Capstone Project-3

Coronavirus-Tweet-Sentiment-Analysis

Nakul Pradeep

Introduction

- Sentiment Analysis is a use case of Natural Language Processing (NLP) and comes under the category of text classification.
- Sentiment Analysis involves classifying a text into various sentiments, such as positive or negative, Happy, Sad or Neutral, etc. Thus, the ultimate goal of sentiment analysis is to decipher the underlying mood, emotion, or sentiment of a text.

Project Objective

- In Industry, sentiment analysis is used to find the product feedback and also can be used to reveal key information into whether the company is doing right or wrong. Companies may use sentiment analysis to assess the effectiveness of a new product, ad campaign, or other marketing initiatives.

- Here we use “Coronavirus Tweet data” set to assess the sentiment of the tweeter from his/her tweet.
- The aim of this project is to analyse the dataset and correctly deduce the correct sentiment for a given tweet.

Data Summary

- **Username :** gives the name of the user (coded)
- **Screenname :** gives the screen name of the user (coded)
- **Location**: gives information regarding the location from which tweet is sent
- **Tweet At**: gives the date of the Tweet
- **Original Tweet** : Tweet Content
- **Label**: shows the sentiment of the Tweet

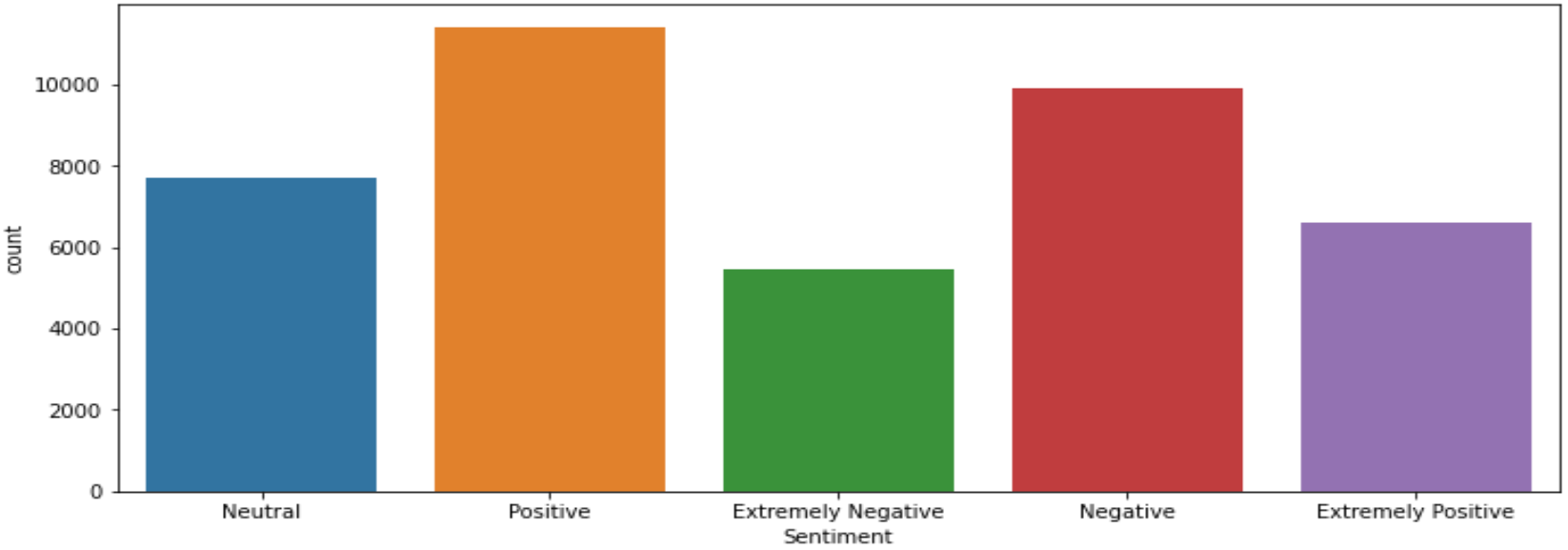
Steps of our Data Analysis

1. Problem Description and Data Description
2. Import the Libraries and Dataset
3. Null value Checking/Treatment
4. Conduct exploratory data analysis with Visualisation
5. Data transformation like cleaning of data, tokenisation, removal of stop words etc.
6. Splitting the dataset into training and test sets.
7. Models and Model Selection.
8. Evaluation of All the models and Comparison
9. Summaries and valuable insights

