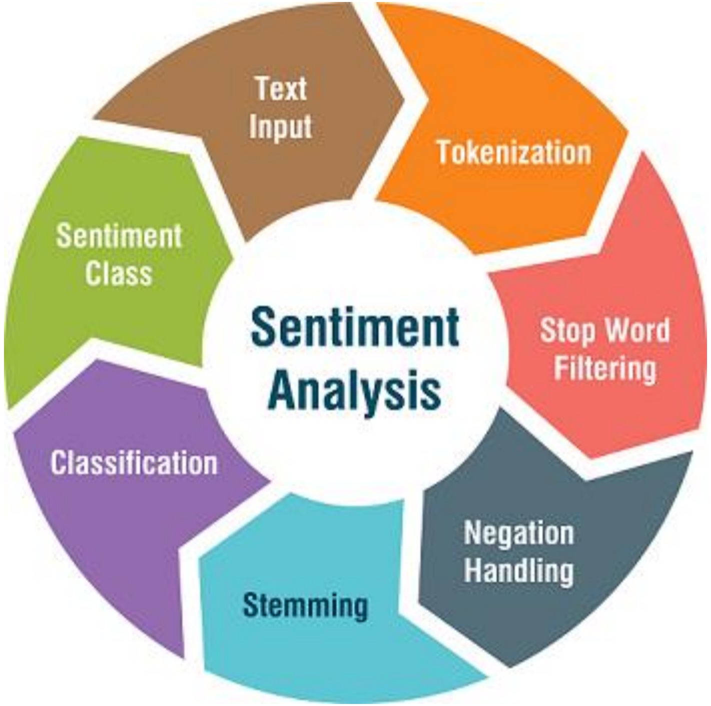
Capstone Project

Corona Tweet Sentiment Analysis

Content

1. Introduction.
2. Exploratory Data Analysis.
3. Data Preprocessing.
4. Vectorization.
5. Classification.
6. Evaluation.
7. Challenges.
8. Conclusion
9. Q&A



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.
- Given the following information:

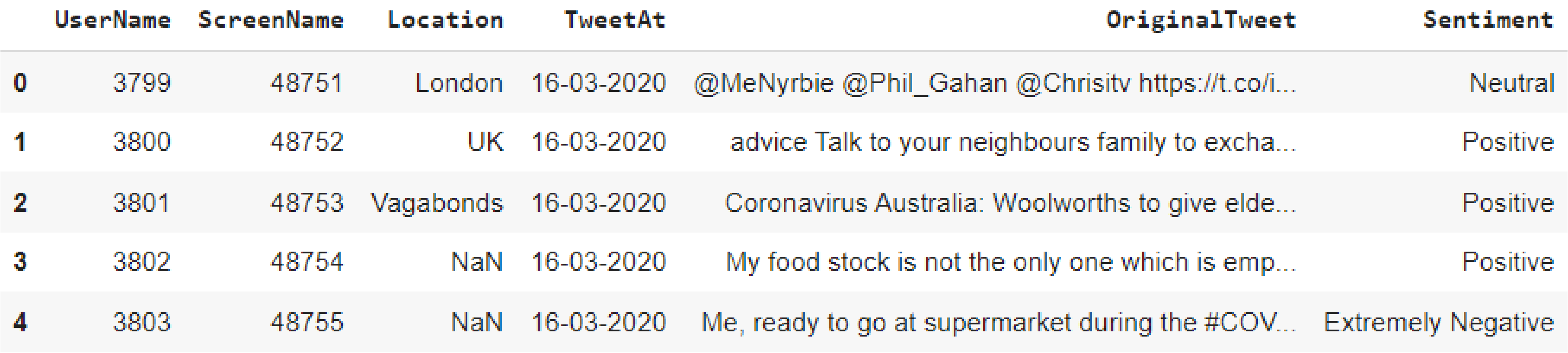
1. Location
2. Tweet At
3. Original Tweet
4. Label

Introduction

- Sentiment Analysis is the **process of determining whether a piece of writing** (product/movie review, tweet, etc.) is positive, negative or neutral. It can be used to identify the customer or follower's attitude towards a brand through the use of variables such as context, tone, emotion, etc.
- Dataset consist of tweets related to covid-19 during pandemic,covid-19 is originally known as Coronavirus Disease of 2019 and declared as a pandemic by World Health Organization (WHO) on 11th march 2020.
- The study analyzes various types of tweets gathered during the pandemic times hence can be useful in policy making to safeguard the countries by demystifying the pertinent facts and information.

Data Summary

- The dataset consist of 6 columns and 41157 rows.
- There are five sentiments: Positive, Extremely Positive, Neutral, Negative, Extremely Negative.
- Tweet data given from 16-03-2020 to 14-04-2020.



