Coronavirus Tweet Sentiment Analysis

BAISHNAVEE MAHATO

Data science trainee,
AlmaBetter, Bangalore

1. 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 The 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.
The first case of COVID-19 was reported Dec. 1, 2019, and the cause was a then-new coronavirus later named SARS-CoV-2. SARS-CoV-2 may have originated in an animal and changed (mutated) so it could cause illness in humans. In the past, several infectious disease outbreaks have been traced to viruses originating in birds, pigs, bats and other animals that mutated to become dangerous to humans. Research continues, and more study may reveal how and why the coronavirus evolved to cause pandemic disease.
As of now, researchers know that the coronavirus is spread through droplets and virus particles released into the air when an infected person breathes, talks, laughs, sings, coughs or sneezes. Larger droplets may fall to the ground in a few seconds, but tiny infectious particles can linger in the air and accumulate in indoor places, especially where many people are gathered and there is poor ventilation. This is why mask-wearing, hand hygiene and physical distancing are essential to preventing COVID-19.
The Rossmann Store Sales dataset is a public dataset, which contains daily historical sales data for 3 Rossmann stores from the 1st January 2013 till the 31st July 2015.
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 the Python environment to implement this project.

2. Problem Statement:

The given 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. I have given information like Location, Tweet At, Original Tweet, and Sentiment.

3. Data Description:

Data set name: Coronavirus tweets.csv

Shape :

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

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

- **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.

4. Approach To Analyze Various Sentiments:

HereBefore we proceed further, One should know what is meant by Sentiment Analysis. 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. Following is the Standard Operating Procedure to tackle the Sentiment Analysis kind of project. We will be going through this procedure to predict what we are supposed to predict:

Exploratory Data Analysis.

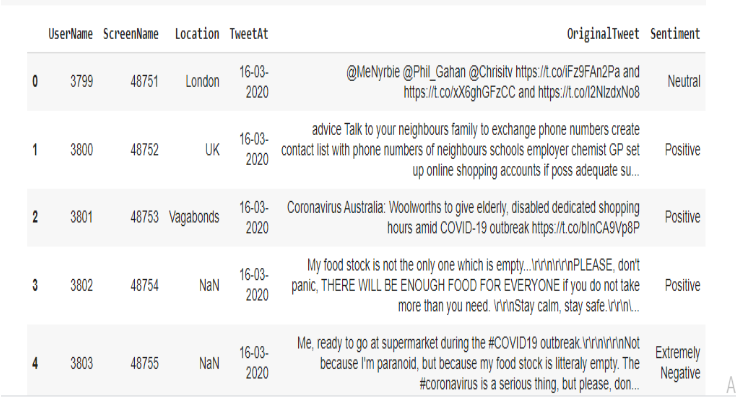
Data Preprocessing.

Classification Models.

Conclusion.

5. Data Summary:

The original dataset has 6 columns and 41157 rows. In order to analyze various sentiments, We require just two columns named Original Tweet and Sentiment. There are five types of sentiments- Extremely Negative, Negative, Neutral, Positive, and Extremely Positive.



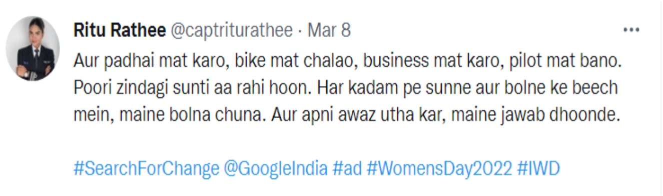
	UserName	ScreenName	Location	TweetAt	OriginalTweet	Sentiment
0	3799	48751	London	16-03-2020	@MeNyrbie @Phil_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8	Neutral
1	3800	48752	UK	16-03-2020	advice Talk to your neighbours family to exchange phone numbers 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

5. Basic Exploratory Data Analysis:

- Maximum tweets came from London.
- The 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.
- Majority 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.

6. 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 that 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.



Ritu Rathee @captriturathee · Mar 8
Aur padhai mat karo, bike mat chalao, business mat karo, pilot mat bano. Poori zindagi sunti aa rahi hoon. Har kadam pe sunne aur bolne ke beech mein, maine bolna chuna. Aur apni awaz utha kar, maine jawab dhoonde.

#SearchForChange @GoogleIndia #ad #WomensDay2022 #IWD

7. Building Classification Models:

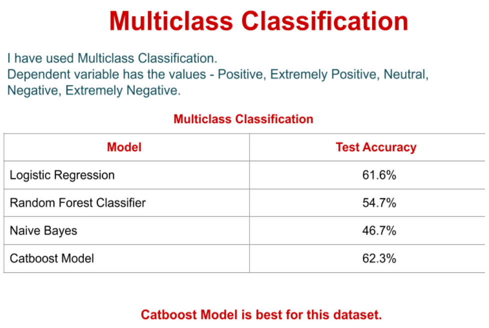
There are five types of sentiments so we have to train our model so that it can give us the correct label for the test dataset. I am going to build different models like Logistic Regression, Random Forest Classifier,Catboost Classifier,Naive Bayes.

Here we got a score of AUC = 0.63 for the classifier (Naive Bayes), we can say that the classifier (Naive

Bayes) is not that so good but can be acceptable. Since the closer to 1 AUC score, the classifier will be Better.

I have used Multiclass Classification.

Dependent variable has the values - Positive, Extremely Positive, Neutral, Negative, Extremely Negative.



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.

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

7. Conclusion:

- The maximum tweets came from London.
- From the ‘Sentiment’ column, I came to know that most of the peoples are having positive sentiments about various issues showing us their optimism during pandemic times. Very few people are having extremely negative thoughts about Covid-19.
- 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.
- From AUC Curve the AUC Score is 0.63 for the Naive Bayes Classifier which is not good but acceptable.
- For multiclass classification, the best model for this dataset is CatBoost with Testing Accuracy Score = 0.62329.

8. References:

https://www.analyticsvidhya.com/blog/2021/02/sentiment-analysis-predicting-sentiment-of-covid-19-tweets/

https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus