Capstone Project 3

Coronavirus Tweet Sentiment Analysis

Team member

Suraj Kumar

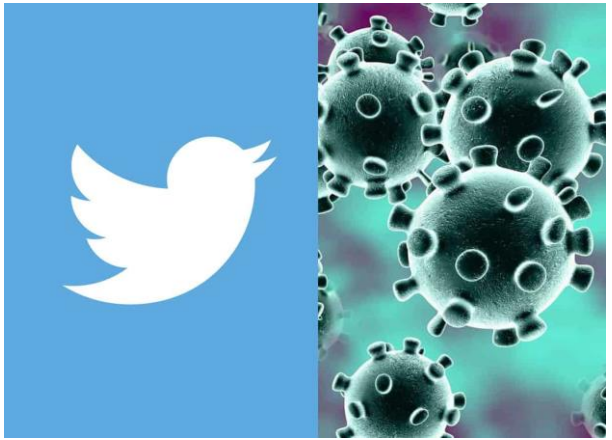
Shreya Ranjan

Sentiment Analysis : Predicting sentiment of COVID-19 Tweets

COVID-19 originally known as Coronavirus Disease of 2019, has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020.

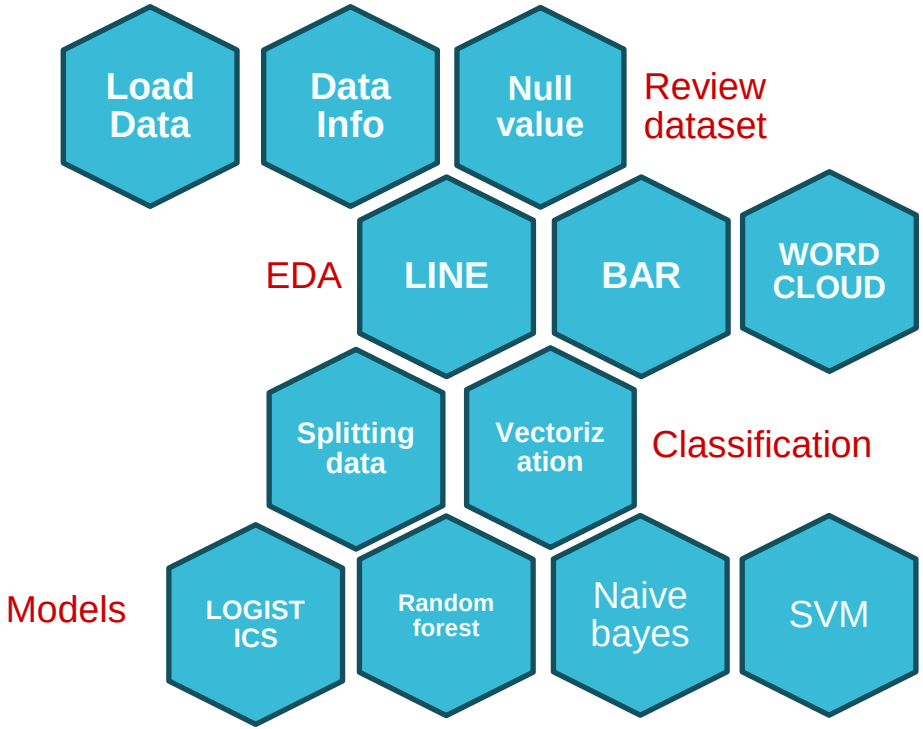
Sentiment Analysis-DataSet

We must create a classification model to forecast the sentiment of COVID-19 tweets for this project. The tweets were downloaded from Twitter after which human tagging was completed.
The names and usernames have been given codes to avoid any privacy concerns.
You are given the following information:
1.Location
2.Tweet At
3.Original Tweet
4.Label



By simplifying the relevant facts and information, the study, which examines various sorts of tweets received during the pandemic, can be helpful in developing policies to protect the countries.