	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

Some sample tweets…

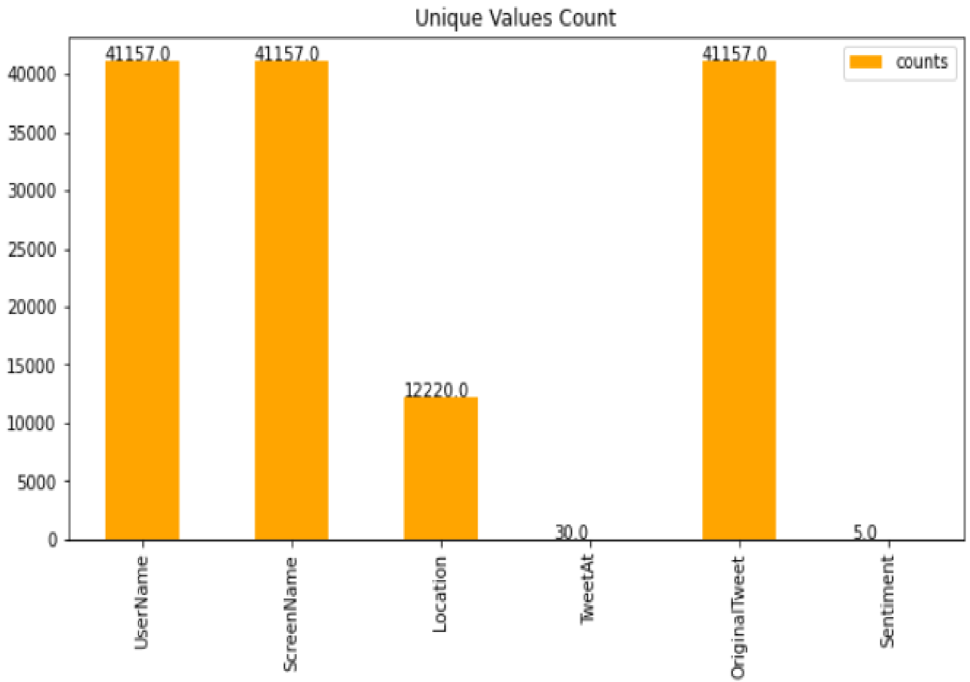
“This morning I tested positive for Covid 19. I feel ok, I have no symptoms so far but have been isolated since I found out about my possible exposure to the virus. Stay home people and be pragmatic. I will keep you updated on how IÂ’m doing ???? No panic. https://t.co/Lg7HVMZglZ”

“For corona prevention, we should stop to buy things with the cash and should use online payment methods because corona can spread through the notes. Also we should prefer online shopping from our home. It's time to fight against COVID 19?. #govindia #IndiaFightsCorona”

EDA

Unique Values:

- The Sentiment column contains five unique values.
- TweetAt column has 30 unique values i.e. 30 days.
- There are no null values in Dataset.



EDA

Sentiment Column:

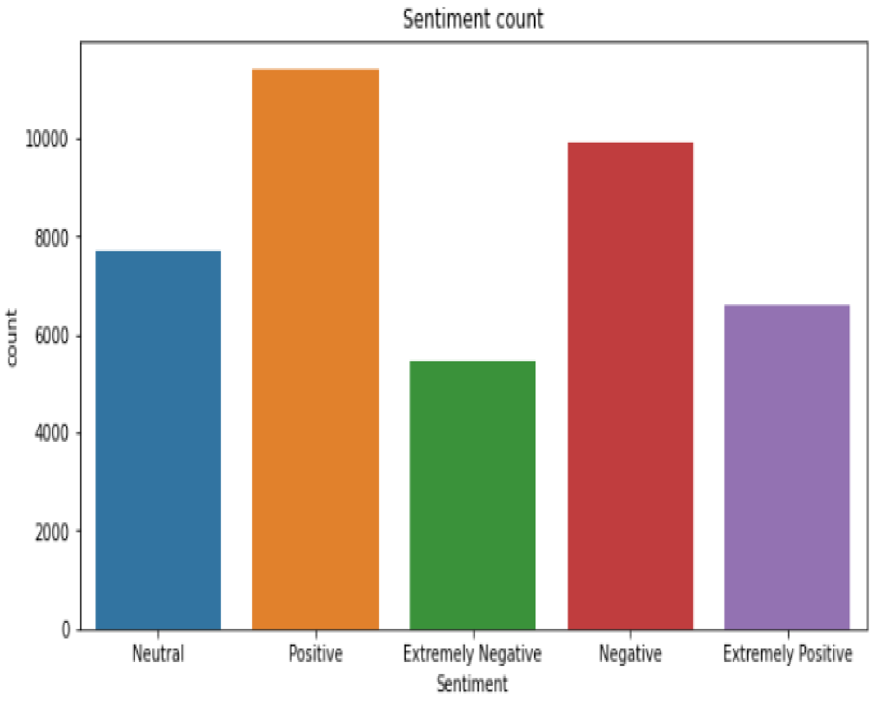
- Sentiment column has total 41157 entries.
- In that, Positive: 11422

