PROJECT DETAILS

Challenge Title: IBM Hack Challenge 2020
Project ID: SPS_PRO_852
Project Title: TWITTER SENTIMENTAL ANALYSIS ON COVID-19
*Application ID: SPS_CH_APL_20200002196*
Team: FHH
Team Member: Anagani Gnanasri,Devaraju Kiranmayi

TWITTER SENTIMENTAL ANALYSIS ON COVID-19

WHO Director-General's opening remarks at the media briefing on COVID-19 "It took 12 weeks for the world to reach 400 thousand cases of COVID-19.Over the weekend, there were more than 400 thousand cases across the globe.There have now been 11.4 million cases of COVID-19 and more than 535,000 lives have been lost. The outbreak is accelerating and we have clearly not reached the peak of the pandemic. While the number of deaths appears to have levelled off globally, in reality some countries have made significant progress in reducing the number of deaths, while in other countries deaths are still on the rise. Where there has been progress in reducing deaths, countries have implemented targeted actions toward the most vulnerable groups, for example those people living in long-term care facilities."
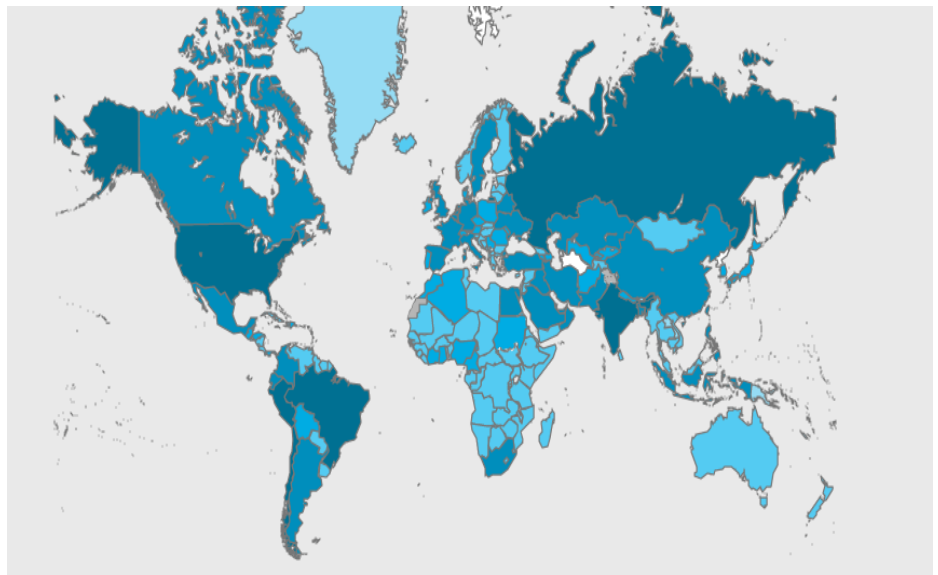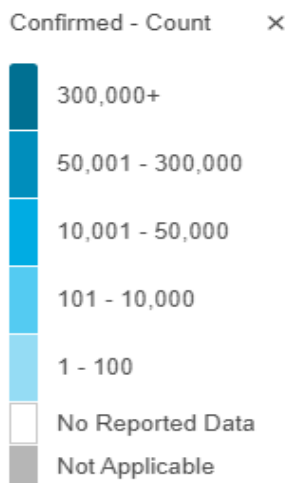
Globally, as of **1:35pm CEST, 12 July 2020**, there have been **12,486,484 confirmed cases** of COVID-19, including **560,814 deaths,** reported to WHO

Table of Contents

1. Introduction

1.1 Overview

What is sentiment analysis ?

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral.It's also known as **opinion mining**, deriving the opinion or attitude of a speaker.

Million of people give their opinion on different topics on a daily basis on social media like Facebook, Twitter, Instagram,etc. It has many applications in different areas of research from social science to business. Twitter nowadays is one of the popular social media which according to the Statists in currently has over 300 million accounts. Twitter is a rich source to learn about people's opinions and sentimental analysis. Each tweet is important to determine the sentiment of the tweet whether is it positive, negative, or neutral. Another challenge with twitter is Recently it has extended the text limitations to 280 characters of each tweet which causes people to use phrases and words which are not in language processing. It contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text . Some of the top abbreviations are FB for Facebook, B4 for before and so on. Therefore sentimental analysis for short texts like Twitter's posts is challenging.

