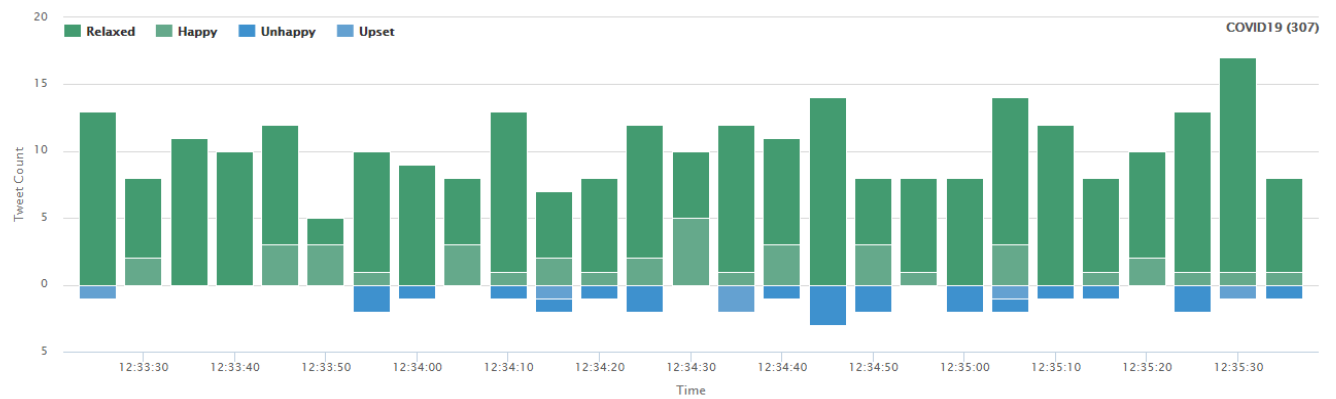
Heatmap:



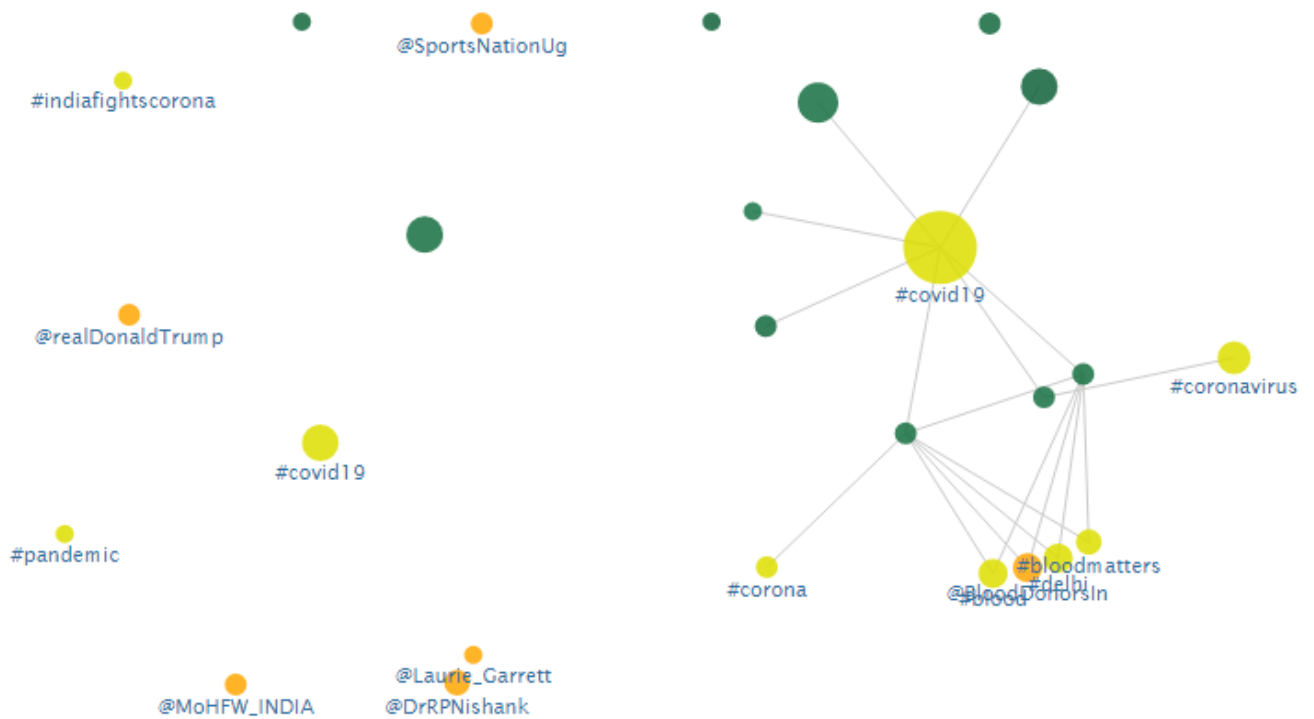
Tag cloud:



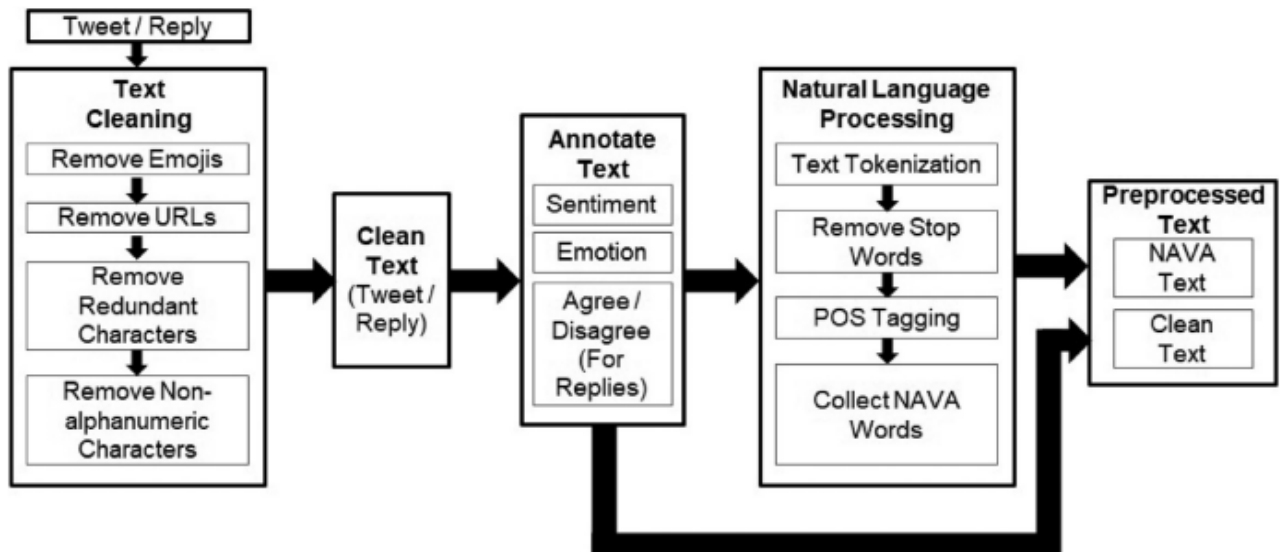
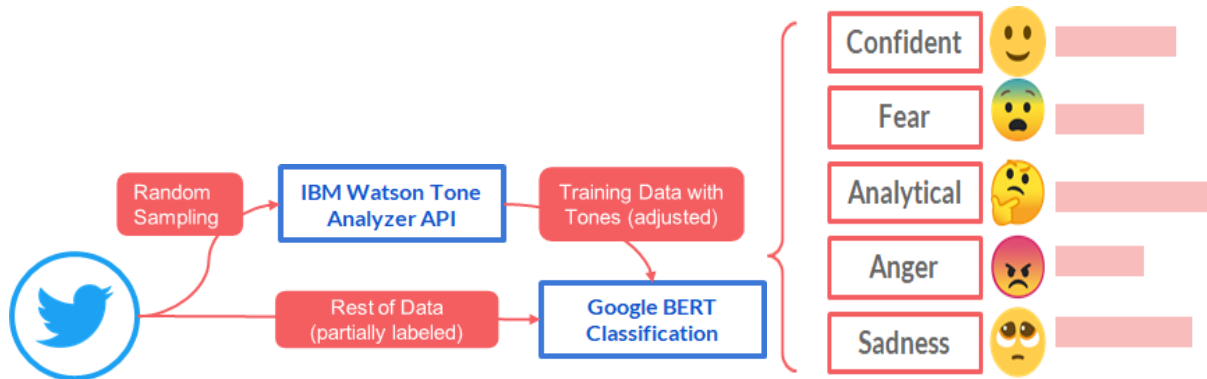
Timeline:

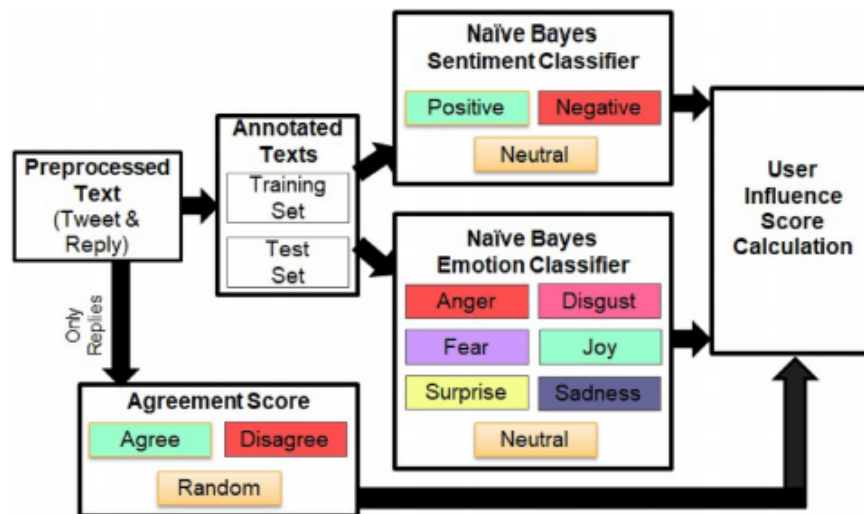


Affinity:



Block diagram



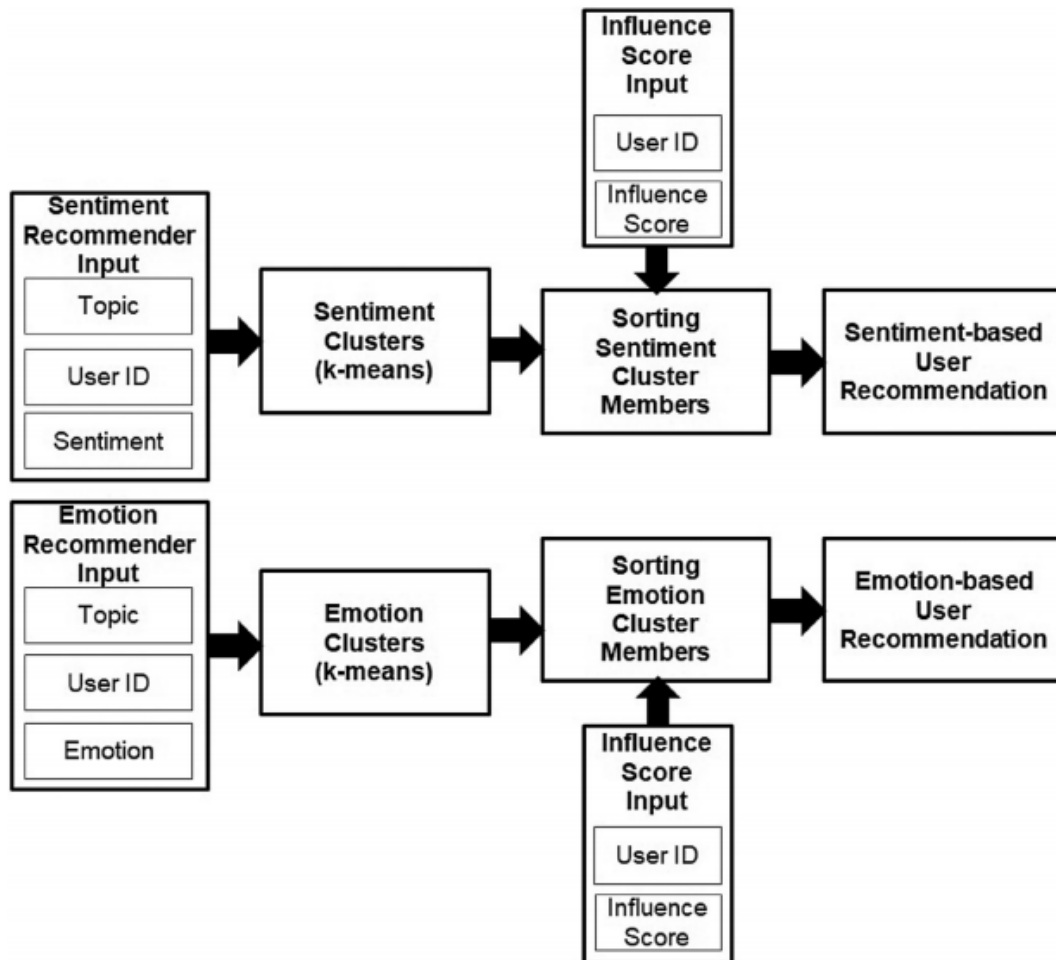
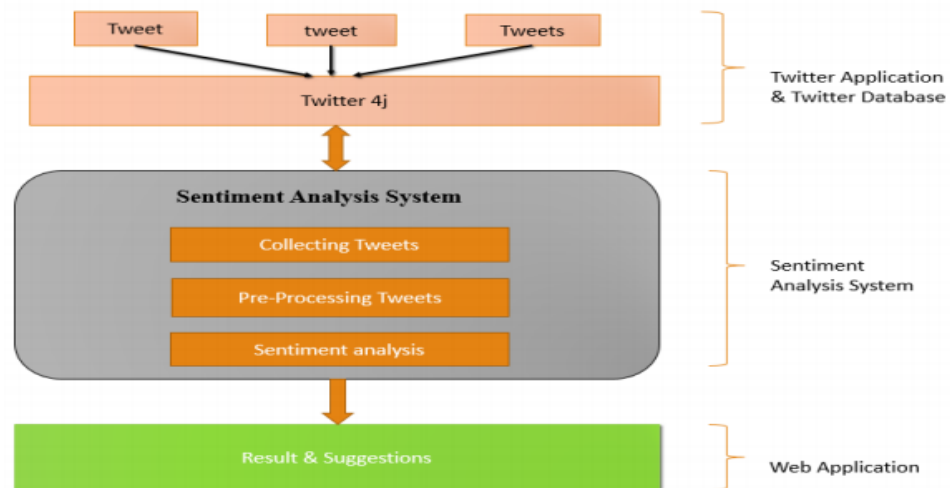


System Design

As twitter provides free APIs, it will be easy to collect and analyze the data using Twitter. Figure shows the basic architectural diagram of the implemented system.

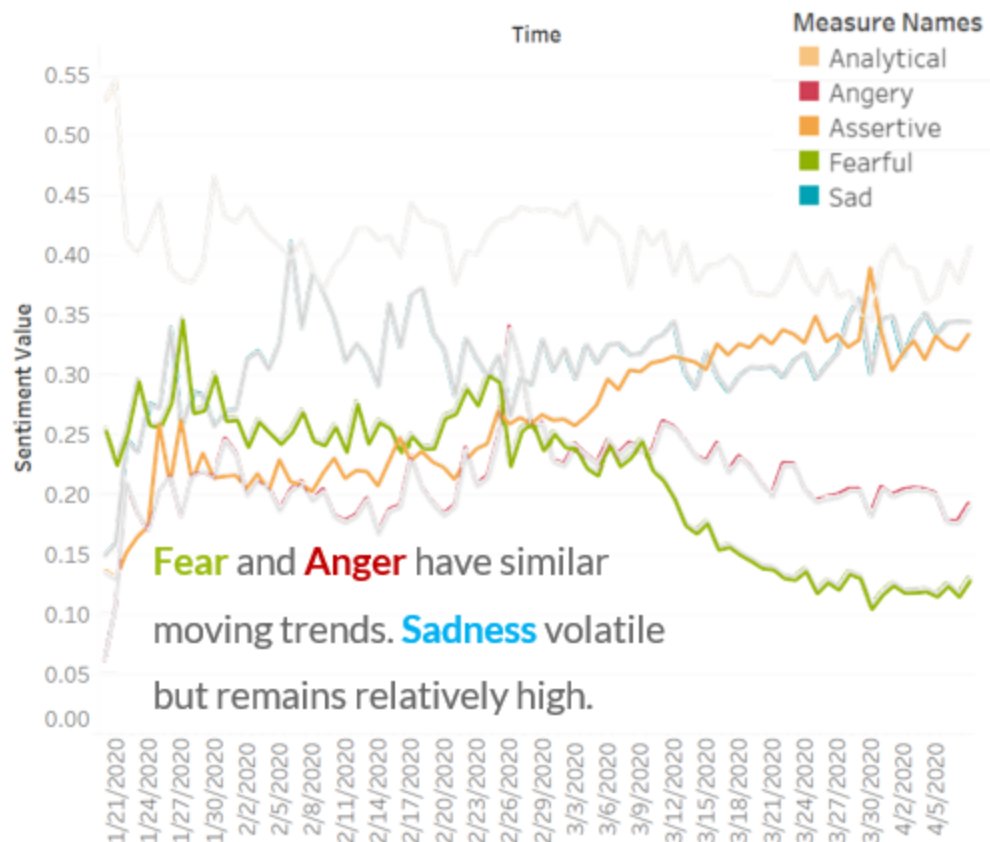
Basically, it consists three modules, they are:

- Twitter application and twitter database
- Sentiment analysis process
- Web application



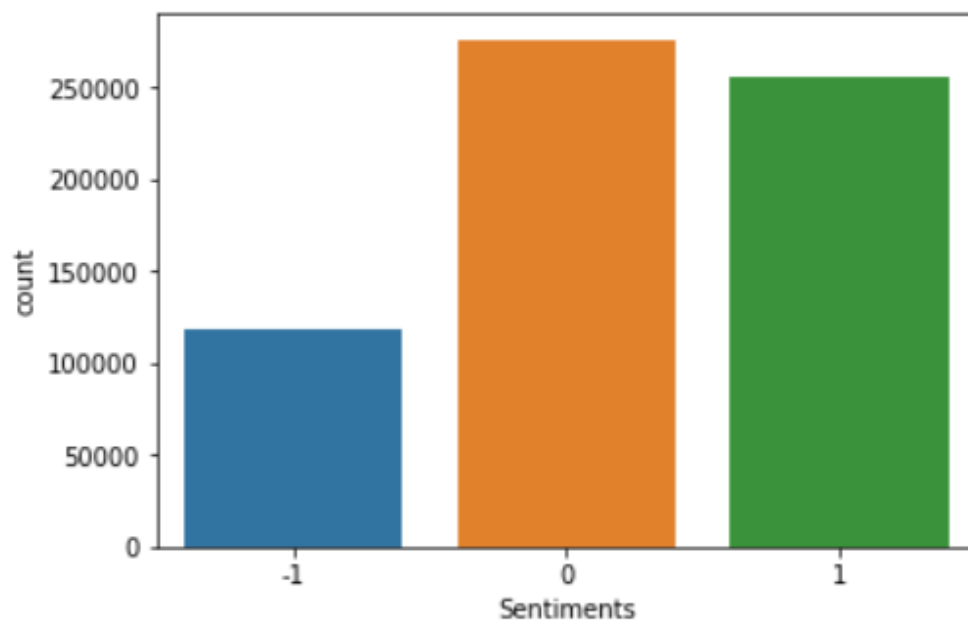
EXPERIMENTAL INVESTIGATIONS

5 sentiment trend lines reflect the trends of social mental status on coronavirus.