  Negative: 9917

  Neutral: 7713

  Extremely Positive: 6624

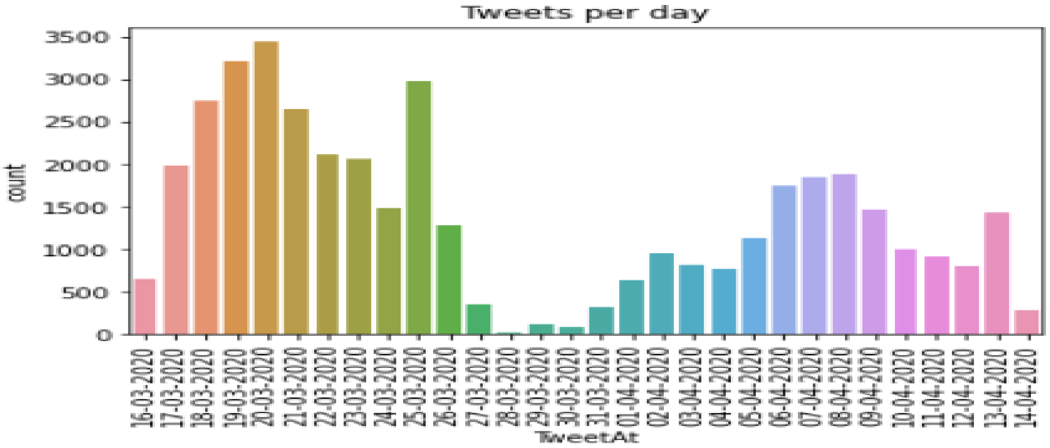
  Extremely Negative: 5481.
- In data most of the tweets are positive as compared to other sentiments.
- There is very less count of extremely negative tweets.

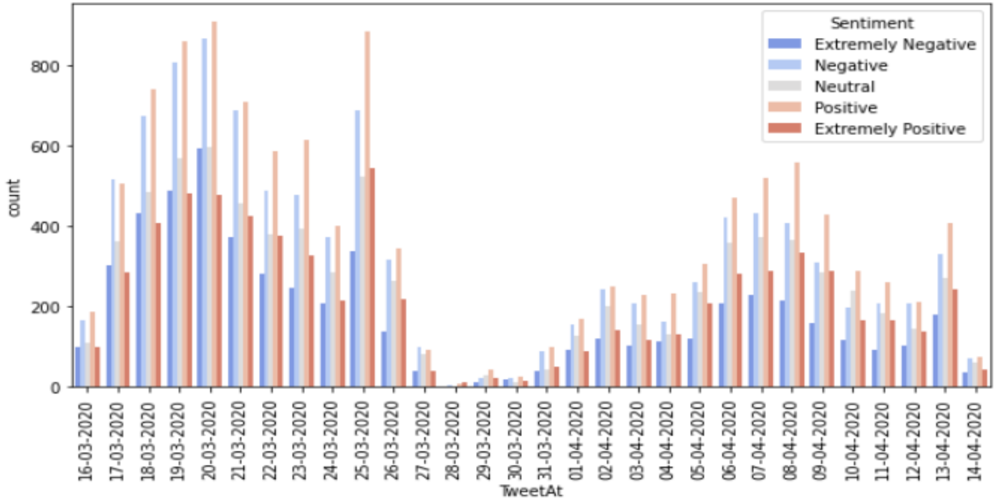


EDA

TweetAt column:

- Graph shows count of tweets per day in a month.
- It shows highest count of tweets is come from march.
- At March ending tweets count decreased dramatically.
- Highest tweet count was on 20th March i.e. 3448 tweets. And lowest count on 28th March i.e. 23 tweets only.
- In 2nd graph, per day tweets with sentiment count is plotted.
- It shows positive sentiment tweets count is always highest in each day.