Data Pipelines



Libraries:

1} NumPy
2} Panda
3} Matplotlib
4} Seaborn
5} Datetime
6} Sklearn
7} NLP

Data Wrangling:

Columns and rows in dataset:

```
Tweet_df
Rows 41157 Columns 6
```

Data Information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   UserName       41157 non-null  int64 
 1   ScreenName     41157 non-null  int64 
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.9+ MB
```

Data Wrangling(cont.)

Null values:

```
UserName            0
ScreenName          0
Location         8590
TweetAt             0
OriginalTweet       0
Sentiment           0
dtype: int64
```

Last 5 Tail count:

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
41152	44951	89903	Wellington City, New Zealand	14-04-2020	Airline pilots offering to stock supermarket s...	Neutral
41153	44952	89904	NaN	14-04-2020	Response to complaint not provided citing COVI...	Extremely Negative
41154	44953	89905	NaN	14-04-2020	You know itÂ□s getting tough when @KameronWild...	Positive
41155	44954	89906	NaN	14-04-2020	Is it wrong that the smell of hand sanitizer i...	Neutral
41156	44955	89907	i love you so much || he/him	14-04-2020	@TartiiCat Well new/used Rift S are going for ...	Negative

Data Wrangling(cont.)

First 10 head count

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative
5	3804	48756	ÃT: 36.319708,-82.363649	16-03-2020	As news of the regionÂs first confirmed COVID...	Positive
6	3805	48757	35.926541,-78.753267	16-03-2020	Cashier at grocery store was sharing his insig...	Positive
7	3806	48758	Austria	16-03-2020	Was at the supermarket today. Didn't buy toile...	Neutral
8	3807	48759	Atlanta, GA USA	16-03-2020	Due to COVID-19 our retail store and classroom...	Positive
9	3808	48760	BHAVNAGAR,GUJRAT	16-03-2020	For corona prevention,we should stop to buy th...	Negative