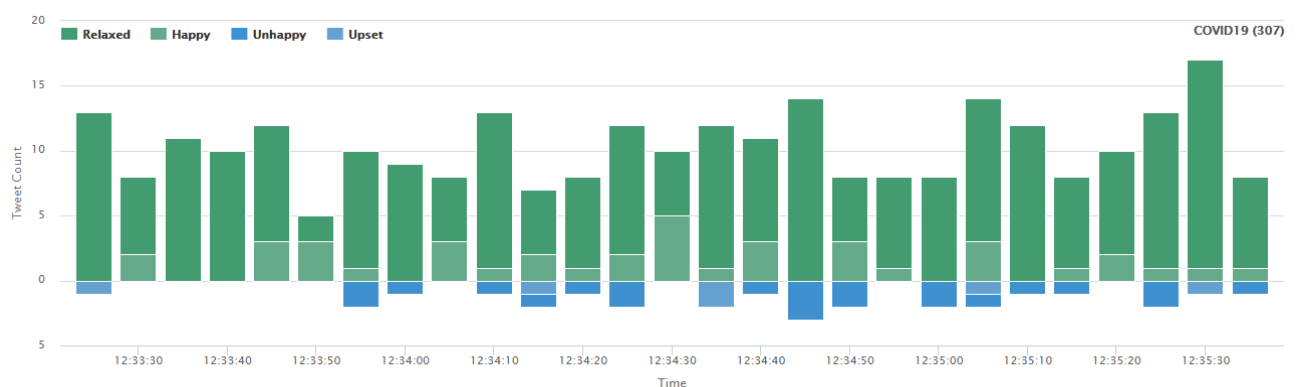


-Change in Each Sentiment: We can measure the change in each sentiments, and gain related news information that could cause the status changes

-Twitter Trending Tags: Twitter tags of the day serve as another reflection of people's attitudes towards the pandemic.



With all the information provided, we can extract the overall public sentiments. When observe a continuous increase or decrease of certain sentiments, we can investigate into it timely with the trending tags at the same time, and take action.



ACTIVITIES

Scope of work:

We are planning to take two datasets and also an application that is based only on Twitter, not on any dataset. In our 1st dataset, we are planning to analyze by using technologies like Google BERT, TextBlob, IBM Watson Tone Analyzer, and Jupyter Notebook. And, In our 2nd dataset, we are planning to analyze by using machine learning techniques like Naive Natural Language Processing (NLP), logistic regression, Naive Bayes, Decision tree, Random forest, and many more. And, at last, an application-independent of a dataset only based on Twitter which shows Tweets related to COVID-19 with image and the link to that particular tweet or image or video on Twitter and also planning to do analyze and display Sentiment analysis of COVID-19 tweets.

Dataset Selection:

For the dataset-1 we collected by using and its links:

- ✓ Tweets on Twitter retrieved with
<https://github.com/taspinar/twitterscraper>
- ✓ Twitter trending topics retrieved by scraping
<https://trendogate.com/>

- ✓ Number of tweets with #COVID-19 shared by <https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/>

To combined the datasets in folders, please try:

```
1 # Read raw datas from the raw data file
2 path = r'-----path-----'
3 files = os.listdir(path)
4 covid_twitter_data = pd.DataFrame()
5 # Concat the Twitters data into one-table
6 for file in files:
7     data = pd.read_csv(str(path) + file)
8     covid_twitter_data = covid_twitter_data.append(data, ignore_index=True)
```

For the dataset-2 we collected by using and its links:

- ✓ This dataset includes CSV files that contain tweet IDs. The model monitors the real-time Twitter feed for coronavirus-related tweets, using filters: language "English", and keywords "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" and the respective hashtag of all these keywords <https://ieee-dataport.org/keywords/covid-19-twitter-sentiment>

Working With IBM Cloud Account:

Create Cloud Account:

The screenshot shows the IBM Cloud Catalog interface. The top navigation bar includes the IBM Cloud logo, a search bar, and links to Catalog, Docs, Support, and Manage. The user's name, BALA KOTESWARA SASTRY DANDIBHOTLA, is displayed in the top right corner. The main content area features a large blue banner with the text "IBM Cloud products" and "Over 190+ products available for you to customize and build the solutions that you need for your business". Below the banner is a search bar labeled "Search the catalog...". On the left side, there is a sidebar with a "Catalog" section and a list of categories: IBM Cloud catalog, Featured, Services, Software, and Consulting. On the right side, there is a user profile section with a circular profile picture and the text "BALA KOTESWARA SASTRY DANDIBHOTLA". Below the profile picture are links for "Profile and settings", "Log in to CLI and API", "Privacy", "Feedback", and "Log out". At the bottom, there is a "Recommended for you" section with two cards: "Machine Learning" and "Watson Studio".

The screenshot shows the IBM Cloud console interface for a resource named "Node RED EJNLD". The top navigation bar is the same as the previous screenshot. The main content area has a header with the resource name "Node RED EJNLD", a status indicator "Running", and links for "Visit App URL" and "Add tags". On the left side, there is a sidebar with a "Resource list" section and a list of categories: Getting started, Overview, Runtime, Connections, Logs, API Management, Autoscaling, and Availability Monitoring. The "Overview" section is selected. The main content area displays the "Instances" section, which shows a "Health" status of "100%" and "1/1 instance(s) are running". Below this is a "MB memory per instance" slider set to 256. To the right, there is a "Runtime" section showing "SDK for Node.js™" and a circular progress indicator for "Total MB allocation" with a value of 256. Below the progress indicator, it says "0 MB still available" and "Used Free". At the bottom, there is a "Runtime cost" section with the text "Current and estimated cost excludes connected services." and a "Connections (1)" section with a link to "node-red-ejnld-cloudant-1593000403858-75232".

Open the Node-RED visual programming editor

Welcome to your new Node-RED instance on IBM Cloud

We know you're eager to start wiring up your flows, but first there are a couple of tasks you should do:

- Secure your Node-RED editor
- Learn how to install additional nodes



Previous

Next

Finish the install

You have made the following selections:

- *Not recommended:* Allow anyone to access the editor and make changes

You can change these settings at any time by setting the following environment variables via the IBM Cloud console:

- `NODE_RED_USERNAME` - the username
- `NODE_RED_PASSWORD` - the password
- `NODE_RED_GUEST_ACCESS` - if set to 'true', allows anyone read-only access to the editor



Previous

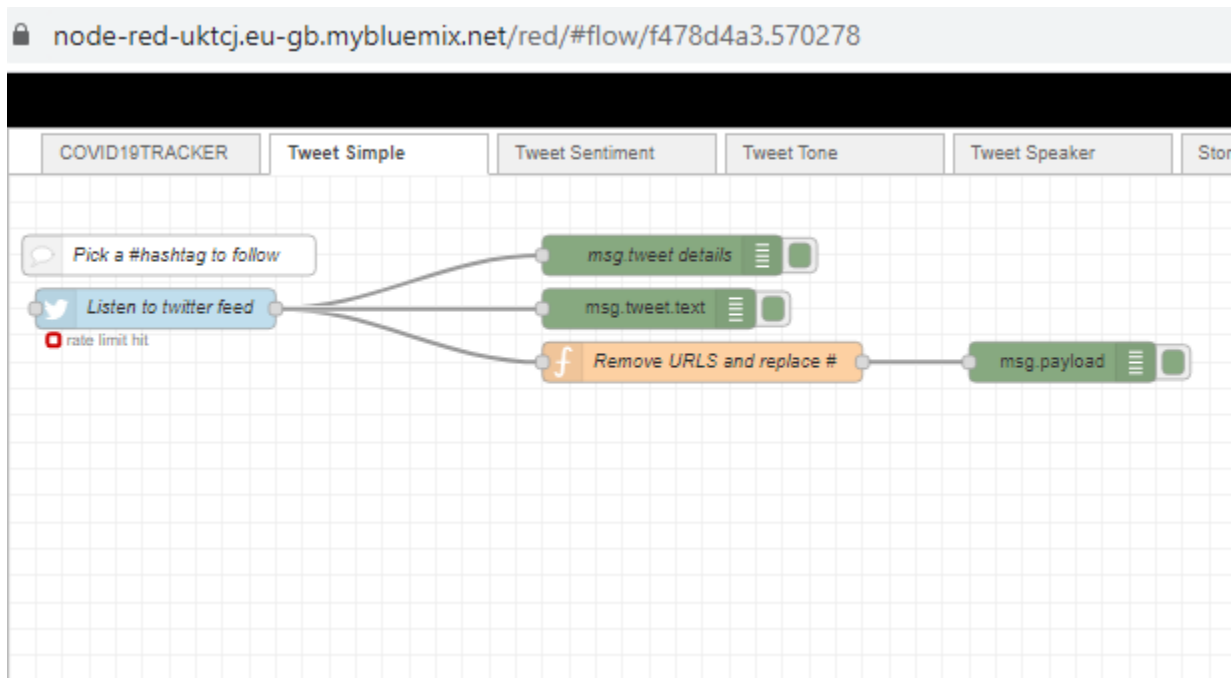
Finish

IBM Watson Tone Analyzer

Watson Tone Analyzer can be used to conduct social listening. Analyze emotions and tones in what people write online, like tweets or reviews. Predict whether they are happy, sad, confident, and more.