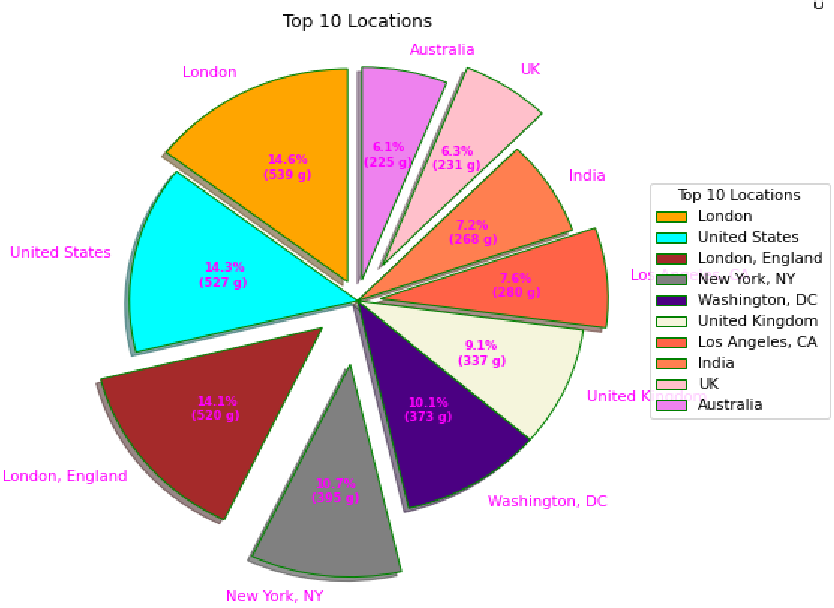


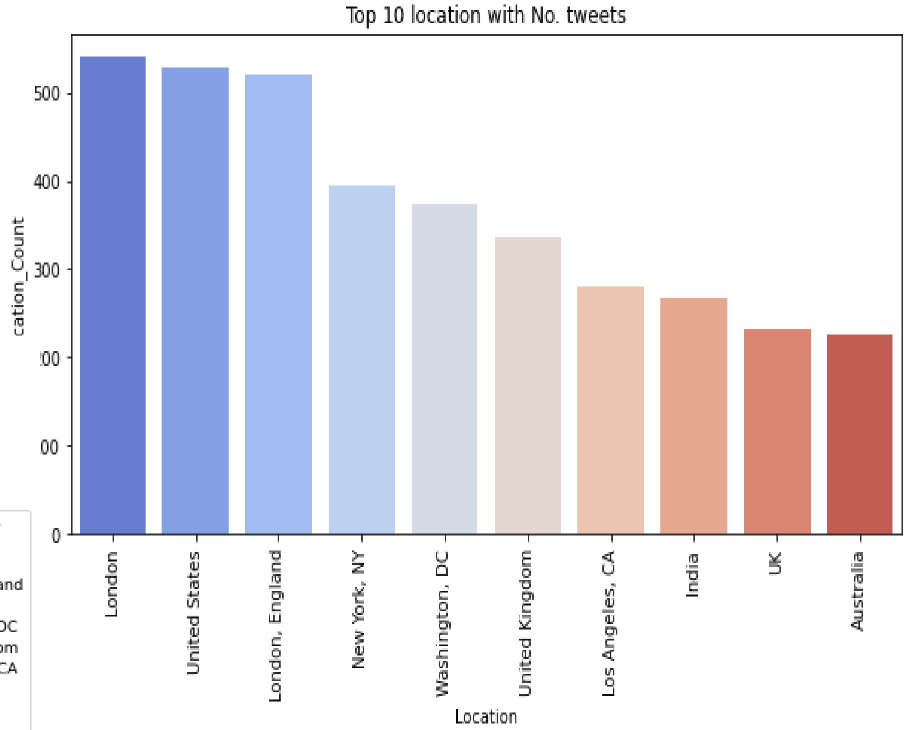


EDA

Location column:

- Graph shows top 10 locations which has highest count of tweets.
- In that London is on the Top.