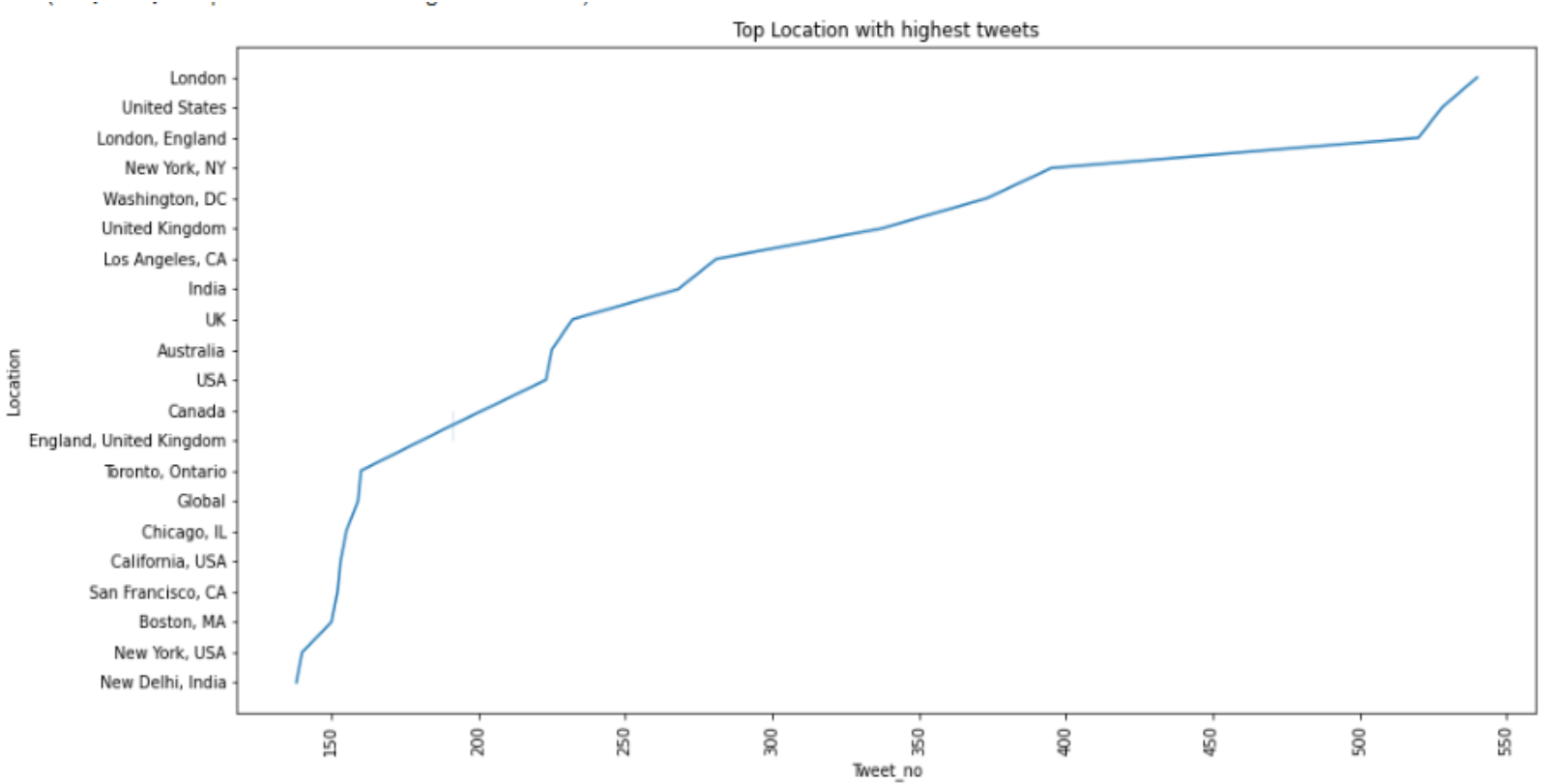
EDA:

Location of tweet from Top 20 location

This table clearly shows that London is the location from which maximum number of tweet are posted and
New Delhi, India is a location from the top 20 location being at the lowest position.

	Location	Tweet_no
0	London	540
1	United States	528
2	London, England	520
3	New York, NY	395
4	Washington, DC	373
5	United Kingdom	337
6	Los Angeles, CA	281
7	India	268
8	UK	232
9	Australia	225
10	USA	223
11	Canada	191
12	England, United Kingdom	191
13	Toronto, Ontario	160
14	Global	159
15	Chicago, IL	155
16	California, USA	153
17	San Francisco, CA	152
18	Boston, MA	150
19	New York, USA	140
20	New Delhi, India	138

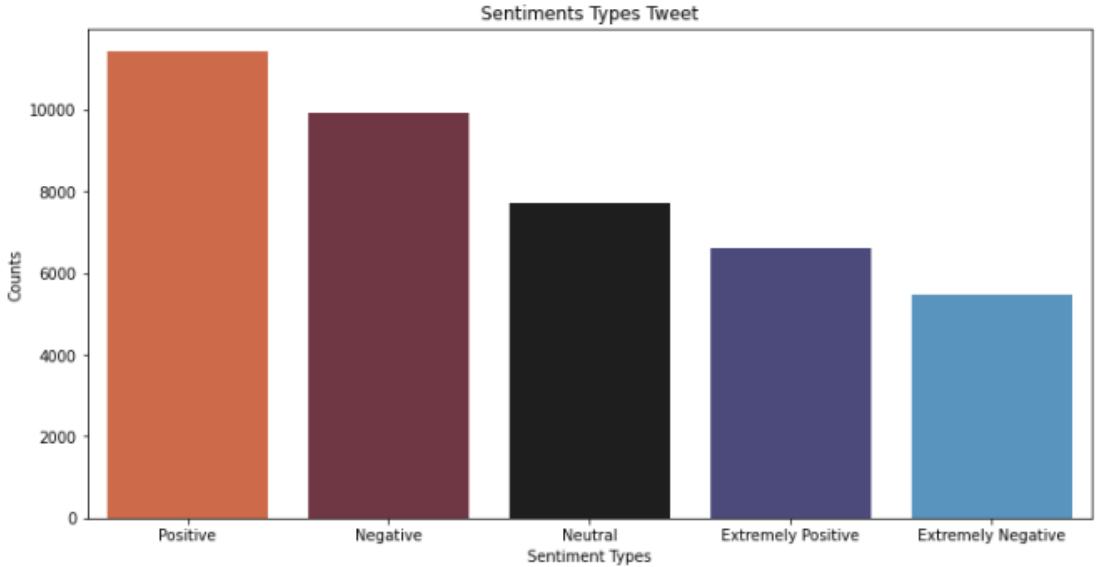
Visualization:



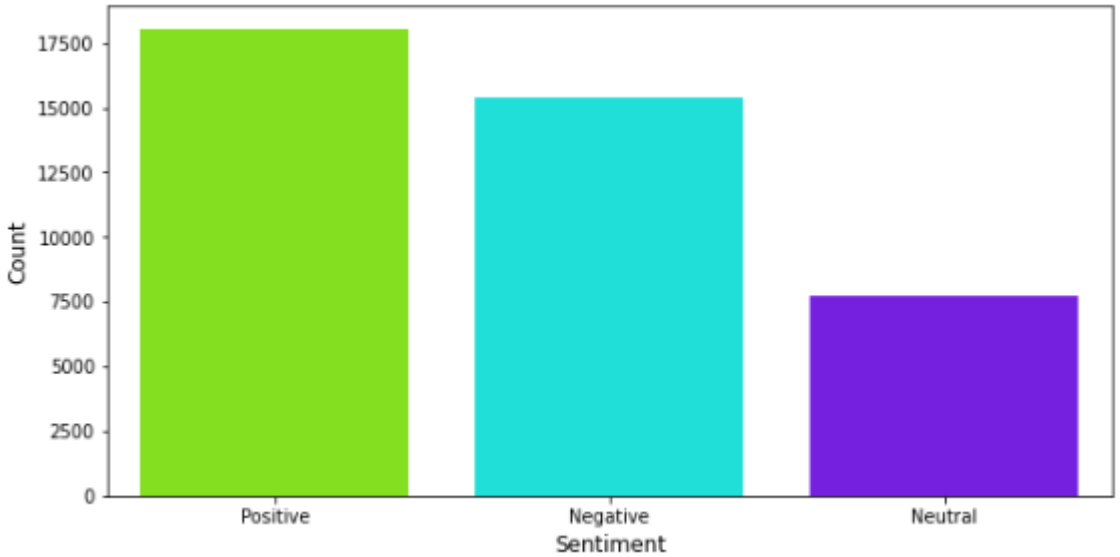
Exploring the Sentiment Column:

This Data clearly shows that positive it the maximum sentiment tweet made in all the locations:

	Sentiment Types	Counts
0	Positive	11422
1	Negative	9917
2	Neutral	7713
3	Extremely Positive	6624
4	Extremely Negative	5481



- **There are 5 subcategories in this case, so we will combine 5-class classification problem into a 3-class classification problem by replace Extremely Positive tweets with positive tweets and Extremely Negative tweets with negative tweets.**

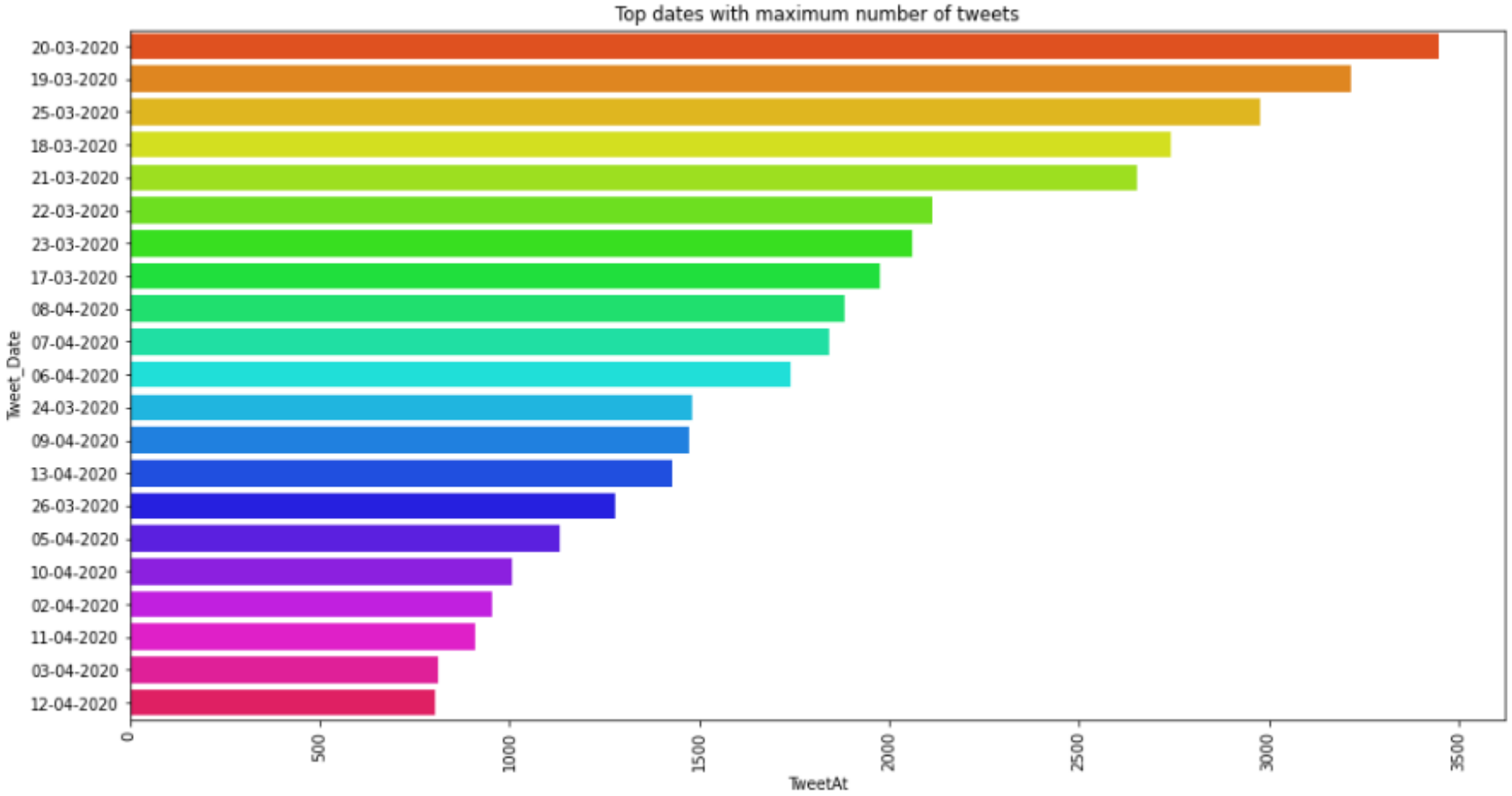


	Sentiment	count
0	Positive	18046
1	Negative	15398
2	Neutral	7713

Top dates with maximum number of tweets:

The table and the graph shows the top dates of the year with the maximum no of tweet.20/03/2022 is the date where maximum number of Tweet was made followed by 19/03/2022.

	Tweet_Date	TweetAt
0	20-03-2020	3448
1	19-03-2020	3215
2	25-03-2020	2979
3	18-03-2020	2742
4	21-03-2020	2653
5	22-03-2020	2114