We have used the Watson Tone Analyzer service within our Node-RED Application. From IBM Cloud Dashboard <https://console.bluemix.net/dashboard/apps>

Created a flow for Displaying Twitter Feed



Edit twitter in node

Delete Cancel Done

▼ **node properties**

Twitter ID

Search

for

Name

Tip: Use commas without spaces between multiple search terms.
Comma = OR, Space = AND.
The Twitter API WILL NOT deliver 100% of all tweets.
Tweets of who you follow will include their retweets and favourites.
Leave **for** blank to set using msg.payload.

info debug dashboard config

all nodes

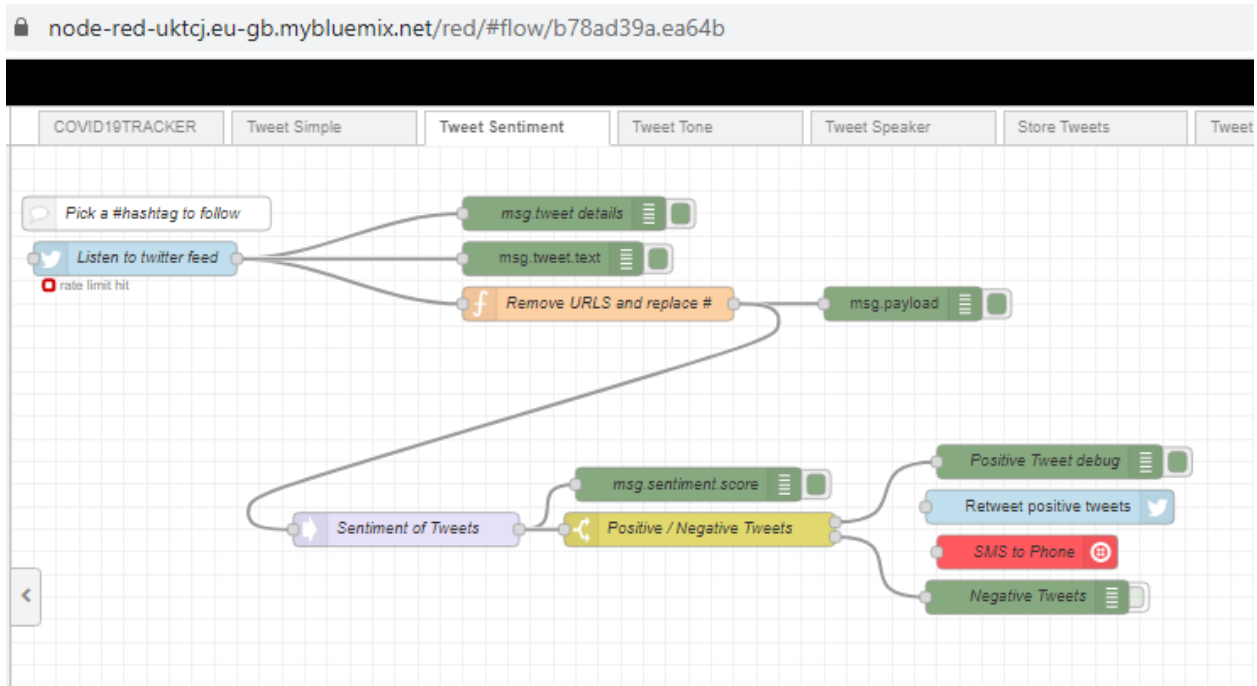
```

tweets/VanessaBenzion1: msg.tweet: Object
  > { created_at: "Tue Jan 30 05:51:14 +0000 2018", id:
    958216010955452400, id_str: "958216010955452416",
    text: "RT @blockweather: IBM Blockcha...", source: "<a
    href='\"http://twitter.com/do...\"' _ }

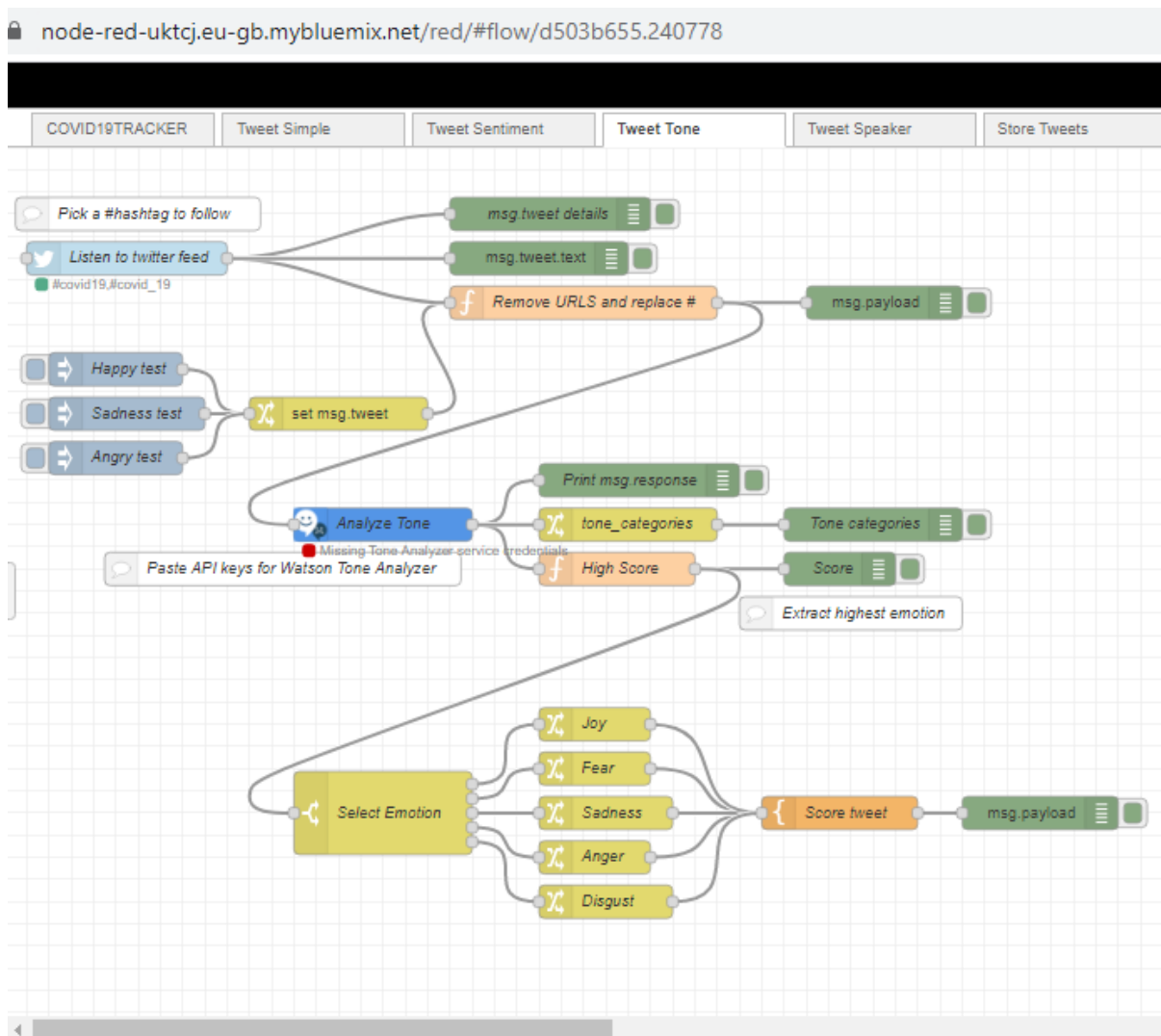
1/30/2018, 12:51:14 AM node: 34a2442.ccef13c
tweets/VanessaBenzion1: msg.payload: string[160]
  > "RT @blockweather: IBM Blockchain Car Lease Demo...
    Hash tag IBM Hash tag blockchain Hash tag car Hash
    tag digitalcurrency Hash tag technology...See short
    URL "

1/30/2018, 12:51:14 AM node: 673ab097.f6d64
tweets/VanessaBenzion1: msg.tweet.text: string[124]
  > string[124]
    RT @blockweather: IBM Blockchain Car Lease Demo
    #IBM #blockchain #car #digitalcurrency #technology
    https://t.co/7IXJdRPtXc
  
```

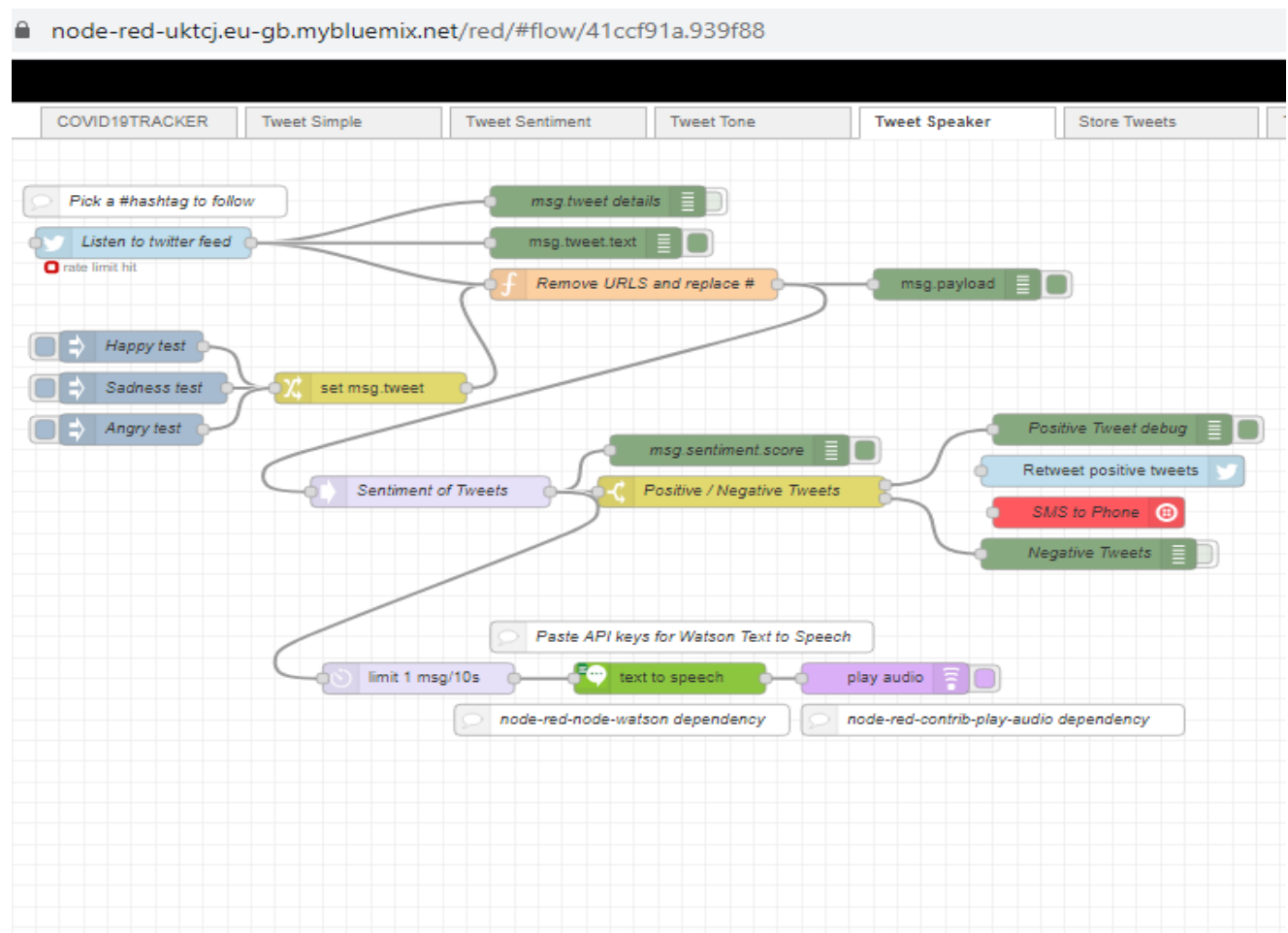
Created a flow for Sentiment Analysis of Tweets



