Capstone Project

Coronavirus Tweet Sentiment Analysis

By

By- Baishnavee Mahato
Data Science Trainee
AlmaBetter, Bengaluru

Content

Introduction:

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. The virus was declared a pandemic by World Health Organization on 11th March 2020. Some people lost their lives, but many of us successfully defeated this new strain i.e. Covid-19.

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is Positive, Negative, or Neutral.

In this project, we are going to predict the Sentiments of COVID-19 tweets. The data gathered from the Tweeter and I'm going to use Python environment to implement this project.

Problem Statement:

The challenge is to build a classification model to predict the sentiment of COVID-19 tweets.The tweets have been pulled from Twitter and manual tagging has been done then.

The names and usernames have been given codes to avoid any privacy concerns.

I have given the following information:

1. Location
2. Tweet At
3. Original Tweet
4. Label

Project Scope:

The project is an **Machine Learning (Classification)** of the Data set of Coronavirus Tweet Sentiment Analysis.

- Copy of Coronavirus Tweets.csv

Libraries such as **Numpy, Pandas, Matplotlib, Seaborn** are used to **Analysis** & **Visualise** the data, and Machine Learning Algorithm Such as Logistic Regression, Random Forest Classifier, Catboost Model, Naive Bayes Classifier are used to build a classification model.

Data Summary

Data set name: Coronavirus tweets.csv

Shape :

- Rows: **41157**
- Columns: **6**

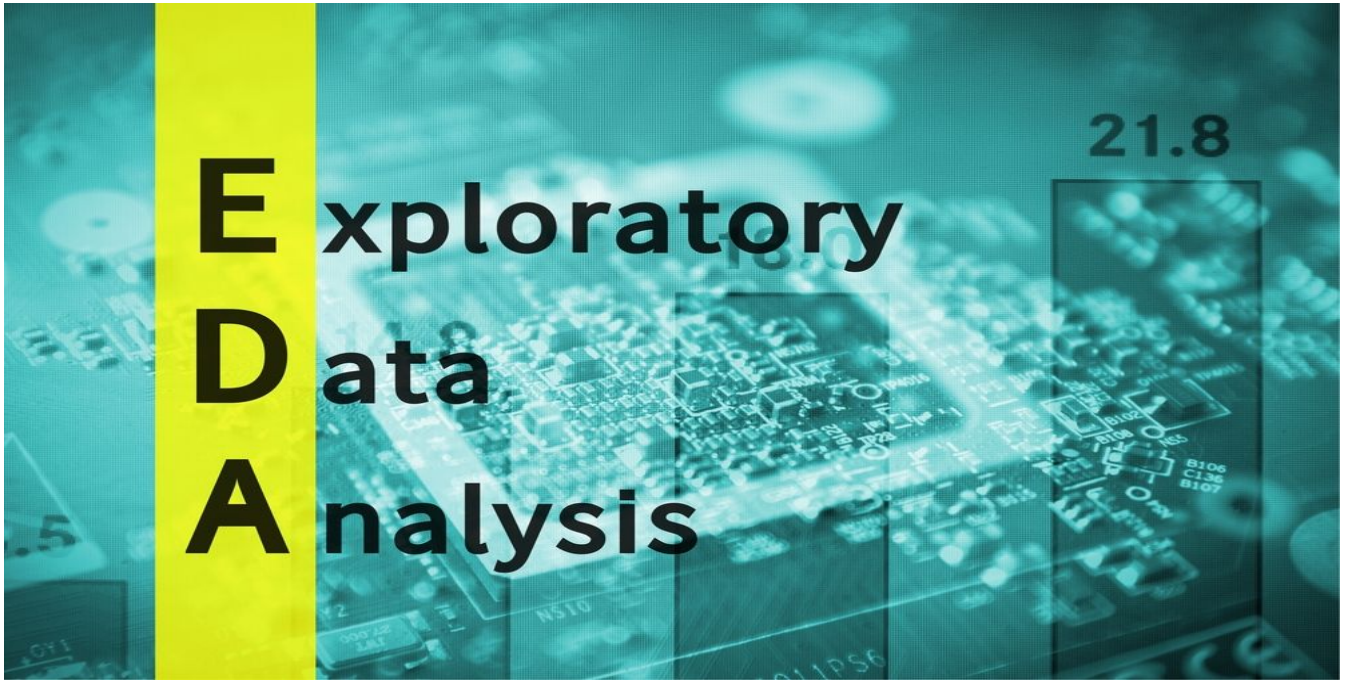
Columns: UserName, ScreenName, Location, TweetAt, OriginalTweet, Sentiment.