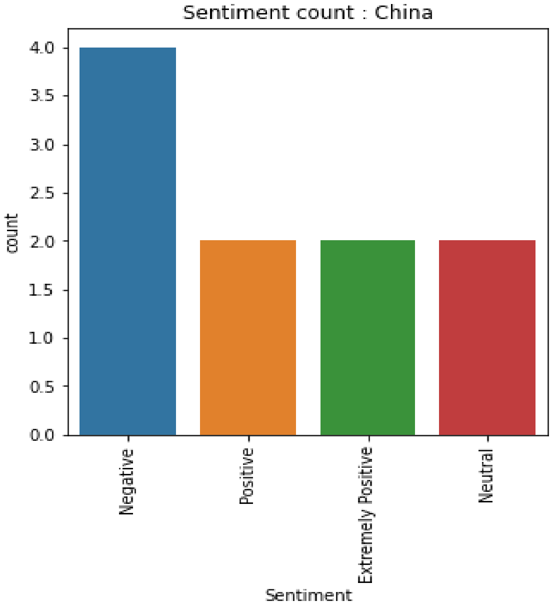


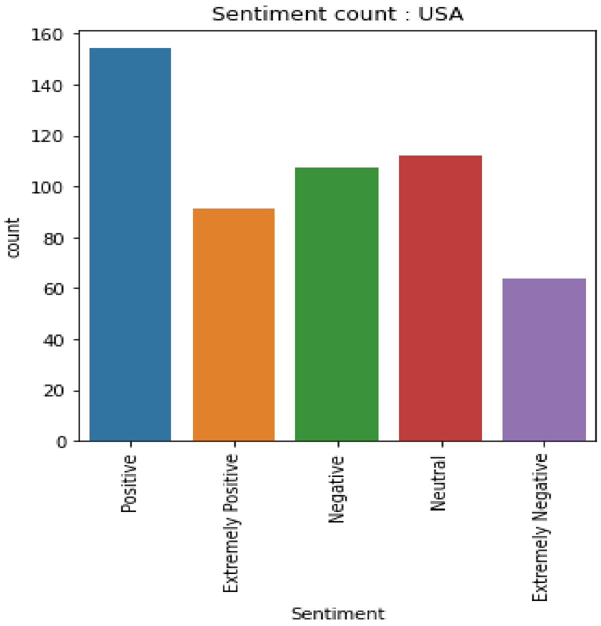


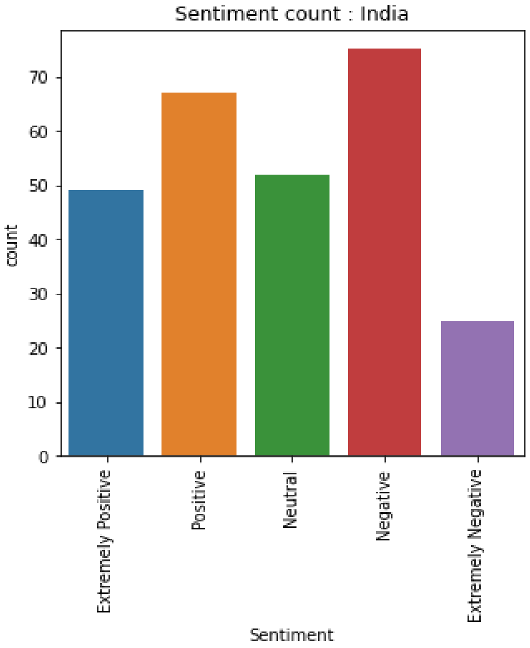
EDA

Sentiment wise tweet count of Three major countries: China, USA, India.

- Most of the tweets from USA are positive but from India and China Negative tweets are high.







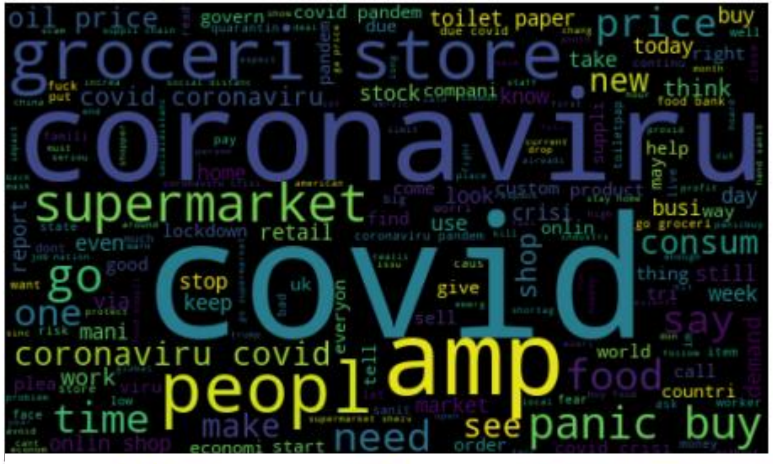
EDA

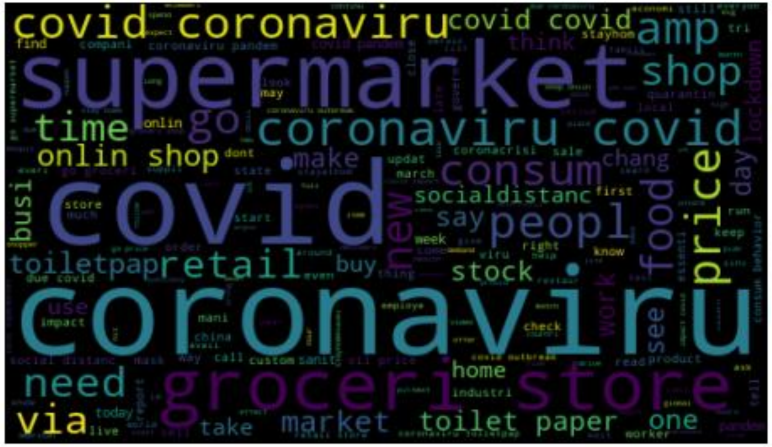
Original tweet column:

- In all sentiment tweets some words like ‘coronavirus’, ‘covid’, ‘people’, ‘supermarket’ has maximum frequency in our dataset.
- There are some common #Hashtags in tweet column like ‘coronavirus’, ‘covid_19’, ‘covid2019’.

Tweet Sample…

“#WestJet is lying. They are NOT lowering prices to get you out of the country.
These are screen shots for the same flight 2am 12pm 7pm
#covid_19 #canada @GlobalNational @globalnews @GlobalCalgary @CTVNews @CNN @JustinTrudeau https://t.co/pfEQVZVRf6”