Created a flow for Tone Analysis of Tweets



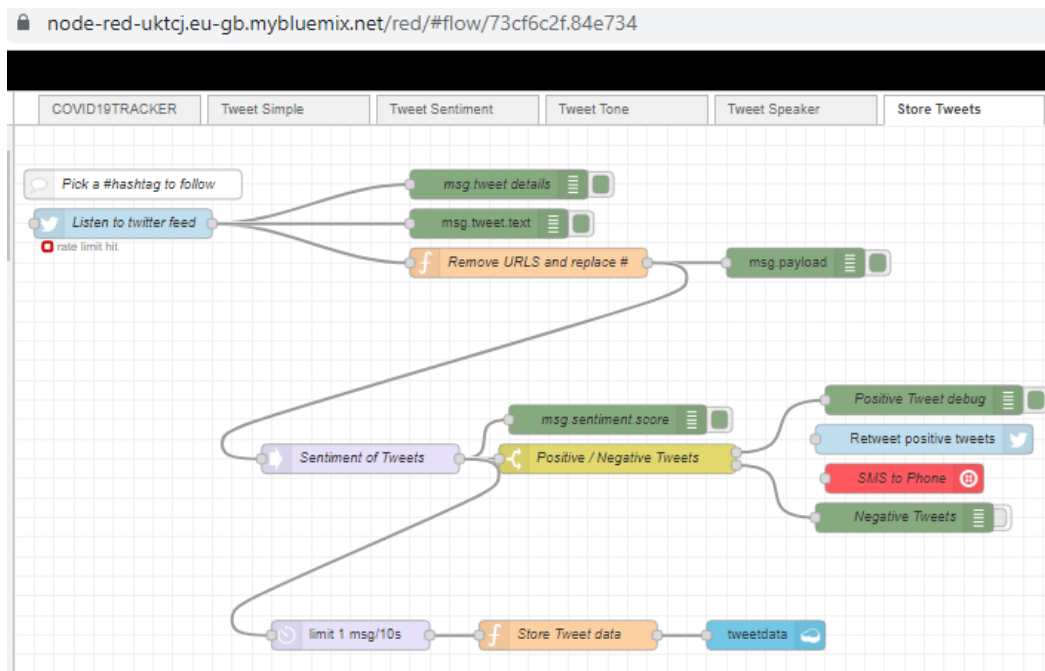
Created a flow for Speak Tweets with Watson Text to Speech



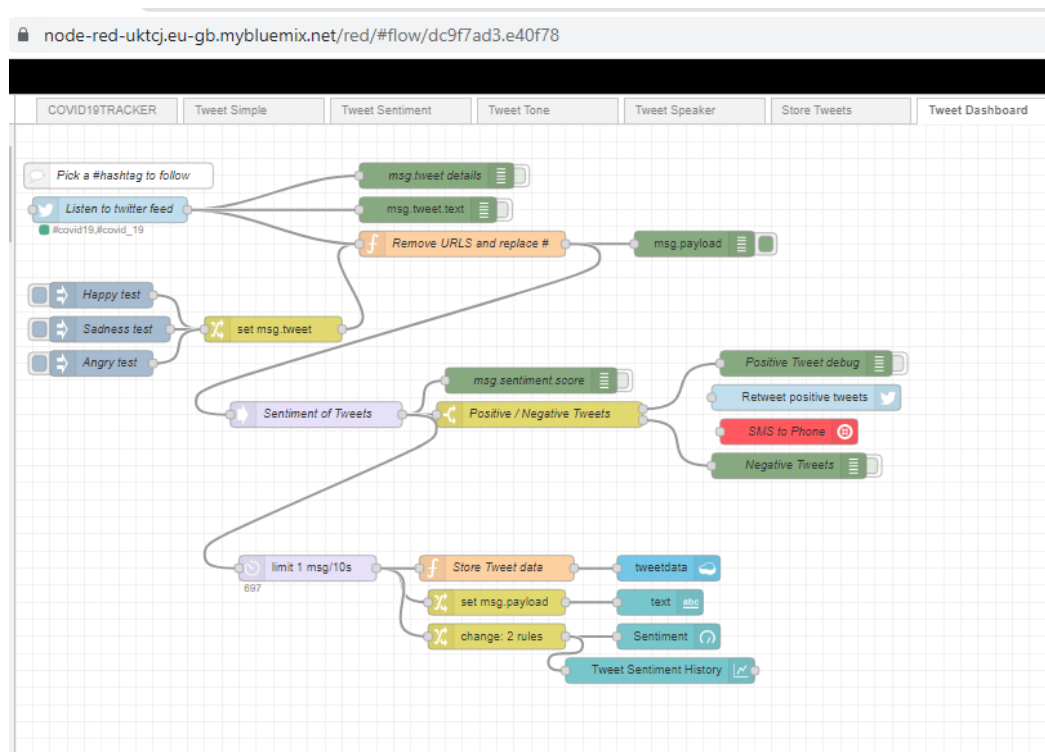
Here,

Add the "**node-red-contrib-play-audio**" package using the Node-RED Manage Palette.

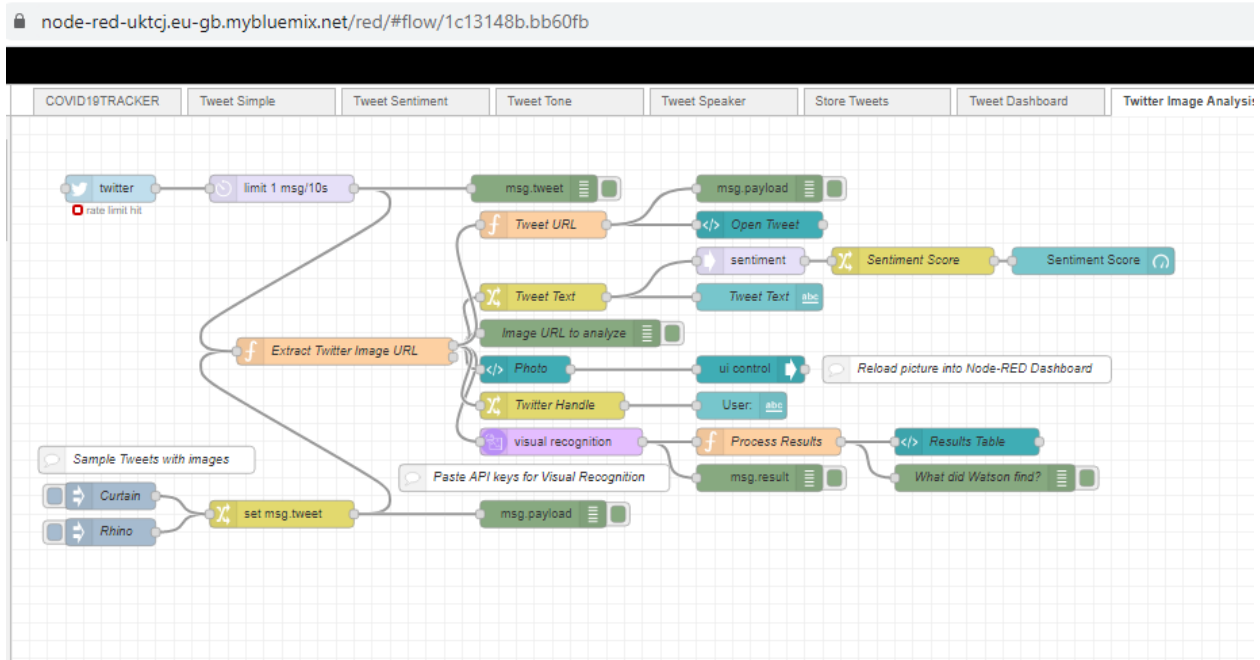
Created a flow for Store Tweets in a Cloudant Database