Exploratory Data Analysis

- The data has 41157 rows and 6 columns or features.
- The data set does not contain duplicate values.
- The dependent variable is 'Sentiment column' and it has 5 distinct features namely Neutral, Positive, Extremely Negative, Negative and Extremely Positive
- The independent variable is 'OriginalTweet' column and it has no null or duplicate values.
- There are null values present in the location column as there are only 32567 non null values for location with rows of 41157 values but there is no need for null value treatment as it is not part of the independent variable.
- The tweets in the dataset are from 16Th March 2020 to 14th April 2020. i.e. for a period of 30 days

Count plot of Sentiment Distribution of our data

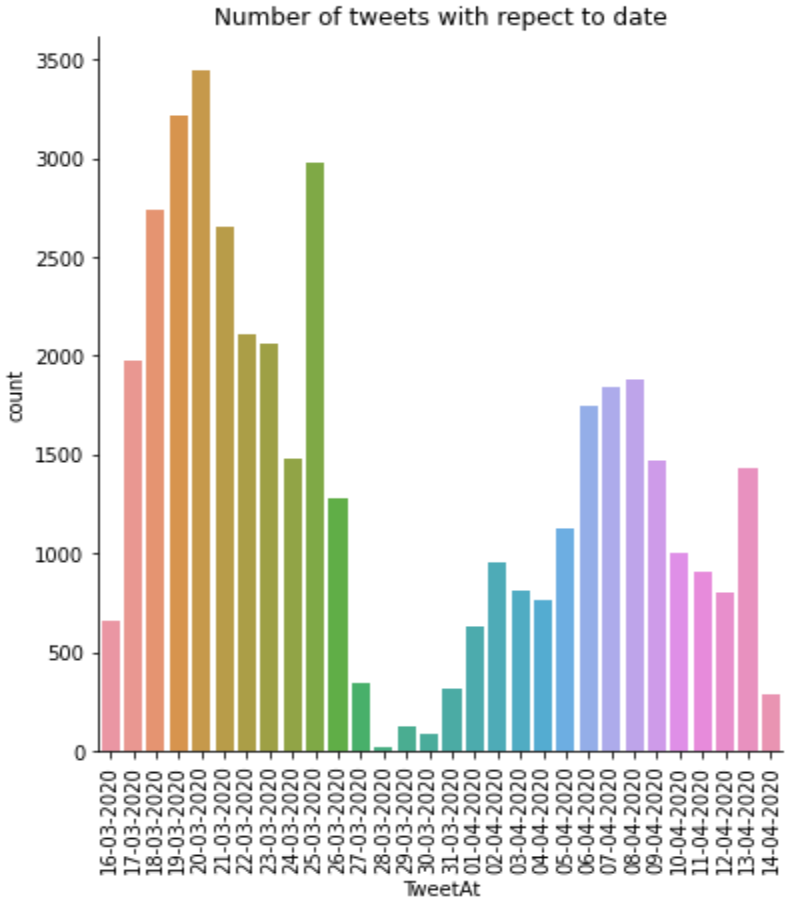


The majority sentiment of the tweeted data is positive and the minority sentiment is extremely negative.