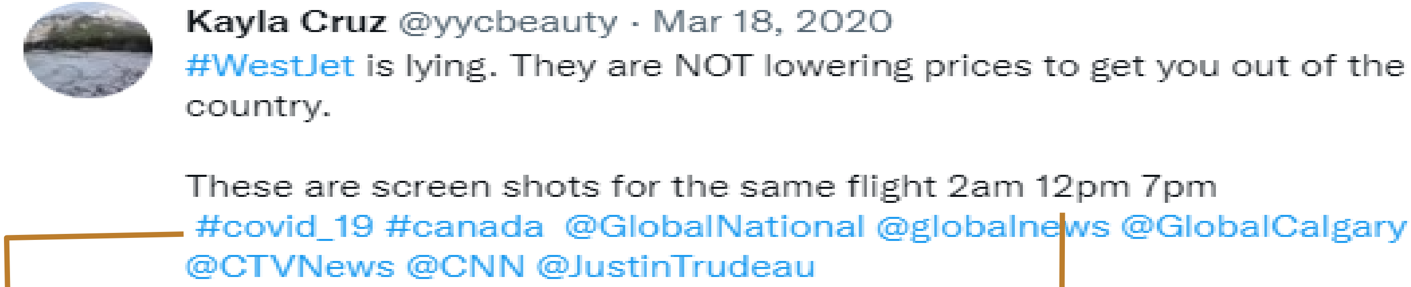


Data Processing

- Data Processing is important step it makes raw text data ready for mining.
- Raw text consist of noise means punctuations, special characters,numbers,symbols and stop words which don’t carry much weightage in context to the next.
- In this process we remove this noise from text and make it proccesible.

Text processing on Tweet:

- Text processing on Tweet:



Kayla Cruz @yycbeauty · Mar 18, 2020

#WestJet is lying. They are NOT lowering prices to get you out of the country.

These are screen shots for the same flight 2am 12pm 7pm #covid_19 #canada @GlobalNational @globalnews @GlobalCalgary @CTVNews @CNN @JustinTrudeau

#Hashtag

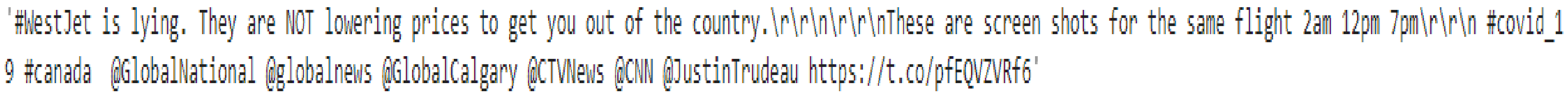
@user

Digits

Removing @user

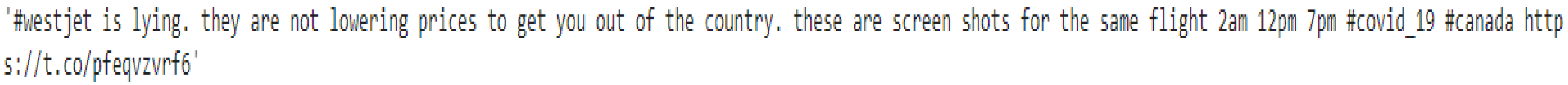
As mentioned earlier, the tweets contain lots of twitter handles @user. We will remove all these twitter handles from the data as they don’t convey much information.

Original tweet



```
'#WestJet is lying. They are NOT lowering prices to get you out of the country.\r\r\n\r\r\nThese are screen shots for the same flight 2am 12pm 7pm\r\r\n #covid_1
9 #canada  @GlobalNational @globalnews @GlobalCalgary @CTVNews @CNN @JustinTrudeau https://t.co/pfEQVZVRf6'
```

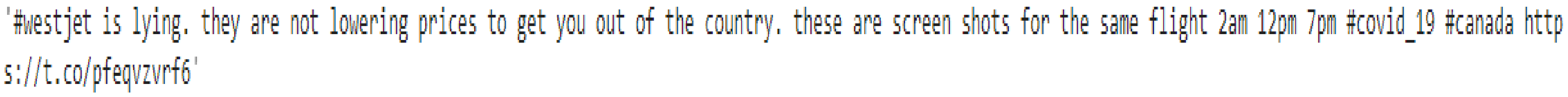
After



```
'#westjet is lying. they are not lowering prices to get you out of the country. these are screen shots for the same flight 2am 12pm 7pm #covid_19 #canada http
s://t.co/pfeqvzvrf6'
```

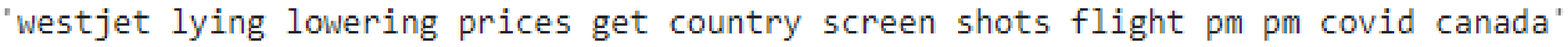
Removing Punctuations, Numbers, Hashtags and stopwords.