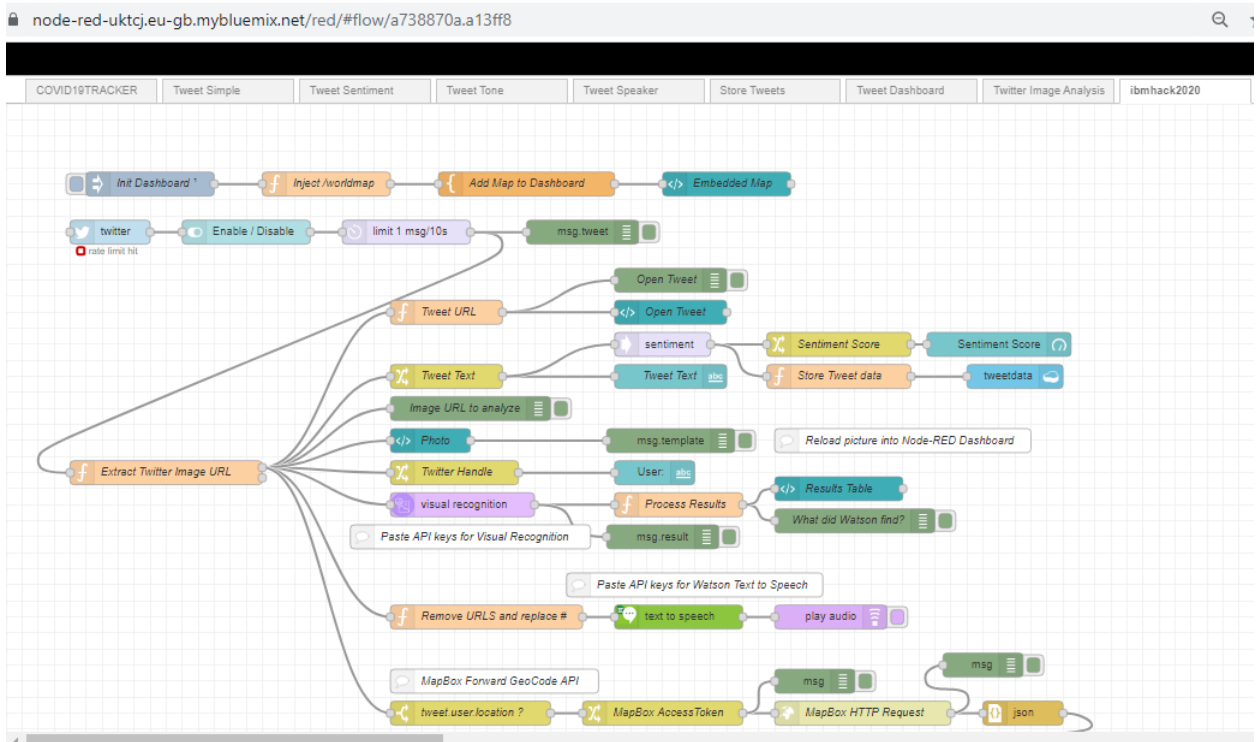
Created a flow for Node-RED Twitter History Dashboard



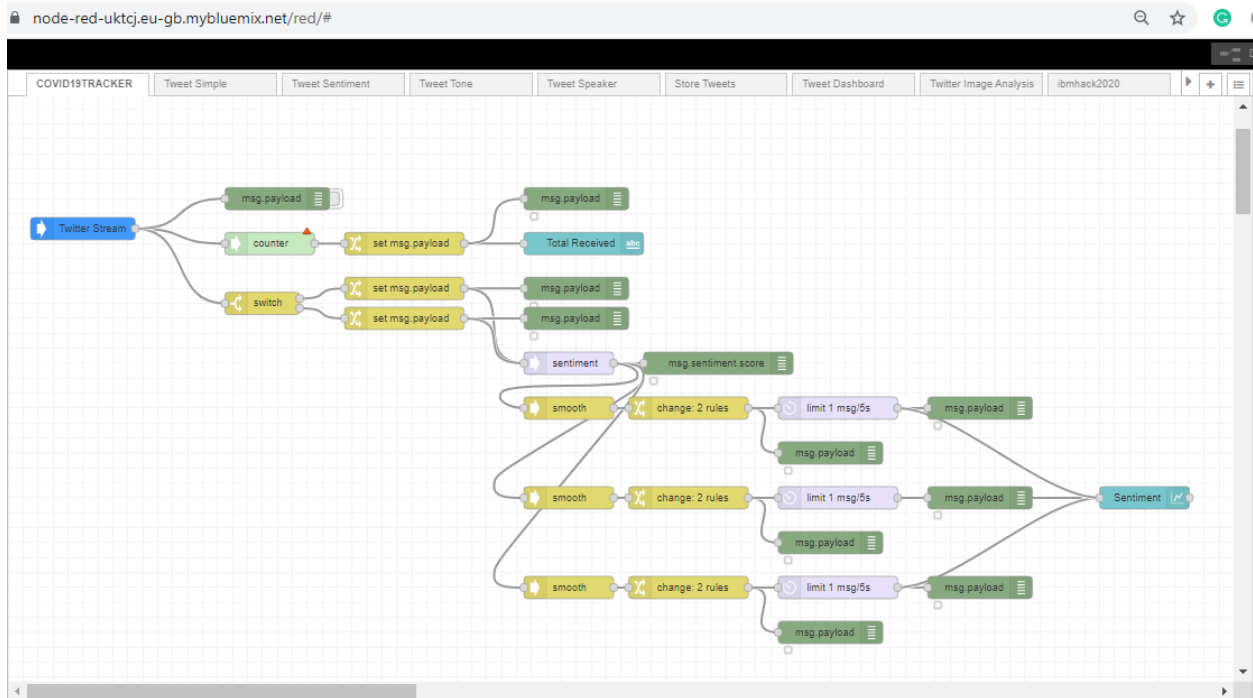
Created a flow for Tweet Image Analysis with Watson Visual



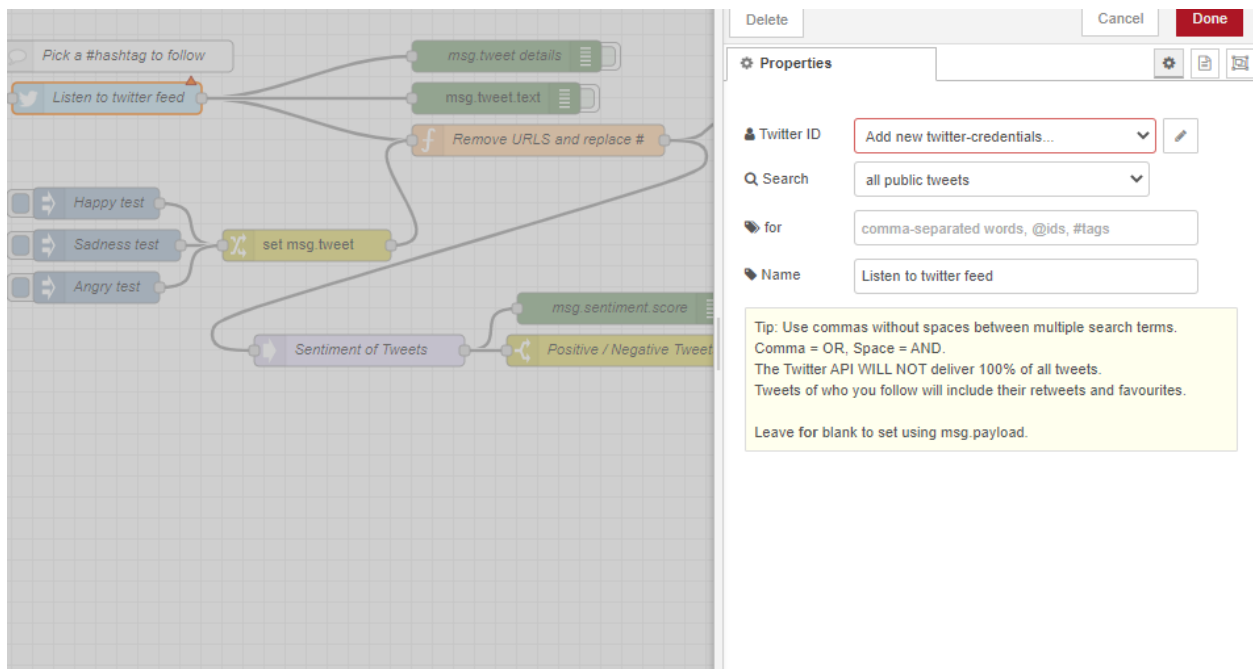
Created a flow for for tracking analysis of survivor requests



Created a flow for Sentiment analysis of covid-19 tweets-visualization Live Tracker for every second



Now, add Twitter-Credentials and Deploy.



Analysis Methods

Sentiment Analysis

TextBlob Polarity & Subjectivity Score:

We utilized TextBlob, a popular NLP library, to conduct sentiment analysis by generating polarity score (negative (-1 ~ +1) positive) and subjectivity score (objective (0 ~ 1) subjective).

IBM Watson Tone Analyzer:

We choose to use IBM's Tone Analyzer (a cloud service) to do the sentiment analysis because it can provide 5 different tones of the text data which is more than positive-negative sentiment analysis. Through this way, we can study the tweets' emotions more specifically.



Google BERT:

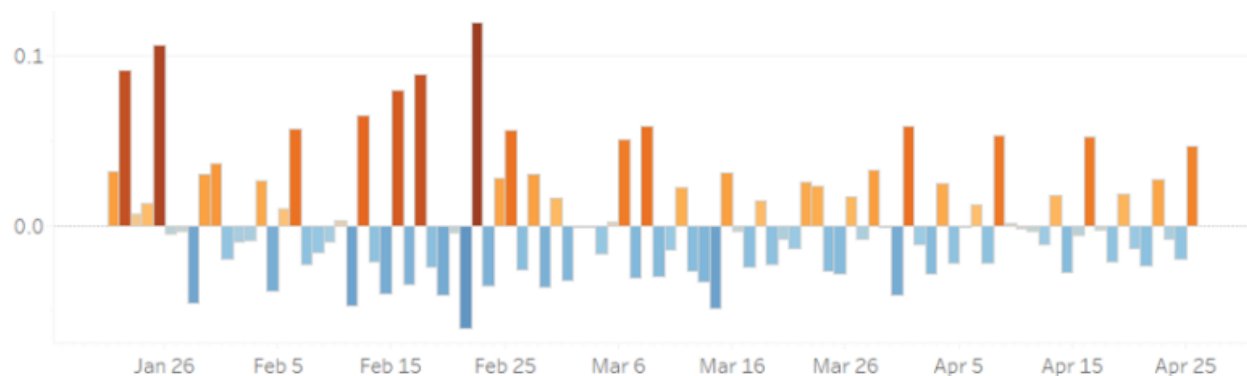
We firstly utilized the Tone Analyzer to labeled a sample of data that we sampling randomly from the whole dataset. With adjustment and also combined with our manually labeled data, we used these data as a training set for the Google BERT model, a state-of-art machine learning technique for classification. Compared to other alternatives, BERT requires much less time and fewer data to train and yields better accuracy. It is a good fit for our case where we have limited training data.

Overview Of Analysis

Overview Of COVID-19 Tweets:

As one of the world's biggest social network platforms, Twitter hosts abundant user-generated posts, which closely reflect the public's reactions towards this pandemic with low latency. By deploying Natural Language Processing (NLP) methods on it, we were able to extract and quantify the public sentiments over time. The tools we used are TextBlob, IBM Watson Tone Analyzer, BERT, and Mallet.