Description of Dataset Columns:

- **Username:** Twitter handle username
- **Screenname:** The name that users choose to identify themselves on the network.
- **Location**: Place (Country) from where the tweets came from
- **TweetAt**: Date in which the tweet has been posted
- **OriginalTweet**: The content of the posted tweet
- **Sentiment**: Behaviour at the tweets of the people.There are five types of sentiments- Extremely Negative, Negative, Neutral, Positive, and Extremely Positive.

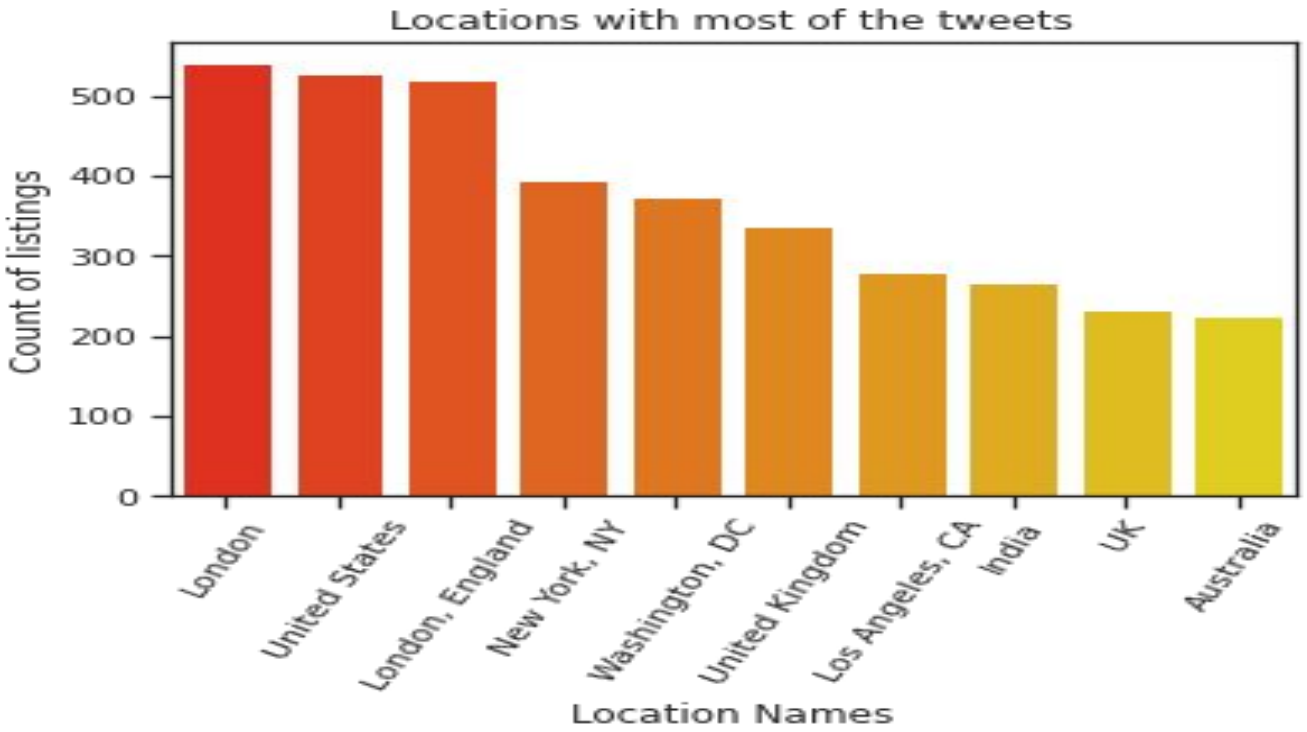
Top 5 Rows of the dataset

	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate su...	Positive
2	3801	48753	Vagabonds	16-03-2020	Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P	Positive
3	3802	48754	NaN	16-03-2020	My food stock is not the only one which is empty...\r\r\n\r\r\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \r\r\nStay calm, stay safe.\r\r\n\...	Positive
4	3803	48755	NaN	16-03-2020	Me, ready to go at supermarket during the #COVID19 outbreak.\r\r\n\r\r\nNot because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don...	Extremely Negative