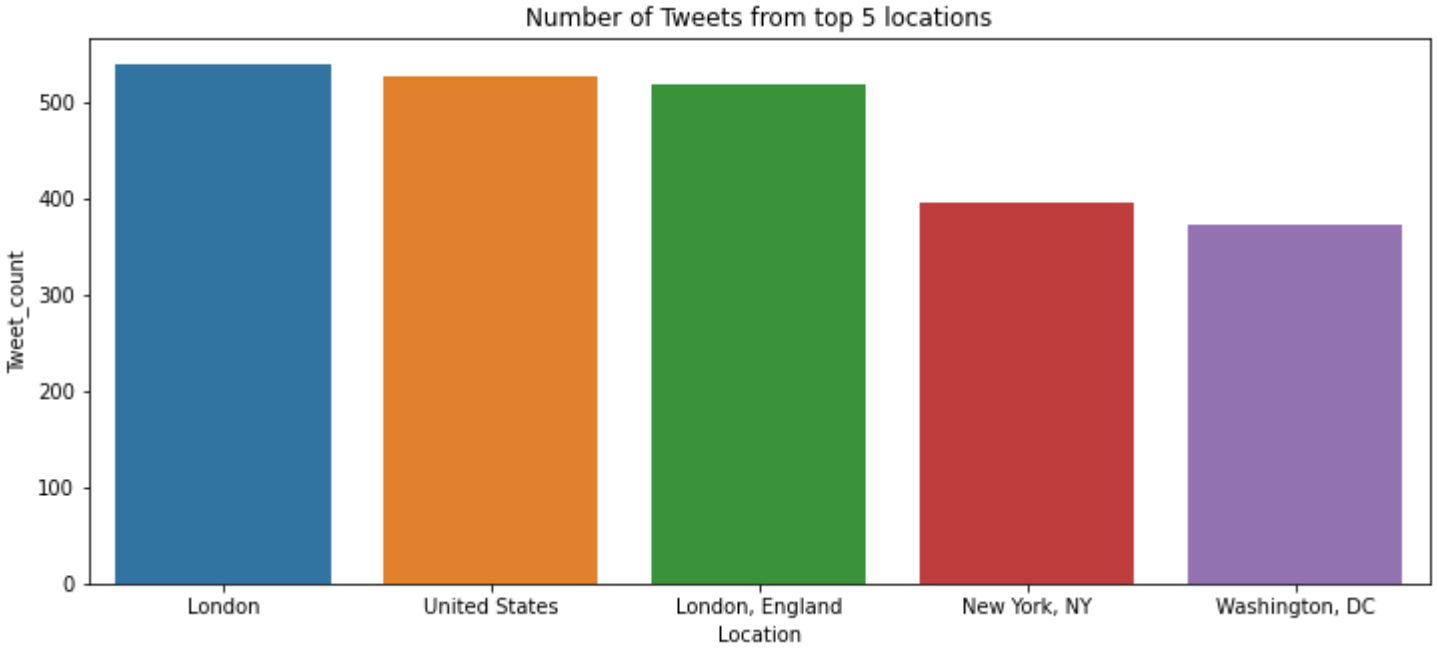
Distribution of the Dates of Tweet

- People tweeted most during 20-03-2020 and least during 28-03-2020.

- The 'Tweet At' column contains values from 16th March 2020 to 14th April 2020. i.e. for a period of 30 days



Count of Tweets from top 5 locations



All of top 5 tweets come from parts of UK and USA

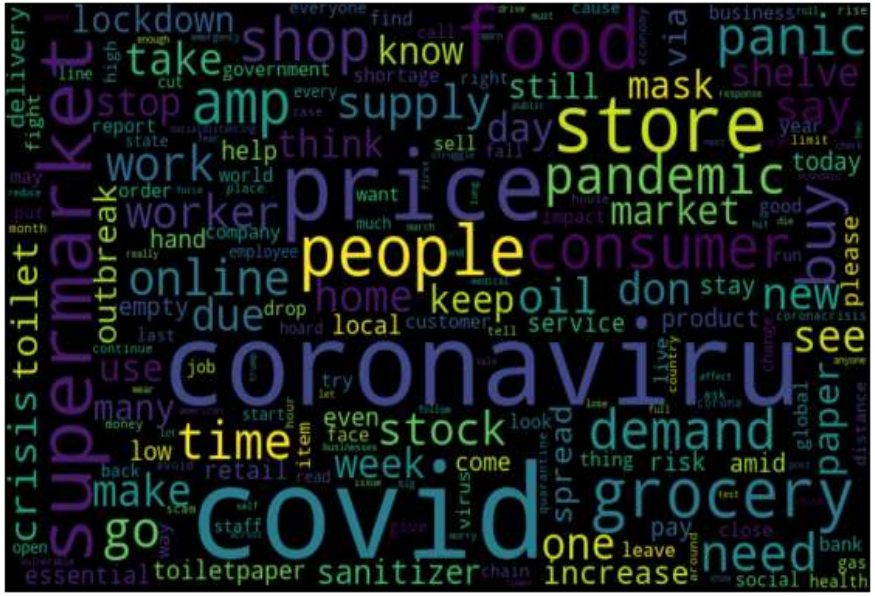
Word cloud

Word cloud for entire dataset

Word cloud for entire dataset with neutral sentiment

Word cloud

Word cloud for extremely negative sentiment



Word cloud for negative sentiment

Word cloud

Word cloud for positive sentiment

Word cloud for extremely positive sentiment

Tweet Cleaning steps

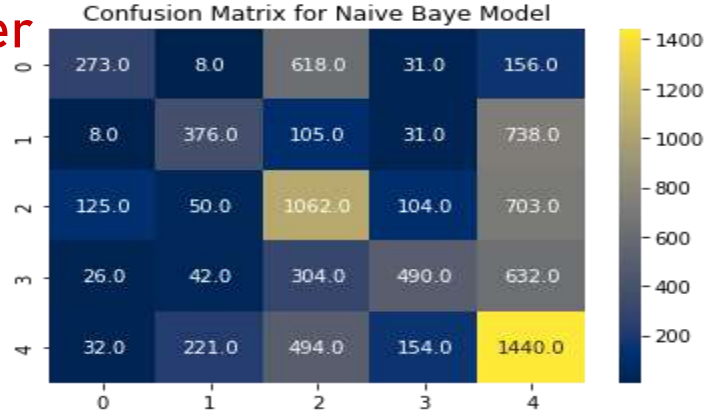
- Removing urls from dataset
- Removing values with @username attached
- Removing punctuations and special character
- Removing words with length less than or equal to 2.
- Removing stopwords
- Lemmatization
- Stemming