6	23-03-2020	2062
7	17-03-2020	1977
8	08-04-2020	1881
9	07-04-2020	1843
10	06-04-2020	1742
11	24-03-2020	1480
12	09-04-2020	1471
13	13-04-2020	1428
14	26-03-2020	1277
15	05-04-2020	1131
16	10-04-2020	1005
17	02-04-2020	954
18	11-04-2020	909
19	03-04-2020	810
20	12-04-2020	803


Top dates with maximum number of tweets
20-03-2020
19-03-2020
25-03-2020
18-03-2020
21-03-2020
22-03-2020
23-03-2020
17-03-2020
08-04-2020
07-04-2020
06-04-2020
24-03-2020
09-04-2020
13-04-2020
26-03-2020
05-04-2020
10-04-2020
02-04-2020
11-04-2020
03-04-2020
12-04-2020
Tweet_Date
0
500
1000
1500
2000
2500
3000
3500
TweetAt

Data Pre-Processing:

Removing Punctuation

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment	TokenizedTweet
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	MeNyrbie PhilGahan Chrisitv httpstcoiFz9FAn2Pa...
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive	advice Talk to your neighbours family to excha...
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia Woolworths to give elder...
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive	My food stock is not the only one which is emp...
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative	Me ready to go at supermarket during the COVID...

Removing Stop words and Stemming:

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment	TokenizedTweet
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	MeNyrbie PhilGahan Chrisitv httpstcoiFz9FAn2Pa...
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive	advice Talk neighbours family exchange phone n...
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive	Coronavirus Australia Woolworths give elderly ...
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive	food stock one empty PLEASE dont panic ENOUGH ...