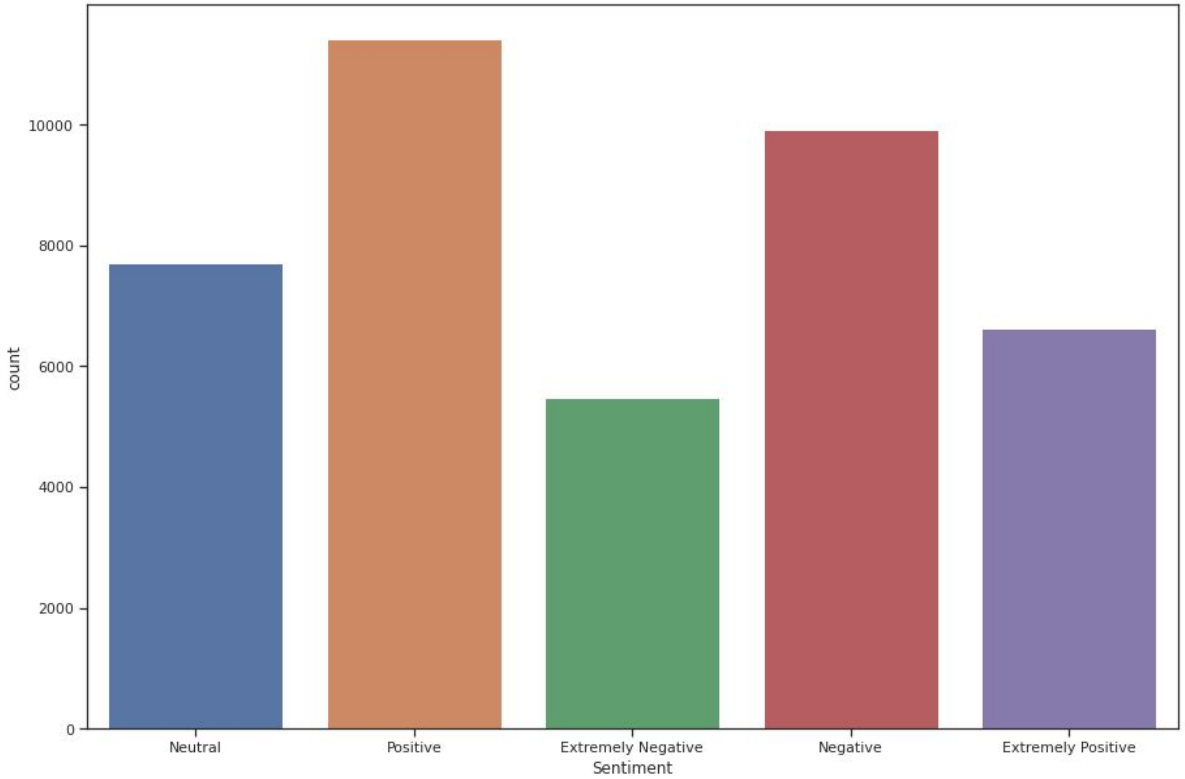
A picture is worth a thousand words.

Top 10 Locations of Tweets



Maximum tweets came from London.