Pre-Processing data for fitting machine learning models

- Train test split
- Vectorization : It is the process of converting messages or text in to vectors that algorithms can work with
- The methods we used are Count Vectorizer and Tf-idf Vectorizer
- Count Vectorizer :It will transform the clean data to columns that represent a word or pair of words in the training dataset and the row count the frequency of each word/pair of words in each sentence.
- Tf-idf Vectorizer :This is a technique to quantify words in a set of documents. We generally compute a score for each word to signify its importance in the document

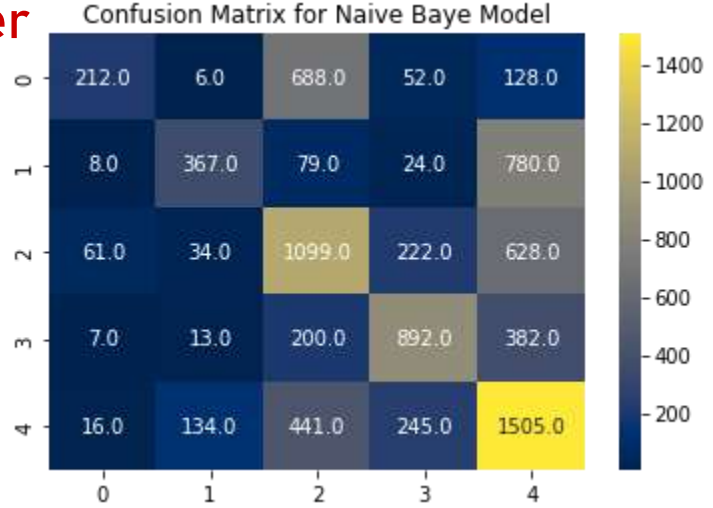
Naïve Bayes (Multinomial)- Tfidf vectorizer

- Accuracy:0.44
- Recall:0.44
- Precision:0.48
- F1-score:0.43



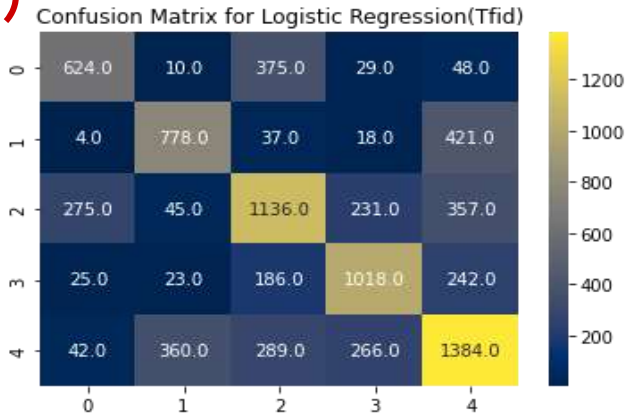
Naïve Bayes (Bernoulli's)- Tfidf vectorizer

Accuracy:0.495
Recall:0.495
Precision:0.54
F1-score:0.48



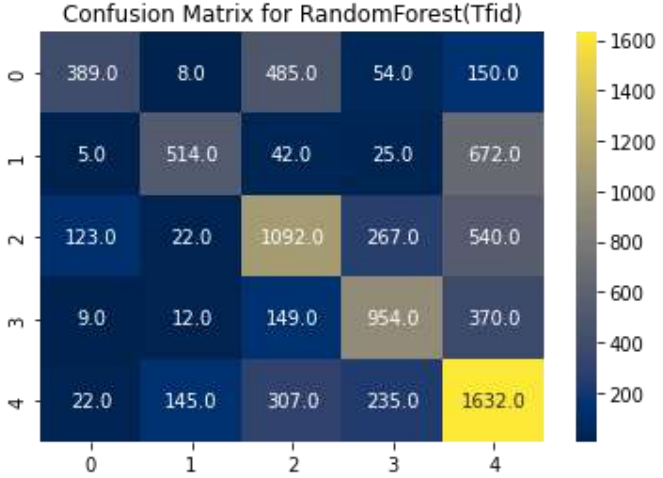
Logistic Regression-(Tf-idf Vectorizer)