- As Punctuations,numbers,special characters, stopwords adds noise in data we are removing it from dataset.



```
'#westjet is lying. they are not lowering prices to get you out of the country. these are screen shots for the same flight 2am 12pm 7pm #covid_19 #canada http
s://t.co/pfeqvzvrf6'
```

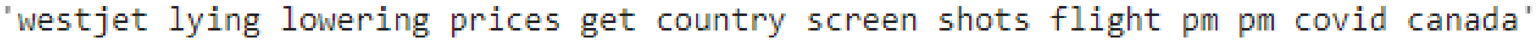
After



```
'westjet lying lowering prices get country screen shots flight pm pm covid canada'
```

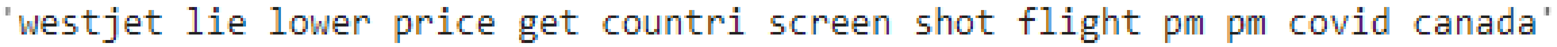
Stemming:

- **Stemming is a technique used to extract the base form of the words by removing affixes from them.**
- **It is just like cutting down the branches of a tree to its stems. For example, the stem of the words *eating, eats, eaten* is *eat*.**



```
'westjet lying lowering prices get country screen shots flight pm pm covid canada'
```

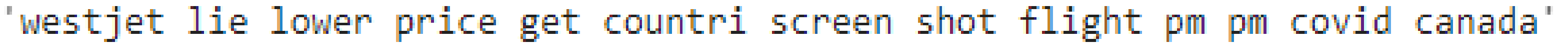
After



```
'westjet lie lower price get countri screen shot flight pm pm covid canada'
```

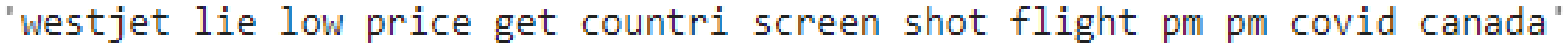
Lemmatization:

- Lemmatization is a linguistic term that means grouping together words with the same root or lemma but with different inflections or derivatives of meaning so they can be analyzed as one item.
- For example, to lemmatize the words “cats,” “cat’s,” and “cats’” means taking away the suffixes “s,” “’s,” and “s’” to bring out the root word “cat.”



```
'westjet lie lower price get countri screen shot flight pm pm covid canada'
```

After



```
'westjet lie low price get countri screen shot flight pm pm covid canada'
```

Tokenization

- **Tokenization basically refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language.**
- **Tokenization can be done in python by using NLTK library's word_tokenize() function.**

Vectorization

- Vectorization is a technique to implement arrays without the use of loops. Using a function instead can help in minimizing the running time and execution time of code efficiently.
- We choose count vectorizer as our vectorizer with minimum document frequency = 10.
- It will creat a sparse matrix of all words and the number of times they are present in a document.

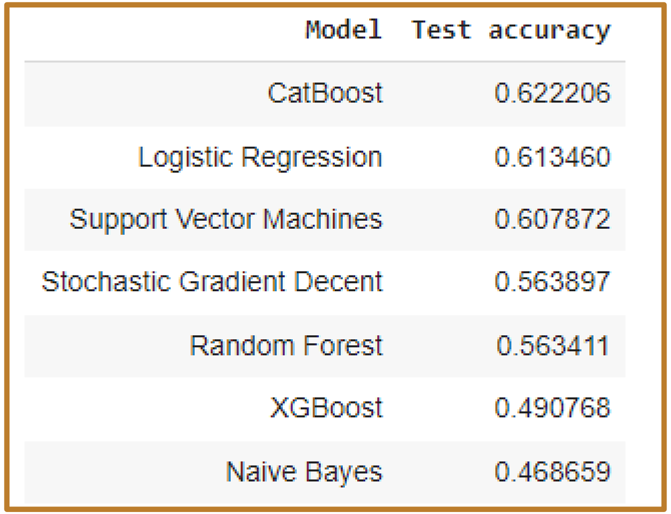
Classification

- **Models Used:**

1. Naïve Bayes
2. Logistic Regression
3. Random Forest
4. XGBoost
5. Support Vector Machines
6. CatBoost
7. Stochastic Gradient Descent

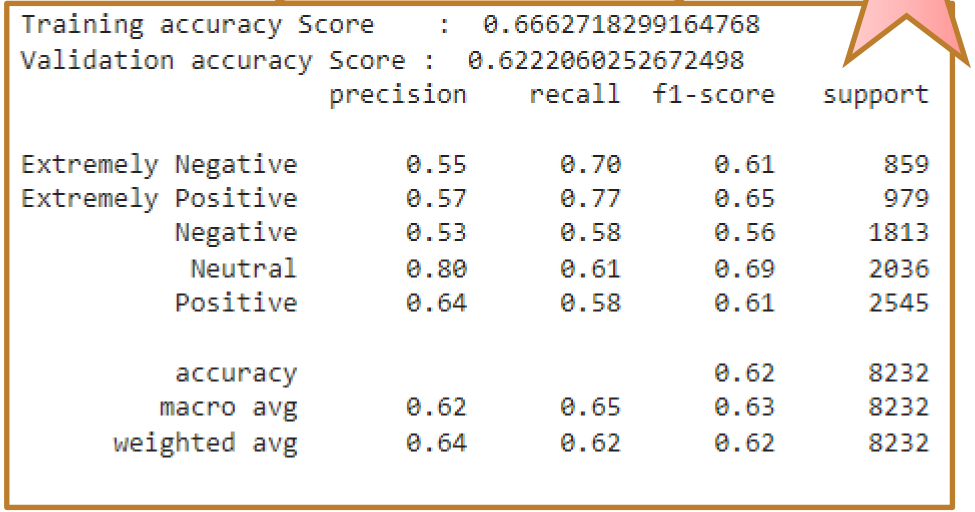
Evaluation

Result for Multi-class Classification



Model	Test accuracy
CatBoost	0.622206
Logistic Regression	0.613460
Support Vector Machines	0.607872
Stochastic Gradient Decent	0.563897
Random Forest	0.563411
XGBoost	0.490768
Naive Bayes	0.468659

