Capstone Project:03

Supervised Machine Learning Classification

Sentiment Analysis : Predicting Sentiment of COVID-19 tweets

Author: Pankaj Beldar
Email id: pankajrbell@gmail.com

Abstract- The coronavirus disease 2019 (COVID-19) pandemic has influenced the everyday life of people around the globe. In general and during lockdown phases, people worldwide use social media networks to state their viewpoints and general feelings concerning the pandemic that has hampered their daily lives. Twitter is one of the most commonly used social media platforms, and it showed a massive increase in tweets related to coronavirus, including positive, negative, and neutral tweets, in a minimal period. The researchers move toward sentiment analysis and analyze the various emotions of the public toward COVID-19 due to the diverse nature of tweets. Meanwhile, people have expressed their feelings regarding the vaccinations' safety and effectiveness on social networking sites such as Twitter. As an advanced step, in this paper, our proposed approach analyzes COVID-19 by focusing on Twitter users who share their opinions on this social media networking site. The proposed approach analyzes collected tweets' sentiments for sentiment classification using various feature sets and classifiers.

1. Introduction

Coronavirus disease 2019 (COVID-19) has severely impacted the daily lives of individuals across the globe. People worldwide use online media to state their viewpoints and general feelings concerning this phenomenon that has assumed control over the world by storm .Social media platforms like Twitter have experienced exponential growth in tweets related to the pandemic in a short period . The social networking site Twitter is a commonly used online media platform. It provides real-time information related to ongoing events concisely and captures the emotions and thoughts of the people. During this pandemic, people use the online media platform Twitter to express their feelings, opinions, emotions, and thoughts related to the worldwide pandemic .The social media users are increasing with time because they depend on social media for informative content, and the volume of data is also increasing; this focused on the use of Natural Language Processing (NLP) with different algorithms of

Artificial intelligence (AI) to extract meaningful information efficiently . NLP and its applications have had a significant impact on social media text analysis and classification; however, the challenges of determining a content's inherent importance using NLP-strategies, such as contextual phrases and words, ambiguity in text or speech, necessitate the use of ML-based algorithms. In this study, we use Twitter data for sentiment analysis to identify public sentiments to investigate the increased fear associated with coronavirus. Many traditional approaches have been used to identify human behavior and nature, which presents the possibility of increasing analyses by quickly doing sentiment classification using NLP techniques. Sentiment analysis and classification of COVID-19 and other disaster-associated scenarios and keywords associated with the Twitter data analysis.

2. Problem Description

This challenge asks you 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. You are given the following information:

1. use_name
2. screen name
3. Location
4. Tweet At
5. Original Tweet
6. Sentiment

Fields	Description
Username	Coded Username
ScreenName	Coded ScreenName
Location	Region of origin
TweetAt	Tweet Timing
OriginalTweet	First tweet in the thread
Sentiment-Target variable	Sentiment of the tweet

3. Natural Language Processing

The Natural Language API breaks up text into its constituent words and punctuation (called tokens) and then provides information on each part. You can use the API to

perform the following tasks on a chunk of text:

- **Syntax analysis-** identify parts of speech
- **Entity recognition-** label entities by type (person, location, event, etc.)
- **Sentiment analysis-** get the overall sentiment of a block of text
- **Content classification-** classify documents into predefined categories Upon reviewing the Natural Language API docs, I became most interested in sentiment analysis. Let's delve deeper into how this feature works