- Accuracy:0.60
- Recall:0.60
- Precision:0.601
- F1-score:0.60



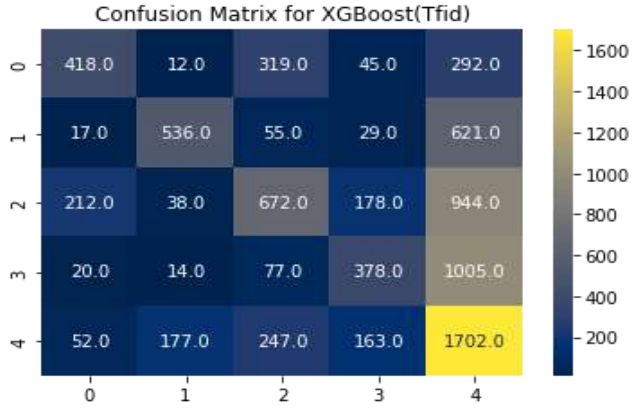
Random Forest(Tf-idf Vectorizer)

- Accuracy:0.557
- Recall:0.557
- Precision:0.587
- F1-score:0.552



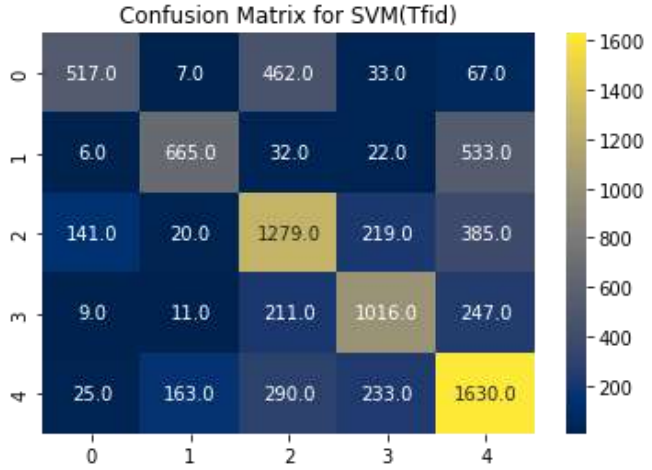
XG Boost-Tfidf Vectorizer

- Accuracy:0.45
- Recall:0.45
- Precision:0.497
- F1-score:0.44



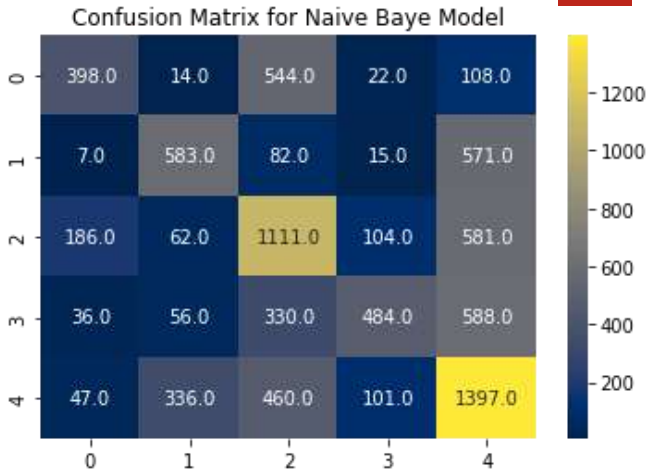
SVM-(Tfidf Vectorizer)

- Accuracy:0.621
- Recall:0.621
- Precision:0.638
- F1-score:0.620



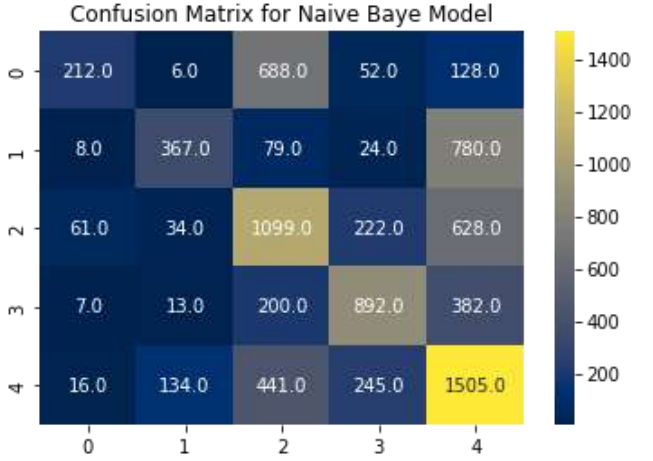
Naïve Bayes (Multinomial)- Count vectorizer