1.2 Purpose

The Corona Virus endangers our physical health indeed, but alongside, social distancing also poses a threat to our emotional stability. Thus, it is crucial to understand public sentiments under COVID-19. We deployed Sentiment Analysis on tweets and Topic Modelling on news to aid the understanding of sentiment trends. Based on these, we

worked on them to present the results.

2.Literature Survey

Abstract: Social networks are the main resources to gather information about people's opinions and sentiments towards different topics and issues. People spend hours daily on social media to share their ideas, opinions, and reactions with others, so in this paper, we analyse the sentiments regarding coronavirus disease(COVID-19) because many peoples of different countries are affected by coronavirus that is very critical issue in these days, so analyse the sentiments of different people's opinion for this disease, we are fetching the twitter streaming tweets related to coronavirus using twitter API and analyse these tweets using machine learning techniques and tools as positive, negative and neutral. In this paper, we run experiments through Python programming on different tweets using twitter API and NLTK library is used for pre-processing of tweets and then analyse the tweets dataset by using Text blob and after that show the interesting results in positive, negative, neutral sentiments through different visualizations.

Keywords: Twitter Sentiment Analysis, Twitter API, TextBlob, NLTK, Tweepy, COVID-19,coronavirus.

Our analysis has shown some relationships between confirmed cases' growth and the trends of sentiments. 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.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.

3.Theoritical Analysis

3.1 Block diagram
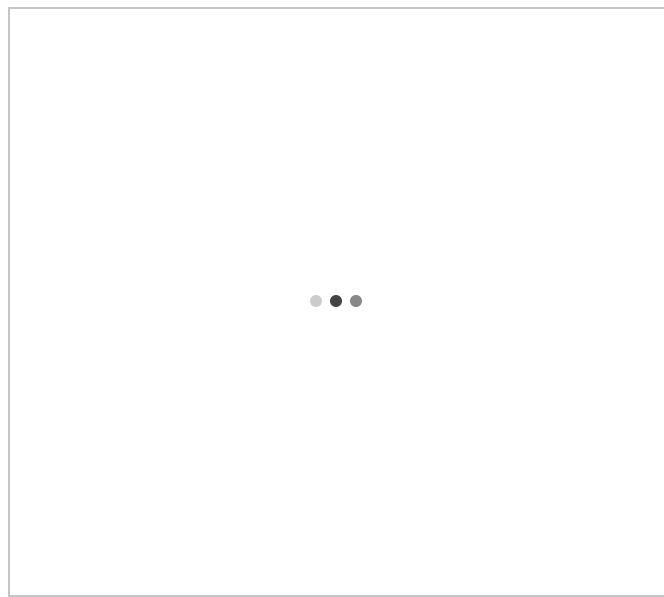
Data Collection Process:
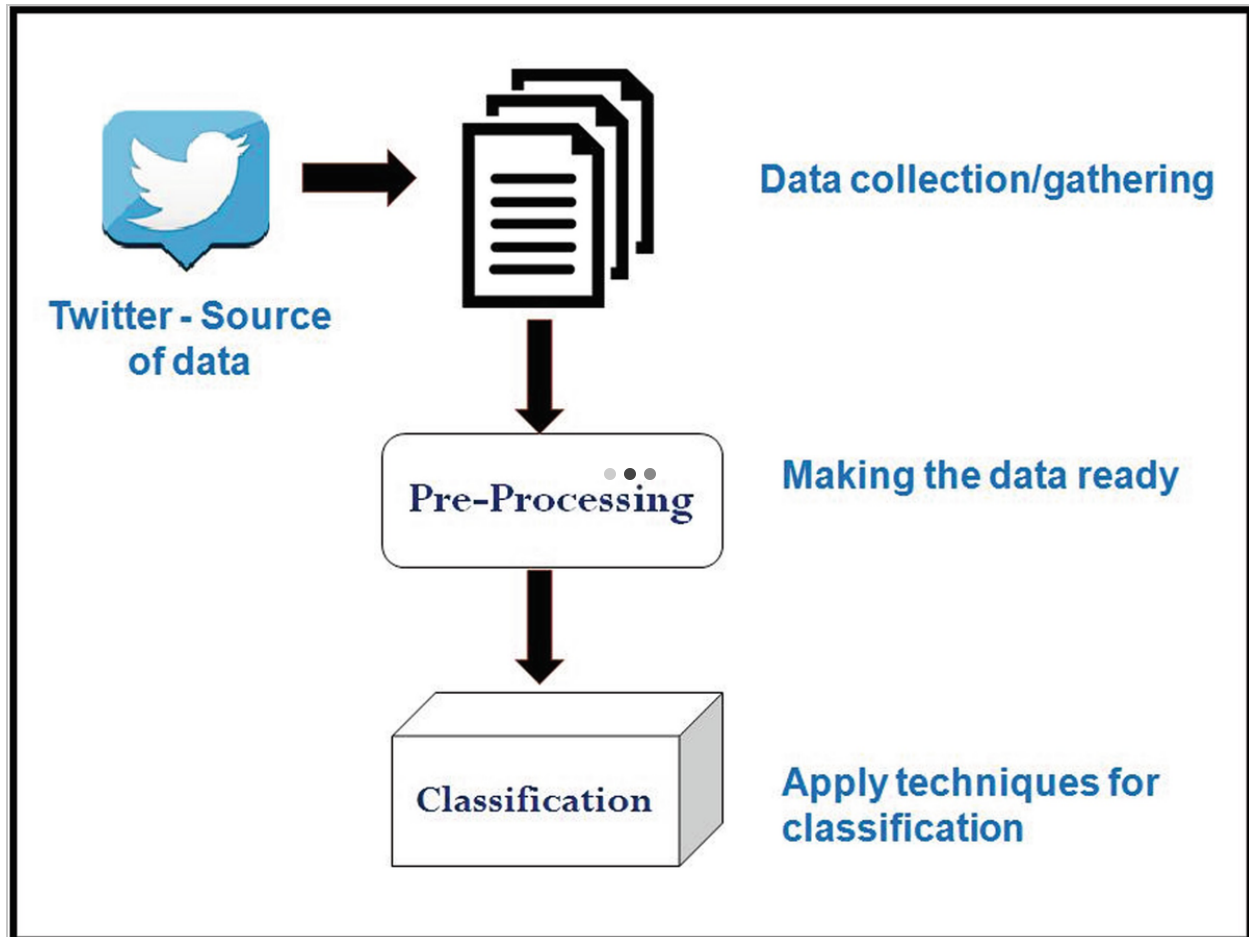
3.2 Design and Implementation