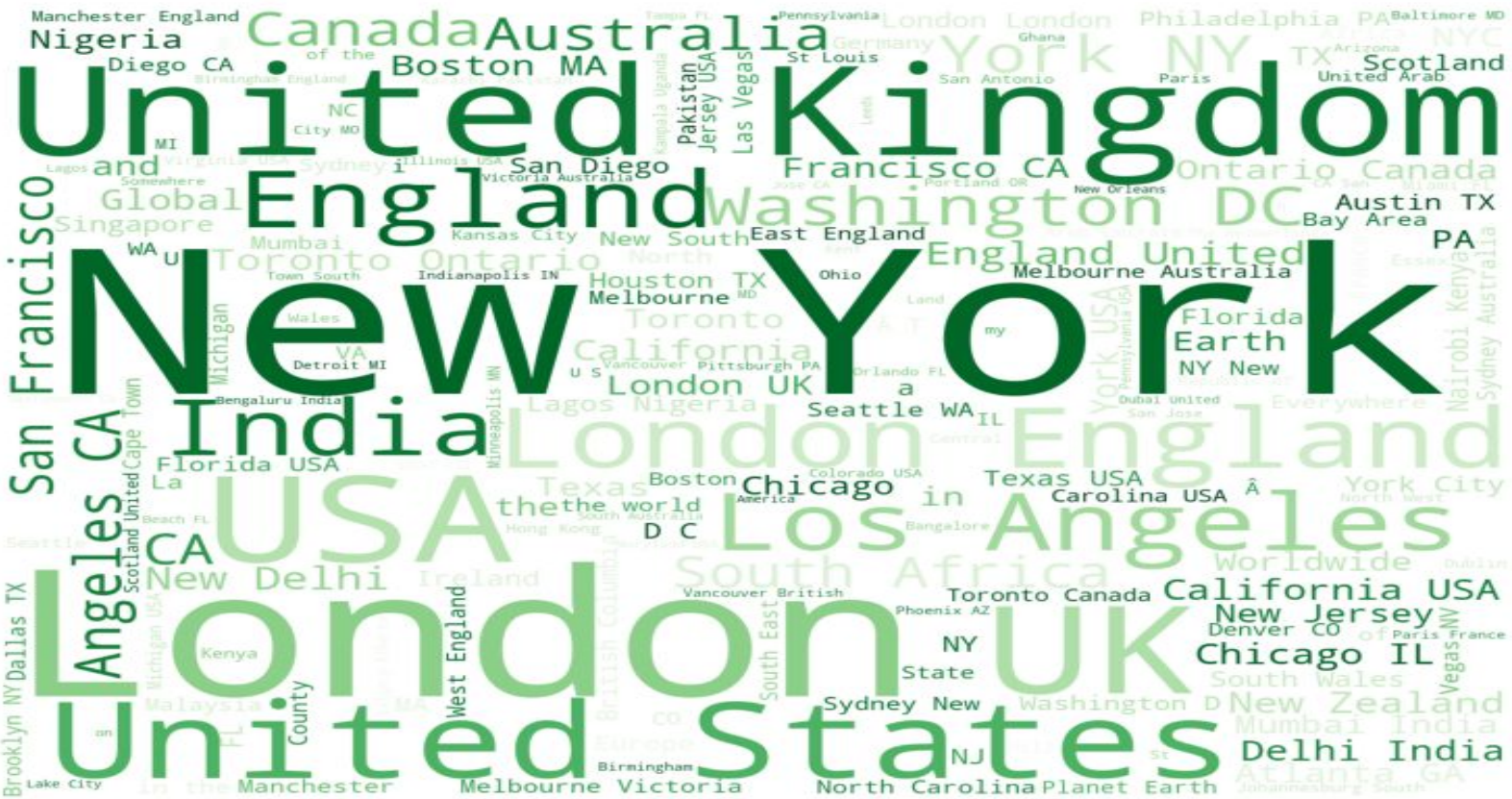
Sentiment Column



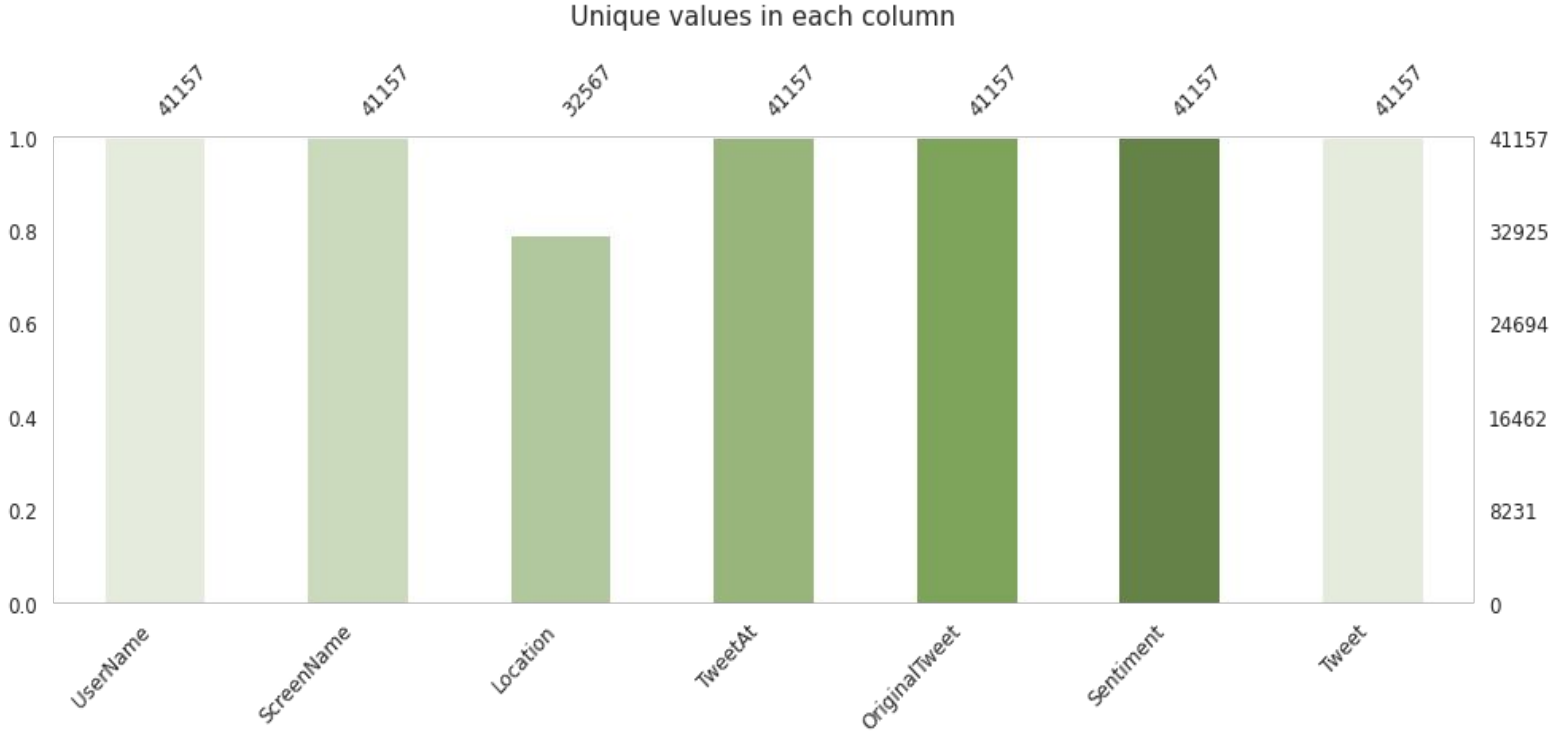
From the ‘Sentiment’ column, I came to know that most of the peoples are having positive sentiments about various issues shows us their optimism during pandemic times. Very few people are having extremely negatives thoughts about Covid-19.

Word Cloud for top 50 prevelant words in India

A WordCloud of the various location in the dataset



Unique Values in each column



Columns such as “UserName” and “ScreenName” do not give any meaningful insights for our analysis. Hence we are not using these features for model building.

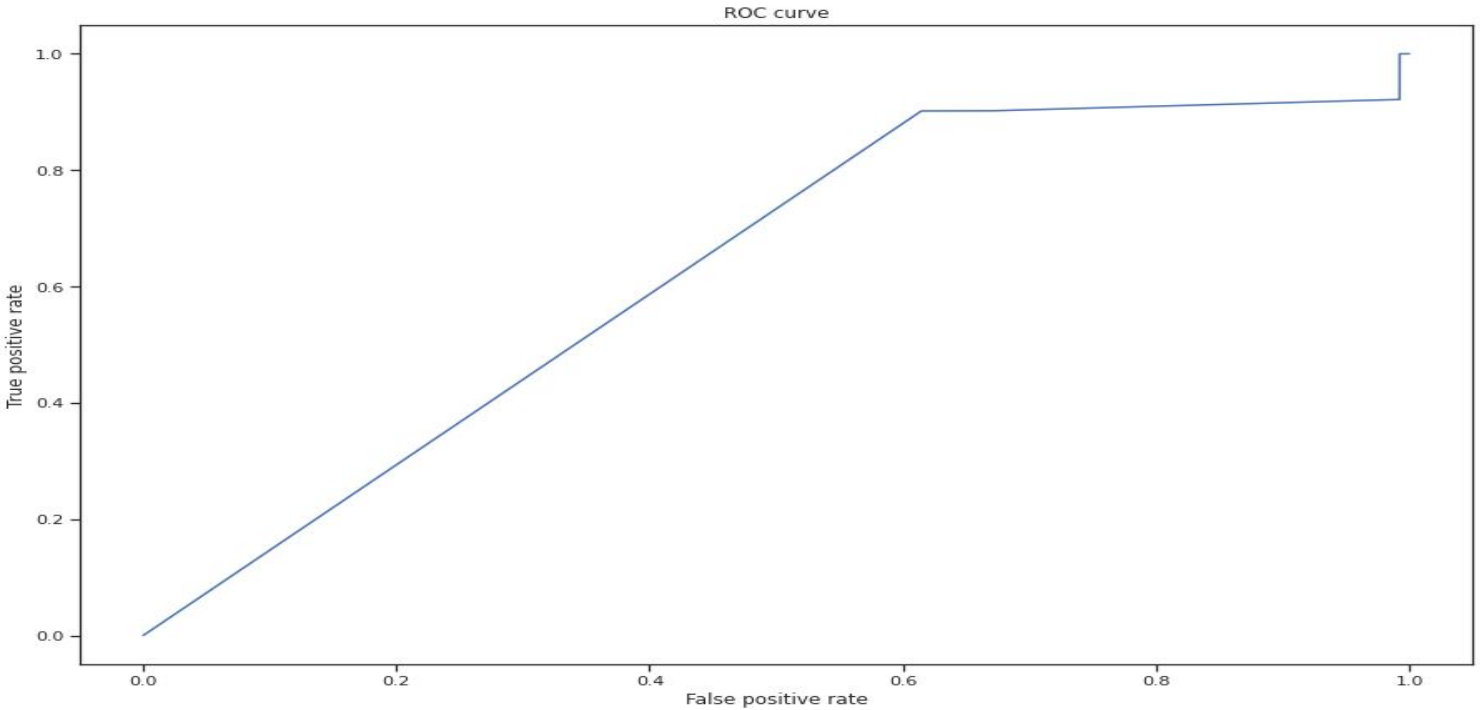
Data Preprocessing

The preprocessing of the text data is an essential step as it makes the raw text ready for mining. The objective of this step is to clean noise those are less relevant to find the sentiment of tweets such as punctuation(.,?,” etc.), special characters(@,%,&,$, etc.), numbers(1,2,3, etc.), twitter handle, links(HTTPS: / HTTP:)and terms which don’t carry much weightage in context to the text.

Building a Classification Model

There are five types of sentiments so we have to train our model so that it can give us the correct label for the test dataset. I am going to built different models like Logistic Regression, Random Forest Classifier,Catboost Classifier,Naive Bayes.

AUC Curve Plotting



Here we got a score of **AUC = 0.63** for **the classifier (Naive Byes)**, we can say that the classifier (Naive Bayes) is not that so good but can acceptable. Since the more nearer to 1 AUC score, the classifier will be better.

Multiclass Classification

I have used Multiclass Classification.
Dependent variable has the values - Positive, Extremely Positive, Neutral, Negative, Extremely Negative.

Multiclass Classification

Model	Test Accuracy
Logistic Regression	61.6%
Random Forest Classifier	54.7%
Naive Bayes	46.7%
Catboost Model	62.3%

Catboost Model is best for this dataset.

Classification Report for Catboost Model

Training accuracy Score : 0.6682763857251329
Testing accuracy Score : 0.6232993197278912

	precision	recall	f1-score	support
Extremely Negative	0.52	0.70	0.60	816
Extremely Positive	0.56	0.77	0.65	966
Negative	0.52	0.58	0.55	1785
Neutral	0.85	0.59	0.70	2248
Positive	0.64	0.60	0.62	2417
accuracy			0.62	8232
macro avg	0.62	0.65	0.62	8232
weighted avg	0.65	0.62	0.63	8232

The best model for this Dataset is Catboost model with Testing Accuracy Score: 0.6232

Conclusion

1. The maximum tweets came from London.
2. From the ‘Sentiment’ column, I came to know that most of the peoples are having positive sentiments about various issues shows us their optimism during pandemic times. Very few people are having extremely negatives thoughts about Covid-19.
3. Columns such as “UserName” and “ScreenName” do not give any meaningful insights for our analysis. Hence we are not using these features for model building.
4. From AUC Curve the AUC Score is 0.63 for the Naive Bayes Classifier which is not good but acceptable.
5. For multiclass classification, the best model for this dataset is CatBoost with Testing Accuracy Score = 0.62329.

Thank You