- Accuracy:0.48
- Recall:0.48
- Precision:0.515
- F1-score:0.479



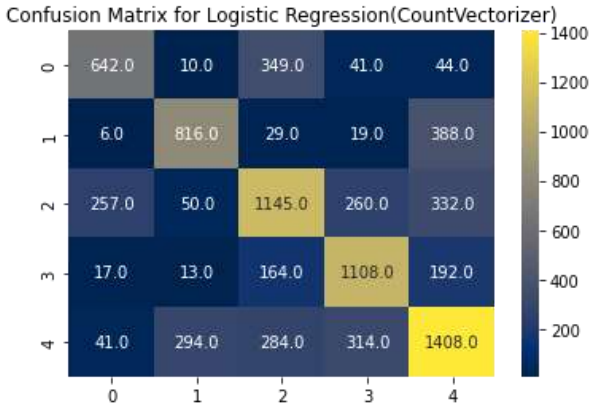
Naïve Bayes (Bernoulli's)- Count vectorizer

- Accuracy:0.495
- Recall:0.495
- Precision:0.54
- F1-score:0.48



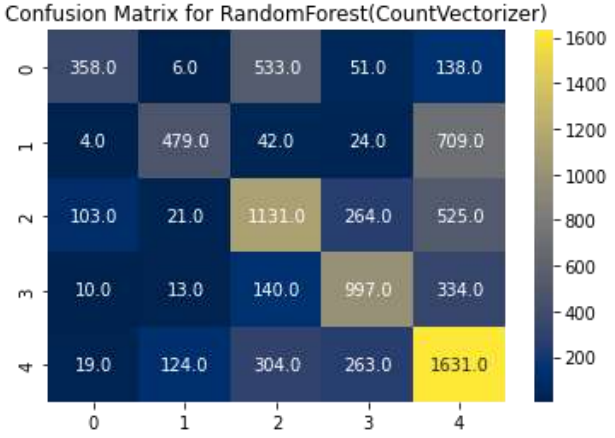
Logistic Regression-(Count Vectorizer)

- Accuracy:0.623
- Recall:0.623
- Precision:0.623
- F1-score:0.622



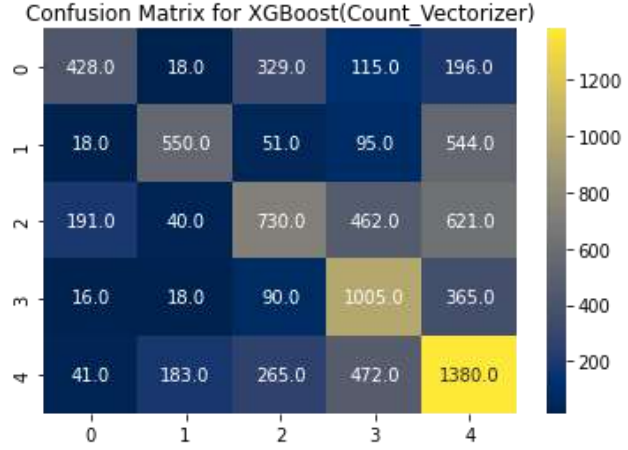
Random Forest(Count Vectorizer)

- Accuracy:0.559
- Recall:0.559
- Precision:0.593
- F1-score:0.552



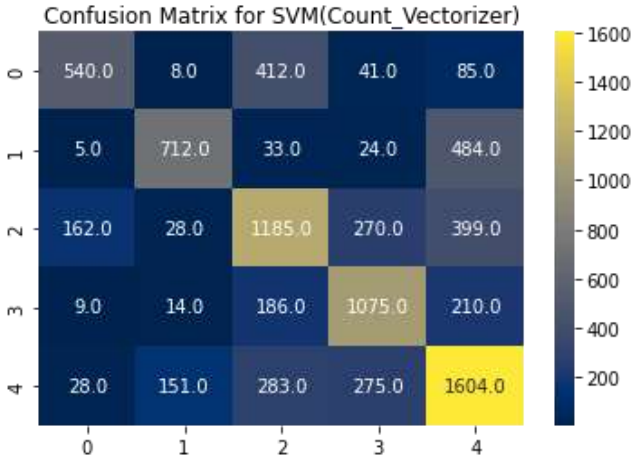
XG Boost-Count Vectorizer

- Accuracy:0.497
- Recall:0.497
- Precision:0.520
- F1-score:0.492



SVM-(Count Vectorizer)