Installation:

Tweepy: tweepy is the python client for the official Twitter API.
Install it using following pip command:

```
1 pip install tweepy
```

TextBlob: textblob is the python library for processing textual data.
Install it using following pip command:

```
1 pip install textblob
```

Also, we need to install some NLTK corpora using following command:

```
1 python -m textblob.download_corpora
```

(Corpora is nothing but a large and structured set of texts.)

WordCloud: WordCloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

```
1 pip install wordcloud
```

Collections: Collections in Python are containers that are used to store **collections** of data, for example, list, dict, set, tuple etc. ... Several modules have been developed that provide additional data structures to store **collections** of data.

```
1 pip install collections
```

re: A regular expression is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern.

```
1 pip install re2
```

Natural Language Toolkit: NLTK is a leading platform for building Python programs to work with human language data

```
1 pip install nltk
```

Authentication:

In order to fetch tweets through Twitter API, one needs to register an App through their twitter account. Follow these steps for the same:

- Open this link and click the button: 'Create New App'
- Fill the application details. You can leave the callback url field empty.
- Once the app is created, you will be redirected to the app page
- Open the 'Keys and Access Tokens' tab.
- Copy 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret'.

Accessing Twitter Data:

After the authentication, we need to connect with Twitter

Streaming API. **Tweepy** a python library enables us to connect with Twitter and download data. Once the Twitter Authentication service authenticates the API, a token is generated and made available to the API for each Twitter server transaction. Using this token, tweets are mined using hashtags or keywords. To access the data, we use the following code, and this downloaded data is stored in .csv files as or dataset.
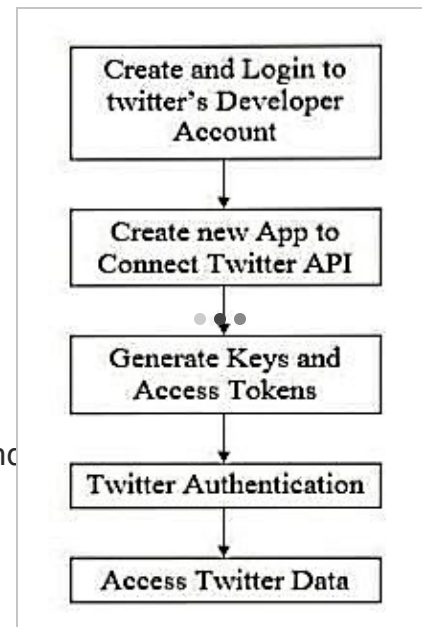
```
1  #Authorization and Search tweets
2  #Getting authorization
3  consumer_key = 'XXXXXXXXXXXXXXXXXXXXXXXX'
4  consumer_key_secret = 'XXXXXXXXXXXXXXXXXXXXXXXXXXXX'
5  access_token = 'XXXXXXXXXXXXXXXXXXXXXXXXXXX'
6  access_token_secret = 'XXXXXXXXXXXXXXXXXXXXXXX'
7  auth=tweepy.OAuthHandler(consumer_key, consumer_key_secret)
8  auth.set_access_token(access_token, access_token_secret)
```

```
9 api = tweepy.API(auth, wait_on_rate_limit=True)
```

Data Preprocessing:

The data which is extracted from twitter is unstructured and unorganized, expressed in various aspects by using various vocabularies, slangs, the context of writing, etc. Therefore, data preprocessing consists of data cleaning and stop word removal.

Data Cleaning:

It includes Removal of unnecessary data such as HTML Tags, white Spaces and Special characters takes place from the Twitter dataset. This noise does not make any sense, therefore, they need to be removed. Data cleaning is achieved by importing **regular expression (RE)** python library. Process of cleaning data for our system is as follows:

- The First step is to remove the URL. URL is not considered as an essential element in the tweets and just for simplicity, URLs are removed.
- Twitter handlers such as '@abs' are also removed as these are not provide any weights in sentiment classification.
- The next step is to remove punctuation. After that, the removal of special character takes place.
- Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated.
- Extra White spaces are also replaced with single white space.
- White spaces from the beginning and end are also removed.

```
1 #Cleaning Data
2 #Removing @ handle
3 def remove_pattern(input_txt, pattern):
4     r = re.findall(pattern, input_txt)
5     for i in r:
6         input_txt = re.sub(i, '', input_txt)
7     return input_txt
8 tweet_dataset['text']=np.vectorize(remove_pattern)(tweet_data
  set['text'], "@[\w]*")
9 tweet_dataset.head()
```

Stop Words :

Removal Stop words are the dictionary-based bag of words. These are the set of commonly used words not only in English but in any other language. Stop words focus on important word only instead of commonly used words in a language. Stop word Removal is done by eliminating the unnecessary words from the Twitter data set, So that, the resultant data set contains only the required information for the analysis.After stop word removal, only important words that could lead to sentiment detection are left. Stop word Removal is achieved by another python library known as **NLTK**.

```
1  #Cleaning Tweets
2  corpus = []
3  for i in range(0, 2000):
4      tweet       =       re.sub('[^a-zA-Z0-9]',      '      ',
   tweet_dataset['text'][i])
5      tweet = tweet.lower()
6      tweet = re.sub('rt', '', tweet)
7      tweet = re.sub('http', '', tweet)
8      tweet = re.sub('https', '', tweet)
9      tweet = tweet.split()
10     ps = PorterStemmer()
11      tweet = [ps.stem(word) for word in tweet if not word in
   set(stopwords.words('english'))]
12     tweet = ' '.join(tweet)
13     corpus.append(tweet)
```

Feature Extraction :

Feature extraction is an important step in opinion mining that generates a list of object, aspect, features, and opinions. The purpose of feature extraction is to extract opinion sentences which contain one or more features, aspects, and opinions. In most of the cases, aspect words are nouns and noun phrases, their opinion words are adjectives and adverbs. In this research work, features are extracted using the **TextBlob** library. After the preprocessing phase, only necessary words are left in tweets which are used for analysis. We extract only nouns and noun phrases from the tweets. These noun phrases are used to know the 'who' in the tweet. After the extraction of nouns and noun phrases, only words that have features or aspects such as adjective and adverb and so on are left. Therefore, in the next phase, these extracted features are classified into

sentiments.