Sentiment Density:

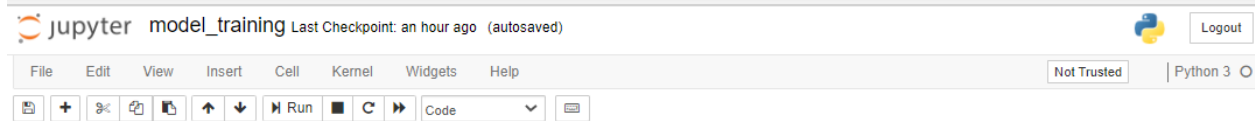


From late February till mid-March, people were undergoing the densest sentiments, especially in terms of the negative feelings, followed by the period of late January to mid-February.

In April, the Sentiment Density decreased and stayed in a lower position, but it was still higher than that of the

beginning.

MODEL BUILDING



Model training through NLP

In [13]: # tokenization; process for creation of bag of words model.

```
corpus = []
for i in range(0, 648958):
    review = re.sub('[^a-zA-Z]', ' ', df['Regenerated_Text'][i])
    review = review.split()

    # The Porter stemming is used for removing the commoner
    # morphological and inflexional endings from words in English.
    ps = PorterStemmer()

    # Stopwords is a process of removing unnecessary english words
    # that are not took part in npl tasks.
    review = [ps.stem(word) for word in review if not word in set(stopwords.words('english'))]
    review = ' '.join(review)
    corpus.append(review)
```

In [14]: len(corpus)

Out[14]: 648958

In [15]: # Creating the Bag of Words model

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=99500)
X = cv.fit_transform(corpus)
y = df.New_Sentiments.values
```

In [16]: print(f'X_shape-->{X.shape}')

X_shape-->(648958, 99500)

Jupyter IBM Tone Analyzer API Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

```
data_api = data_api.set_index(pd.Index(range(0,len(data_api))))
data_test = data_test.set_index(pd.Index(range(0,len(data_test))))

# Save the dataset
data_api.to_csv('data_api.csv',index=False)
data_test.to_csv('data_test_0120_0220.csv',index=False)
```

IBM Tone Analyzer

In [150]: data_api = pd.read_csv('data_api.csv')

In [151]: data_api_12000 = data_api.iloc[9500:12000,]

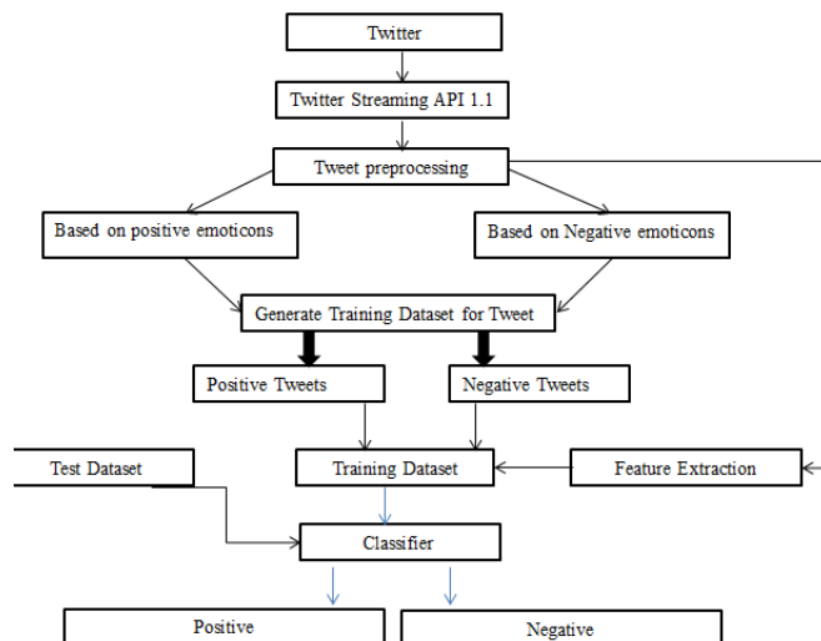
In [153]: data_api_12000.head()

Out[153]:

	has_media	hashtags	img_urls	is_replied	is_reply_to	likes	links	parent_tweet_id	replies	reply_to_users
9500	False	[]	[]	False	False	0	[https://www.wsj.com/articles/china-marshals-...	NaN	0	[]
9501	False	['nCoV', 'coronavirus']	[]	False	False	1	[https://twitter.com/CDPHDirector/status/1224...	NaN	0	[]
9502	False	['Gold', 'China', 'RiskAppetite', 'Commodities...']	[]	False	False	1	[https://s68mv.app.goo.gl/SwgKs]	NaN	0	[]
9503	False	[]	[]	False	False	0	[https://www.bitchute.com/video/ZFKykJUTYIM/]	NaN	0	[]
9504	False	[]	[]	False	True	0	[https://twitter.com/CoronaTurkey/status/1226...	1.226254e+18	0	{['screen_name': 'russian_market', 'user_id': ...]}

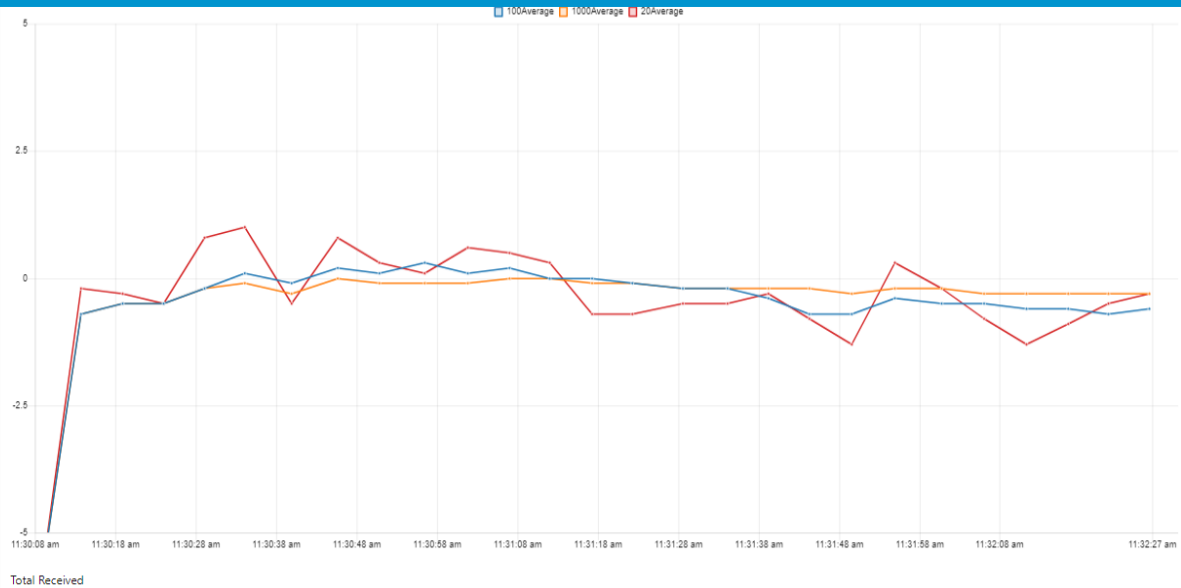
5 rows x 22 columns

FLOWCHART



RESULT

IBM HACK CHALLENGE 2020 [TEAM: SUPERSONICS] Sentiment Analysis Of COVID-19 Tweets-Visualization Dashboard

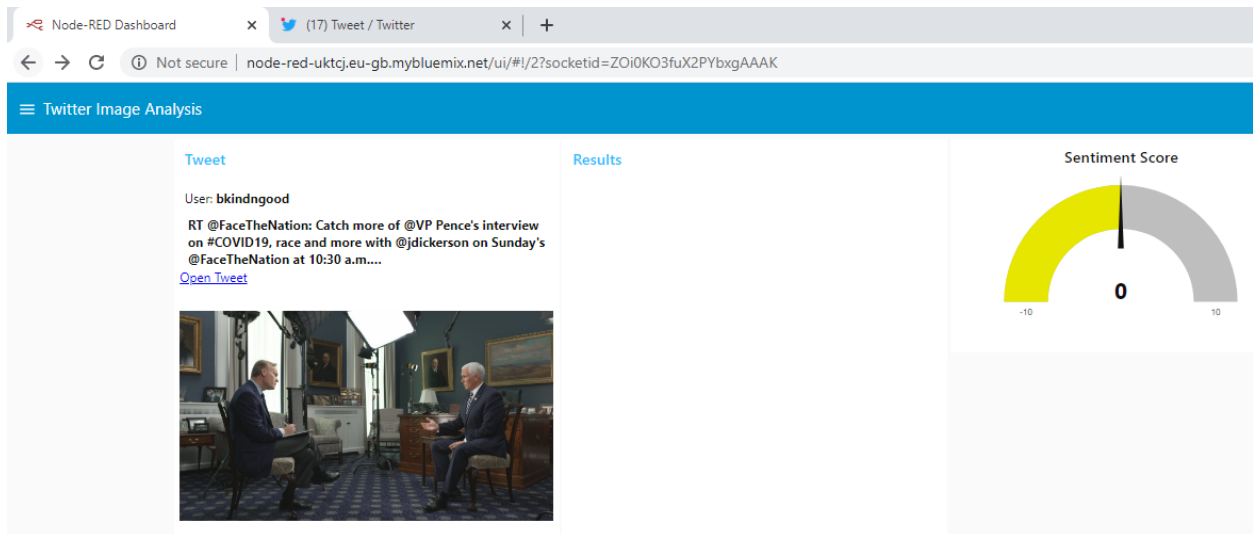


Search

for

Name

Tip: Use commas without spaces between multiple search terms.
Comma = OR, Space = AND.
The Twitter API WILL NOT deliver 100% of all tweets.
Tweets of who you follow will include their retweets and favourites.
Leave for blank to set using msg.payload.



twitter.com/bkindngood/status/1277124569854119936

Home

Explore

Notifications

Messages

Bookmarks

Lists

Profile

More


Tweet

Tweet

Face The Nation @FaceTheNation · 9h

WATCH: @jdickerson presses @VP Pence on why the administration and @RealDonaldTrump has muddled messaging to #maskup during #COVID

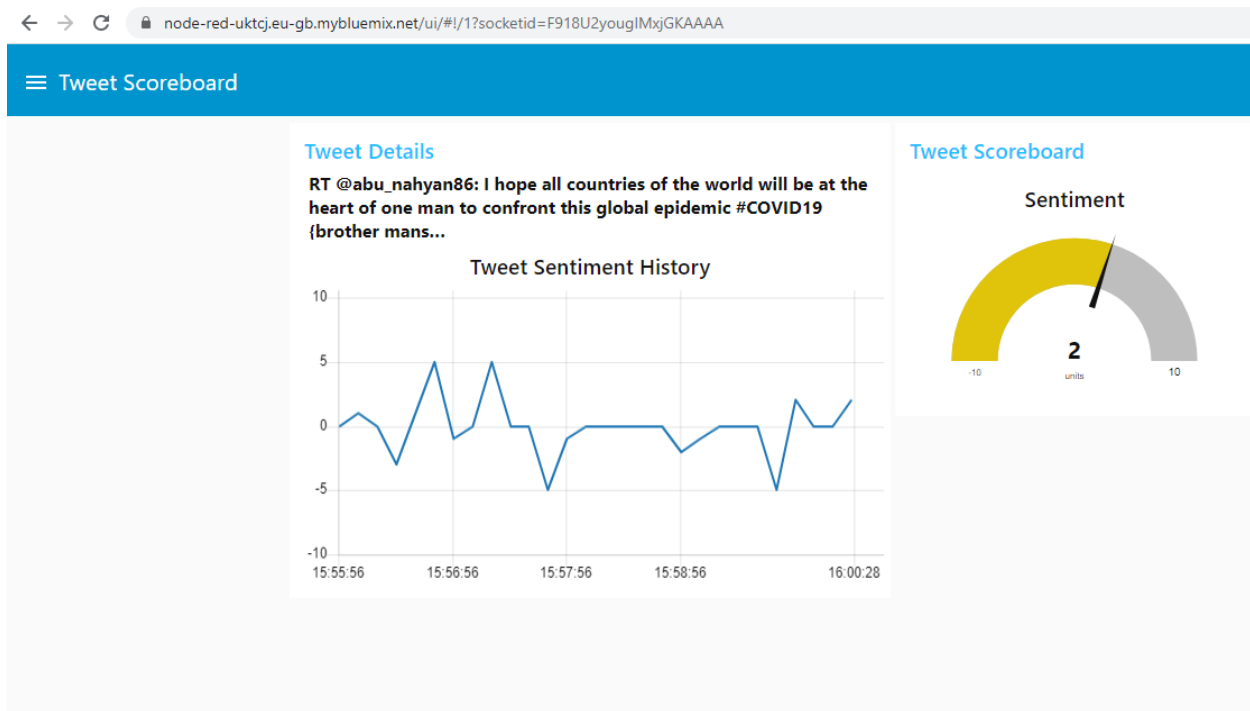
@VP: "One of the elements of the genius of America is the principle of federalism, of state and local control...we want to defer to local officials"



197 168 189

Face The Nation @FaceTheNation · 9h

Catch more of @VP Pence's interview on #COVID19, race and more with @jdickerson on Sunday's @FaceTheNation at 10:30 a.m. EST



ADVANTAGES & DISADVANTAGES

During this COVID19, The use of this information can be applied to make wiser decisions related to the use of resources, to make improvements in organizations, providing better products/services, and ultimately to improve the citizen lifestyle and human relations in order to achieve a better society. Social media like Twitter is the current environment for data collection and analysis of sentiments of people. People can share and comment on everything, from personal thoughts to common events or topics in society.

Despite the possible positive outcomes shown, there are some disadvantages in applying automatic analysis due to the difficulty to

implement it because of the ambiguity of natural language and also the characteristics of the posted content. The analysis of tweets is an example of this, for they are usually coupled with hashtags, emoticons, and links, creating difficulties in determining the expressed sentiment. In addition, there is a need for automatic techniques that require large datasets of annotated posts or lexical databases where emotional words are associated with sentiment values. Another important aspect is that analyses are suitable for the English language, in which there is a limitation for other languages. In the field of sentiment analysis are some challenges in a range of scenarios, in terms of architecture and application domains with unclear or scarce datasets. Also, there is a lack of labeled data, which can pose a barrier to the advancements in this area.

So, we proposed a solution which is only depended on Twitter. And we have also done analysis with datasets.

APPLICATIONS

Our analysis and visualization can be used by Twitter, medical institutions, and business owners.

Twitter: As a social media, Twitter takes the responsibility to control negative rumors spreading during this period

for the social good. Twitter can monitor the sentiment trends and study the abnormal emotion peaks like what we did.

Medical Institutions:

Our analysis can help medical institutions know the emotion changes during the COVID-19. Doctors can provide help to people who potential have mental health problem.

Business Owners: Keeping a watchful eye on trending topics and people's emotion change can help business owners run marketing campaign appropriately and find out potential business opportunities, such as new services that needed by people.

CONCLUSION

We have addressed issues surrounding public sentiment reflecting deep concerns about COVID-19,we analyzed the sentiments of COVID-19-related tweets in several ways.The overall trend shows that the public has been more optimistic over time.Besides,the Sentiment Density indicates that the public turned out to be less loaded

with emotions. At last, the topics behind the sentiments unfolded more details. To fight 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.

Corporations and small businesses can also benefit through such analyses and machine learning models to better understand consumer sentiment and expectations.

FUTURE SCOPE

Web-Application should be converted to Mobile application. COVID19 Sentimental Analysis may be implemented in future for accuracy purposes. There is a plenty of information on social media which can be analysed, and this information is very sparse. Predictive models from the hybrid of different machine learning algorithms can be used to correctly access the sentiment analysis, which is in fact one of the most difficult problem statement in machine learning world. The accuracy of models for non-English statements

should be improved in future.

APPENDIX

Source code

Training

```
1 from sklearn.feature_extraction.text import TfidfVectorizer
2 from sklearn.naive_bayes import GaussianNB
3 from sklearn.metrics import classification_report
4 from sklearn.metrics import accuracy_score
5 import scipy.sparse as sp
6 import numpy as np
7 import re
8 def read():
9     with open("training.txt") as f:
10         contents = f.readlines()
11         ytrain = []
12         lines = []
13         for content in contents:
14             frags = re.split('\t', content.strip())
15             ytrain.append(frags[0])
16             lines.append(frags[1])
17         return (lines, np.array(ytrain))
18 def split_test_train(x, y):
19     posIndex = y == "1"
20     negIndex = y == "0"
21     posX = x[posIndex]
22     negX = x[negIndex]
23     posY = y[posIndex]
24     negY = y[negIndex]
25     xtrain =
26     sp.vstack((posX[:int(posX.shape[0]*0.8)], negX[:int(negX.shape[0]*0.8)]
27 ), format='csr')
28     ytrain =
29     np.concatenate((posY[:int(posX.shape[0]*0.8)], negY[:int(negX.shape[0]*
30 0.8)]))
31     xtest =
32     sp.vstack((posX[int(posX.shape[0]*0.8):], negX[int(negX.shape[0]*0.8):]
33 ), format='csr')
34     ytest =
35     np.concatenate((posY[int(posX.shape[0]*0.8):], negY[int(negX.shape[0]*0
36 .8)]))
37     return (xtrain, ytrain, xtest, ytest)
```

```

38 def train_test():
39     (lines,y) = read()
40     vect = TfidfVectorizer()
41     vect.fit(lines)
42     x = vect.transform(lines)
43     (xtrain,ytrain,xtest,ytest) = split_test_train(x,y)
44     clf = GaussianNB()
45     clf.fit(xtrain.toarray(),ytrain)
46     ypred = clf.predict(xtest.toarray())
47     t = ["positive","negative"]
48     print "accuracy:"
49     print(accuracy_score(ytest, ypred))
50     print(classification_report(ytest, ypred, target_names=t))
51 def train():
52     (lines,y) = read()
53     vect = TfidfVectorizer()
54     vect.fit(lines)
55     x = vect.transform(lines)
56     clf = GaussianNB()
57     clf.fit(x.toarray(),y)
58     return (vect,clf)

```

Sentiment Analysis

```

1  from TwitterAPI import TwitterAPI
2  import train
3  import warnings
4  access_token_key = "#####"
5  access_token_secret = "#####"
6  consumer_key = "#####"
7  consumer_secret = "#####"
8  def get_score(query):
9      (vect,clf) = train.train()
10     api = TwitterAPI(consumer_key, consumer_secret, access_token_key,
11     access_token_secret)
12     r = api.request('search/tweets', {'q':query})
13     tweets = []
14     for item in r:
15         tweets.append(item['text'] if 'text' in item else item)
16     x = vect.transform(tweets)
17     ypred = clf.predict(x.toarray())
18     print "total no. of tweets : " + str(len(ypred))
19     print "no. of positive tweets : "+str(sum(ypred=="1"))
20     print "no. of negative tweets : "+str(sum(ypred=="0"))

```

NodeRedApp.URL:

<https://node-red-uktcj.eu-gb.mybluemix.net/ui/>

Google_Drive_Link for Python codes:

https://drive.google.com/drive/folders/1mybil7B2_6ryzAJEPL22YPjudmStWU7C?usp=sharing

BIBILOGRAPHY

- <https://ieee-dataport.org/keywords/covid-19-twitter-sentiment>
- <https://www.tweetbinder.com/blog/covid-19-coronavirus-twitter/>
- <https://github.com/taspinar/twitterscraper>
- <https://towardsdatascience.com/twitter-sentiment-analysis-based-on-news-topics-during-covid-19-c3d738005b55>
- <https://www.degruyter.com/view/journals/jisys/28/3/article-p377.xml>
- <https://github.com/johnwalicki/Node-RED-Twitter-Workshop>
- <https://github.com/Lewuathe/COVID19-SIR>
- <https://arxiv.org/ftp/arxiv/papers/1601/1601.06971.pdf>
- <https://trendogate.com/>

~~~THANK YOU~~~