- Accuracy:0.622
- Recall:0.622
- Precision:0.635
- F1-score:0.622



Data frame of 5 sentiment class prediction models with metrics

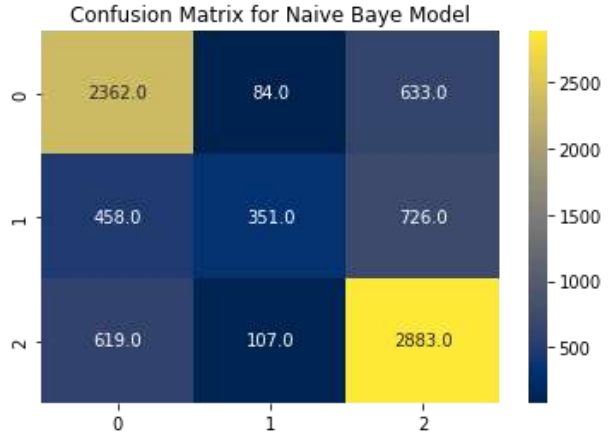
	MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE
0	Naive_Baye(Multinomial)-Tfid	0.442782	0.442782	0.484076	0.433177
1	Naive_Baye(Bernoulli's)-Tfid	0.495561	0.495561	0.540520	0.481637
2	Logistic_Regression-Tfid	0.600754	0.600754	0.601522	0.600702
3	Random Forest-Tfid	0.557096	0.557096	0.587769	0.552286
4	XG Boost-Tfid	0.450687	0.450687	0.497011	0.440018
5	SVM-Tfid	0.621063	0.621063	0.638449	0.620341
6	Naive_Baye(Multinomial)-CountVectorizer	0.483157	0.483157	0.515819	0.479435
7	Naive_Baye(Bernoulli's)-CountVectorizer	0.495561	0.495561	0.540520	0.481637
8	Logistic_Regression-CounntVectorizer	0.622522	0.622522	0.623094	0.621626
9	Random Forest-CountVectorizer	0.558920	0.558920	0.592864	0.551693
10	XG Boost-(Count_Vectorizer)	0.497750	0.497750	0.520772	0.492846
11	SVM-(Count_Vectorizer)	0.622157	0.622157	0.635547	0.621631

Multiclass with 3 labels

- The dependent variable Sentiment is converted to 3 sentiment class from 5 sentiment class to check if better accuracy can be obtained.
- The Extreme positive class is converted to positive sentiment class and the extreme negative class is converted in to negative class.
- The 3 labels for the dependent variable sentiment are positive , negative and neutral.

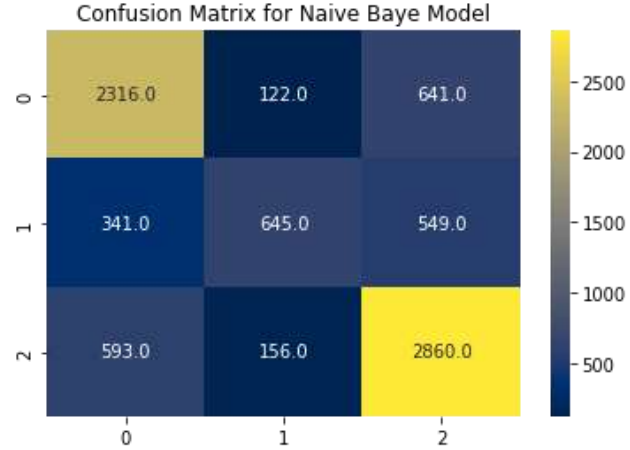
Naïve Bayes (Multinomial)- Count vectorizer

- Accuracy:0.68
- Recall:0.68
- Precision:0.676
- F1-score:0.656



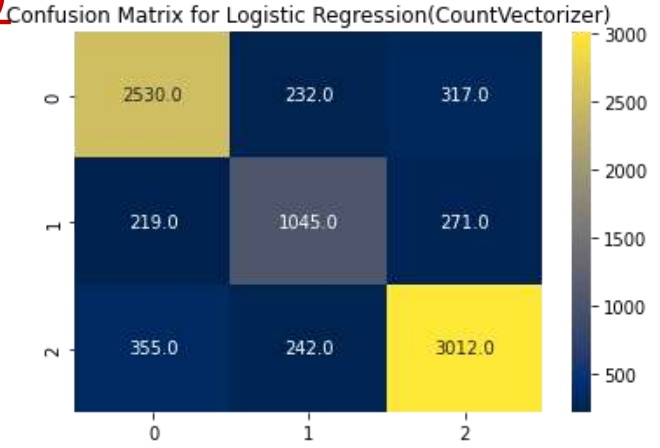
Naïve Bayes (Bernoulli's)- Count vectorizer