Sentiment Identification :

After feature extraction, we identify the **positive** and **negative** orientation of words. These features are searched into opinion word list from the huge set of corpora in **TextBlob** Library to find out the sentiments. Features are searched into positive and negative word list of the dictionary. If the word is present in the positive opinion word list, then the positive Sentiment is assigned to the corresponding feature. If the word is present in the negative Word list, then the negative sentiment is assigned to the corresponding feature. If the word is not present in both the word list, then the sentiment is considered as **neutral**.

Output Presentation:

The final phase of our sentiment analysis is the visualization of the results. We use Pie charts, Trend Graphs and tables to view the results. We use a python library Matplotlib to plot Pie charts and a trend graph to plot these charts and charts. Its numerical extension to mathematics is NumPy. And we're using Pandas for table visualization. Pandas is a software library for manipulation and analysis of data written for the Python programming language. It provides data structures and operations to manipulate numerical tables and time sequences.
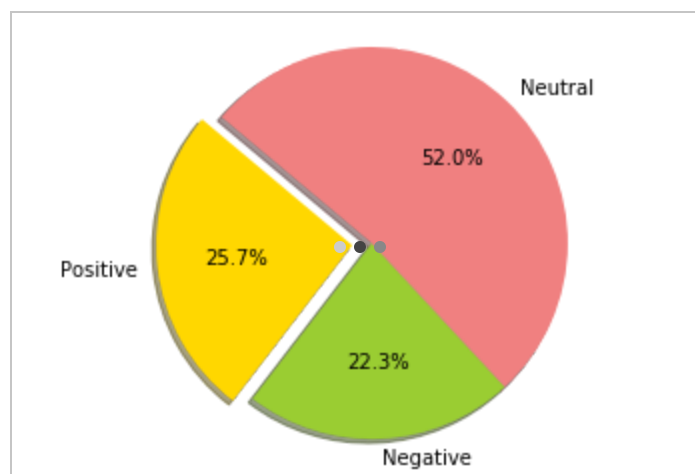

chart:results of twittersentimental analysis on covid-19

This figure shows the result of tweets for 'coronavirus(Covid 2019)' based on 2000

tweets from Twitter.

Positive tweets percentage 25.7%
Negative tweets percentage 22.3%
Neutral tweets percentage 52.0%

Experiment And Results:

Data is collected from twitter for analyzing tweets on any popular topic using search keyword. In our experimental work, we store tweets in one dataset. We choose the trending keyword **covid**.According to pie chart for covid, we observed that **25.7**% tweets are positive, **22.3%** tweets are negative, and **52.0%** are Neutral. These results show that the majority of People are neutral sentiment towards **COVID-19**

Advantages:

The analysis of texts to determine the writers' or speakers' opinion and attitude expressed, and how the results can be used.

sentiment: a point of view, opinion, feeling, or attitude held or expressed by a person

Sentiment analysis provides some answers into what the most important issues are, from the perspective of people, 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.

Word Cloud of Tweets from India
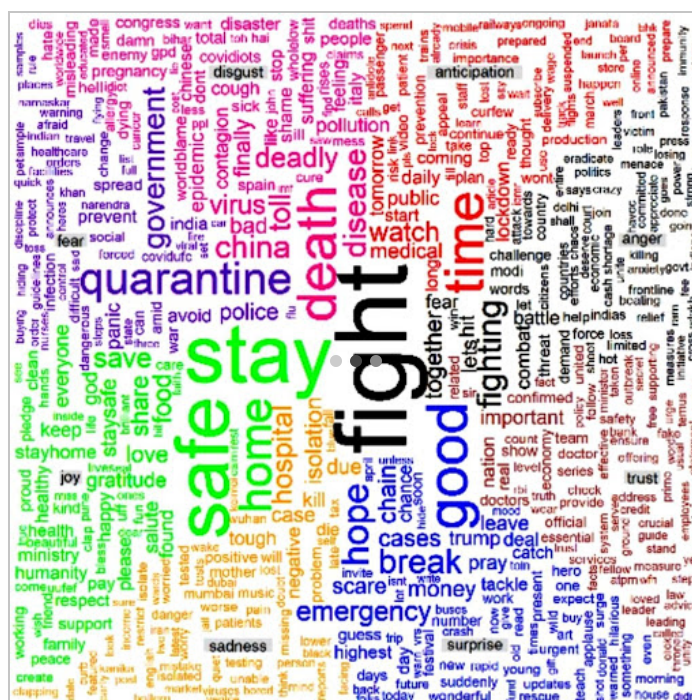
As we can see there was a hope,positivity and unity among the people to fight covid-19.