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative	ready go supermarket COVID19 outbreak Im paran...

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment	TokenizedTweet
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i...	Neutral	menyrbi philgahan chrisitv httpstcoifz9fan2pa ...
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to excha...	Positive	advic talk neighbour famili exchang phone numb...
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elde...	Positive	coronavirus australia woolworth give elder dis...
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is emp...	Positive	food stock one empti pleas dont panic enough f...
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COV...	Extremely Negative	readi go supermarket covid19 outbreak im paran...

Word Cloud:

After removing punctuation, stop words and using stemming to words cropped to save space from the tweets we want to know which words are mostly used by the people.

Building Classification Models:

- We will be using tf-idf method for vectorising the text.
- Then, we will implement 4 models:
  - * Logistic Regression
  - * Random Forest Classifier
  - *Naive Bayes Classifier
  - *Support Vector Machine(SVM)
- We will determine which model has the highest accuracy score before selecting it for model building.

Splitting the data set:

Independent Variable- TokenizedTweet

Dependent Variable- Sentiment

X_train - (32925)

X_test - (32925)

y_train - (8232)

y_train - (8232)

Vectorization:

Creating an object of TfidfVectorizer, the test data was normalised, and stored in the variables X_test and X_train, and also both predicting actual and predicted values.

X_train= (32925, 63453)

X_test= (8232, 63453)

Y_train= (32925)

Y_test= (8232)

Implementing Logistic Regression:

```
Training accuracy Score :  0.8091
Testing accuracy Score :  0.5804
                    precision    recall  f1-score   support

Extremely Negative       0.53      0.67      0.59       860
Extremely Positive       0.57      0.67      0.61      1128
          Negative       0.53      0.53      0.53      1989
           Neutral       0.65      0.65      0.65      1542
          Positive       0.61      0.51      0.56      2713

          accuracy                           0.58      8232
         macro avg       0.58      0.61      0.59      8232
      weighted avg       0.58      0.58      0.58      8232
```

Implementing Random Forest Classifier:

```
Training accuracy Score :  0.999969627942293
Testing accuracy Score :  0.5252672497570456
                    precision    recall  f1-score   support

Extremely Negative       0.35      0.72      0.47       534
Extremely Positive       0.36      0.69      0.47       686
          Negative       0.49      0.49      0.49      1969
           Neutral       0.73      0.55      0.63      2039
          Positive       0.60      0.46      0.52      3004

          accuracy                           0.53      8232
         macro avg       0.51      0.58      0.52      8232
      weighted avg       0.57      0.53      0.53      8232
```

Implementing Naive Bayes Classifier:

```
Training accuracy Score :  0.5162642369020501
Testing accuracy Score :  0.3595724003887269
                    precision    recall  f1-score   support

Extremely Negative       0.01      0.88      0.03        17
Extremely Positive       0.02      0.83      0.03        24
          Negative       0.39      0.40      0.39      1907
           Neutral       0.05      0.80      0.09        91
          Positive       0.91      0.34      0.49      6193

          accuracy                           0.36      8232
         macro avg       0.27      0.65      0.21      8232
      weighted avg       0.77      0.36      0.46      8232
```

Implementing Support Vector Machine(SVM)

```
Training accuracy Score :  0.9659529233105543
Testing accuracy Score :  0.5954810495626822
                    precision    recall  f1-score   support

Extremely Negative       0.48      0.72      0.58       730
Extremely Positive       0.53      0.75      0.62       936
          Negative       0.58      0.54      0.56      2135
           Neutral       0.63      0.67      0.65      1451
          Positive       0.68      0.52      0.59      2980

          accuracy                           0.60      8232
         macro avg       0.58      0.64      0.60      8232
      weighted avg       0.61      0.60      0.59      8232
```

- Support Vector Classifier has performed slightly better than the Logistic regression and got the highest test accuracy score around 60%.
- Multinomial Naive Bayes performed the worst with test accuracy score of just 0.35.

Conclusion:

- The majority of the tweets were around 250 characters long, indicating that there was a lot of interest in COVID-19 among the general public.
- More positive tweets than neutral or negative ones were tweeted globally.
- People tweeted more in March than in April since many nations imposed lockdown during this time.
- The United States and London (England) were the two countries with the most tweets.
- We saw inconsistent responses from Australia during the pandemic, with nearly equal numbers of positive and negative tweets.
- Words like COVID19, grocery, supermarket, shop, price, etc. are frequently used in tweets, indicating that throughout the pandemic, individuals were mostly concerned about food supply and their costs.
- Support Vector Classifier has performed slightly better than the Logistic regression and got the highest test accuracy score around 60%.
- Multinomial Naive Bayes performed the worst with test accuracy score of just 0.35.

Thank You