- Accuracy:0.708
- Recall:0.708
- Precision:0.707
- F1-score:0.6997



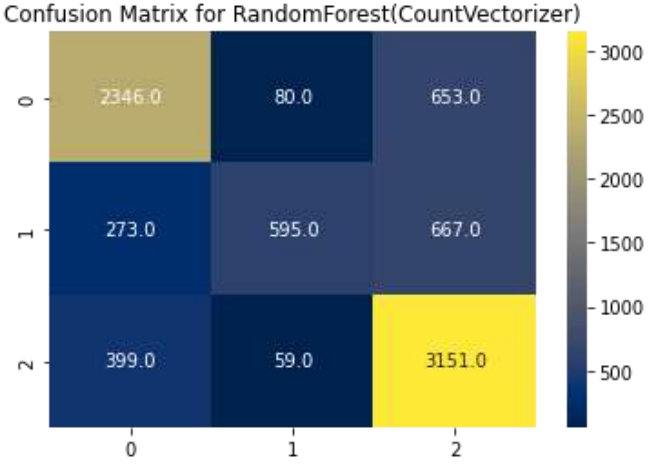
Logistic Regression-(Count Vectorizer)

- Accuracy:0.80
- Recall:0.80
- Precision:0.80
- F1-score:0.80



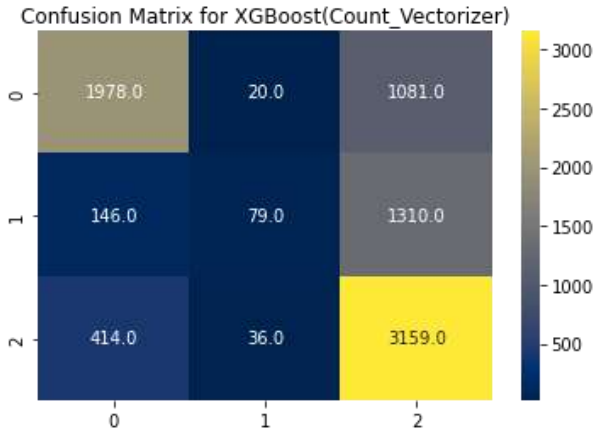
Random Forest(Count Vectorizer)

- Accuracy:0.74
- Recall:0.74
- Precision:0.74
- F1-score:0.728



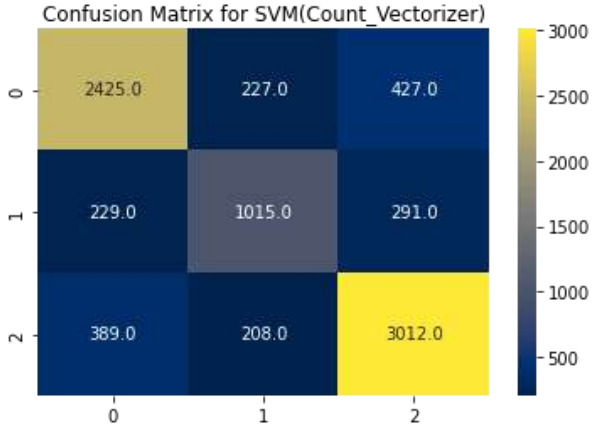
XG Boost-Count Vectorizer

- Accuracy:0.634
- Recall:0.634
- Precision:0.651
- F1-score:0.584



SVM-(Count Vectorizer)

- Accuracy:0.785
- Recall:0.785
- Precision:0.783
- F1-score:0.784



Data frame of 3 sentiment class prediction models with metrics

	MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE
0	Naive_Baye(Multinomial)-CountVectorizer	0.680530	0.680530	0.676347	0.656805
1	Naive_Baye(Bernoulli's)-CountVectorizer	0.707892	0.707892	0.707210	0.699786
2	Logistic_Regression-CounntVectorizer	0.801046	0.801046	0.800822	0.800925
3	Random Forest-CountVectorizer	0.740849	0.740849	0.751699	0.728367
4	XG Boost-(Count_Vectorizer)	0.634318	0.634318	0.650869	0.584127
5	SVM-(Count_Vectorizer)	0.784628	0.784628	0.783471	0.783838

Conclusion

- Logistic Regression and SVM gives the best accuracy, recall, precision and F1 score compared to all other models in multiclass (5 sentiment) dataset. Both of them gives a 0.62 F1 score.

- Logistic Regression gives a best accuracy, recall, precision and F1 score compared to all other models in multiclass (3 sentiment) dataset. It gives a 0.80 F1 score.

THANK YOU