CatBoost Result



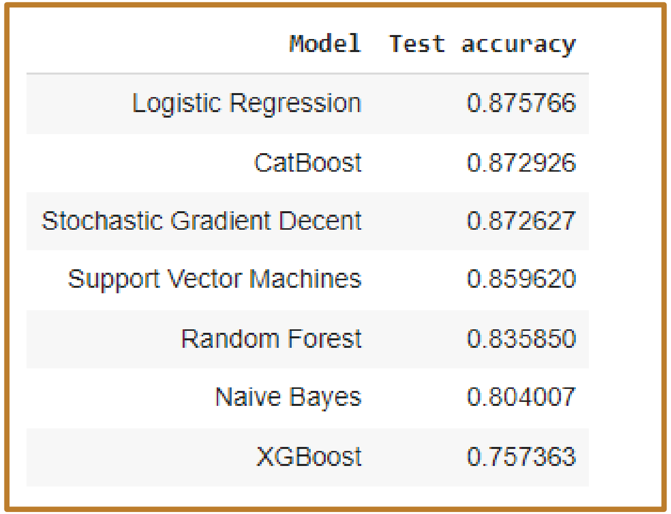
```
Training accuracy Score    :  0.6662718299164768
Validation accuracy Score :  0.6222060252672498
                    precision    recall  f1-score   support

Extremely Negative       0.55      0.70      0.61       859
Extremely Positive       0.57      0.77      0.65       979
          Negative       0.53      0.58      0.56      1813
           Neutral       0.80      0.61      0.69      2036
          Positive       0.64      0.58      0.61      2545

          accuracy                           0.62      8232
         macro avg       0.62      0.65      0.63      8232
      weighted avg       0.64      0.62      0.62      8232
```

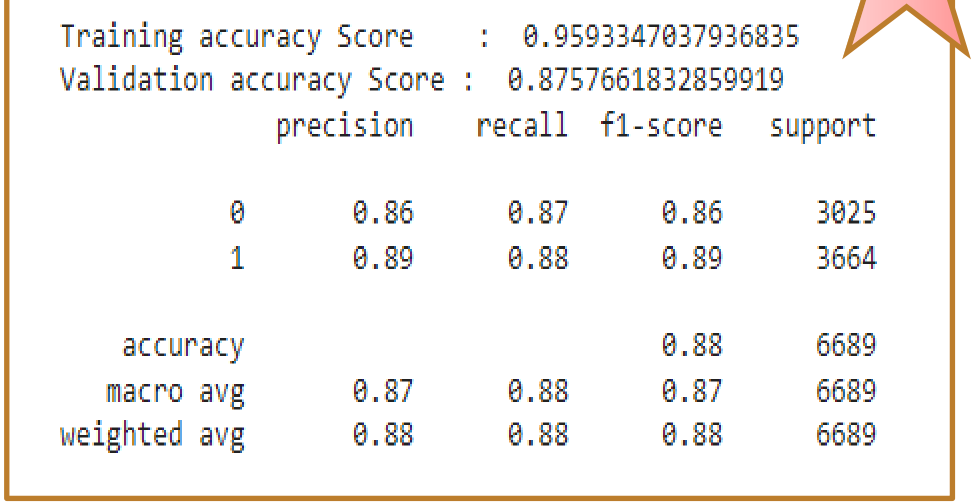
Evaluation

Result for Binary Classification



Model	Test accuracy
Logistic Regression	0.875766
CatBoost	0.872926
Stochastic Gradient Decent	0.872627
Support Vector Machines	0.859620
Random Forest	0.835850
Naive Bayes	0.804007
XGBoost	0.757363

Logistic Regression Result



```
Training accuracy Score    :  0.9593347037936835
Validation accuracy Score :  0.8757661832859919
              precision    recall  f1-score   support

           0       0.86      0.87      0.86      3025
           1       0.89      0.88      0.89      3664

    accuracy                           0.88      6689
   macro avg       0.87      0.88      0.87      6689
weighted avg       0.88      0.88      0.88      6689
```

Challenges

- **High computation time.**
- **Too many locations with irrelevant format.**
- **Sarcastic tweets.**

Conclusion

- For multiclass classification, the best model for this dataset would be CatBoost which gives highest accuracy 62%.
- For binary classification, the best model for this dataset would be Logistic Regression which gives accuracy 87%.

Q&A