Conclusion :
In this research, the opinion of different peoples of different countries on coronavirus the affect various countries has been discussed. The main focus of this paper is on Twitter ,Twitter API and have implemented the python programming language and write code in Jupternotebook to implement the sentimental analysis as positive,negative and neutral. The results are shown by using Matplotlib library. It has been realized that the neutral sentiments are significantly high which shows there is a need to improve Twitter sentiment analysis.

Further Scope:
Our analysis has shown some relationships between confirmed cases' growth and the trends of sentiments. 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.
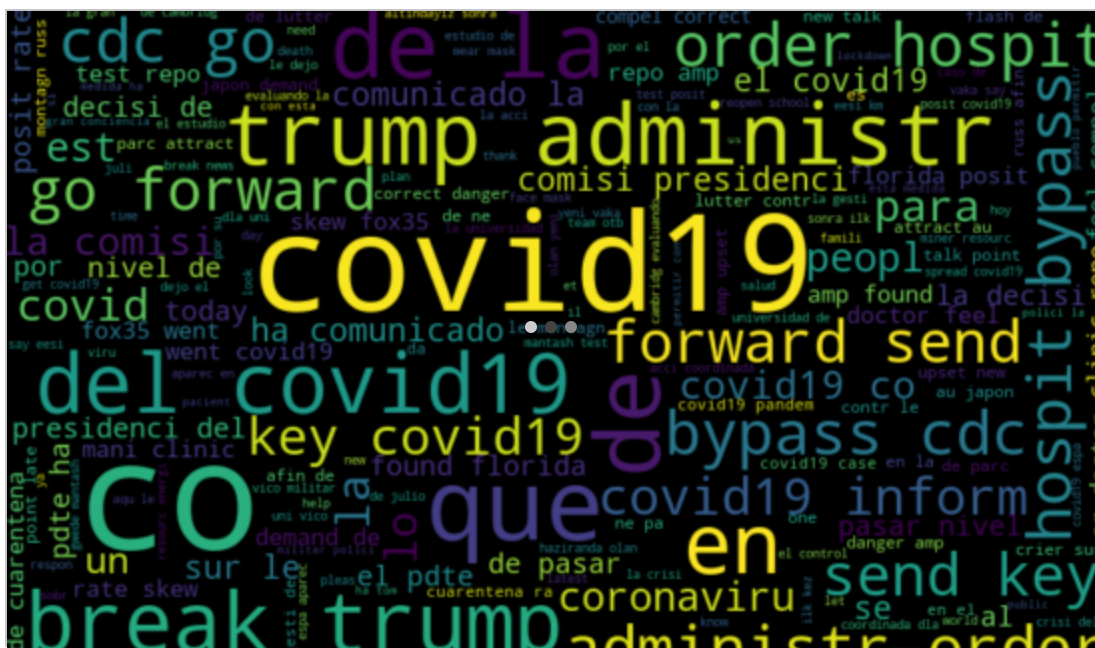
This research work aimed at analyzing the sentiments and emotions of the people during the pandemic COVID19. During the study, it was revealed that countries like Belgium, India and Australia were tweeting about COVID19 with a positive sentiment, people in China had negative sentiments about the same. Similarly, while analyzing the

word clouds of different countries, it was concluded that people are tweeting words like Pandemic, Death, Quarantine, Hope, Stay Safe, Government, Political, Fight and Masks with different emotions. The name of the USA President, Donald Trump was amongst one of the most tweeted words not only in USA, but across all the twelve countries considered for the study.

So here by I conclude it.Thank you