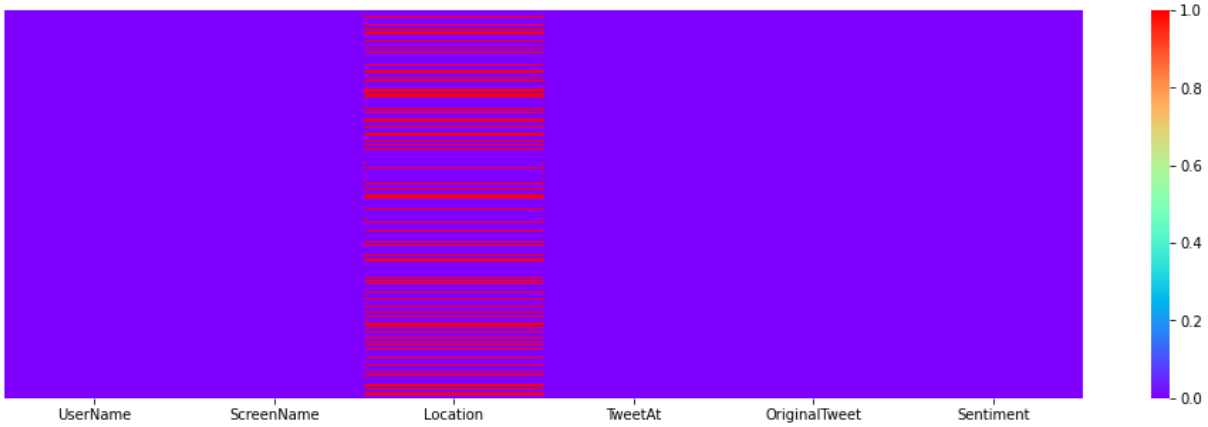
4. Sentiment Analysis

The sentiment score is a numerical interpretation of the overall emotional leaning of the text. Score values range from negative sentiment to positive sentiment.Text analytics is the process of synthesizing unstructured data to help discover patterns and enable decision making. Until recent years text analytics had to be performed the old fashioned way i.e. eyeballing and manual categorisation of text, which is inefficient and time consuming. Also this is not a practical solution when dealing with millions of documents such as twitter data.Twitter data (also known as tweets) is a rich source of information on a large set of topics. This data can be used to find trends related to a specific keyword, measure brand sentiment or gather feedback about new products and services. This post will provide a step by step guide for text analytics on twitter data.

The sentiment is a useful indicator, The data when extracted may lead to very useful insights on your product or company. So in the next step we cover the usage of named entity recognition algorithms, which is designed to extract this information.

5. DATA Preprocessing

Dataset consists of 41157 entries and 6 attributes. There are no Missing values in all columns except 'Location'. The attribute 'Location' has 8590 missing values. As we want to do sentiment analysis, the 'Location' column is not so important.



Null values are present in 'Location'. Almost 8590 null values are present in the "location" variable.

In data cleaning we have done following things-

1. Removed @users from tweets
2. Removed Punctuations
3. Removed Stopwords
4. Removed HTML tags
5. Removed URLs
6. Removed Short words
7. Removed Emojies
8. Convert text to Lower Case
9. Stemming
10. Tokenizing
11. Spelling Correction

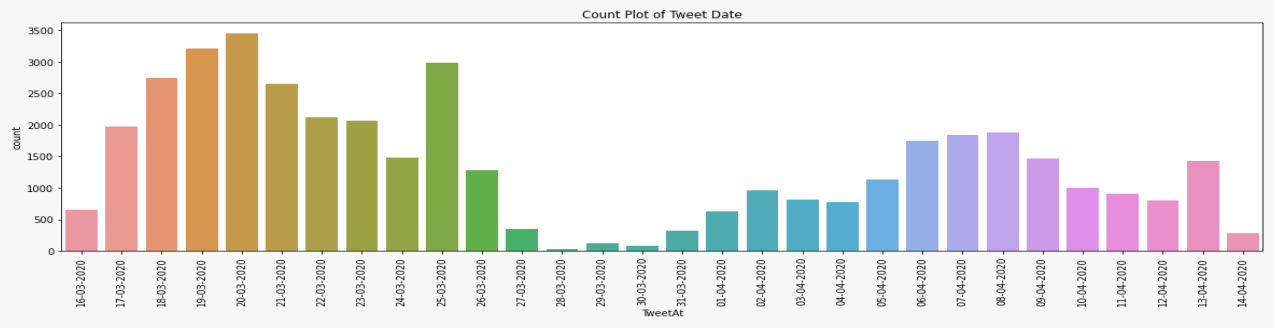
6. Exploratory data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations

We have try to figure out the following question through Exploratory Data Analysis

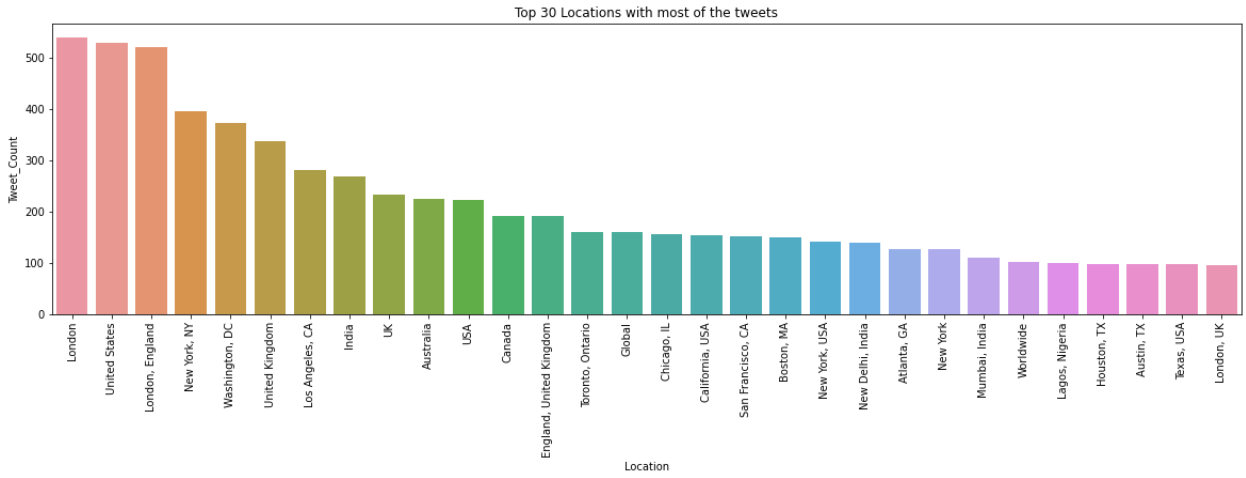
1. How many Tweets per day for the given time period?
2. Which are the Top 30 Locations from where maximum tweets were done?
3. What are the Sentiment types and its Distribution?
4. What is the relationship between the number of words, characters, sentences and different types of Sentiments?
5. What are the most frequent Extremely Positive Sentiment Words?
6. What are the most frequent Positive Sentiment Words?
7. What are the most frequent Extremely Negative Sentiment Words?
8. What are the most frequent Negative Sentiment Words?
9. What are the most frequent Neutral Sentiment Words?
10. Finding out top 10 most frequent words based on sentiment.

- **How many Tweets per day for the given time period?**



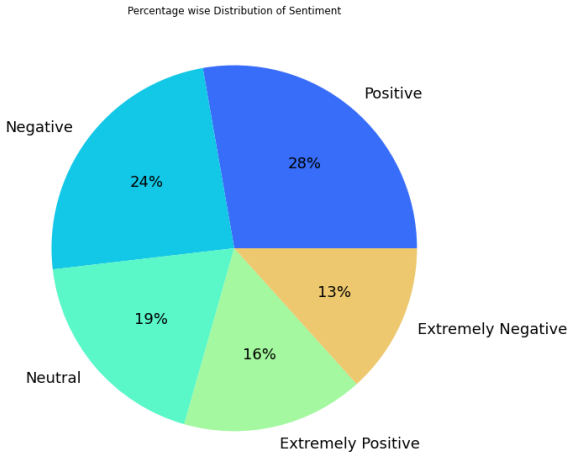
We can see that maximum tweets were done on 20 march 2020, when the first lockdown was declared. People were more active on Twitter in the month of March because it was the early stage of Corona virus Pandemics and people wanted to know more about this disease.

- **Which are the Top 30 Locations from where maximum tweets were done?**



We can see London, United States, New York, Washington DC, United Kingdom ,India, Australia, USA are the top locations as far as count of tweets are concerned.

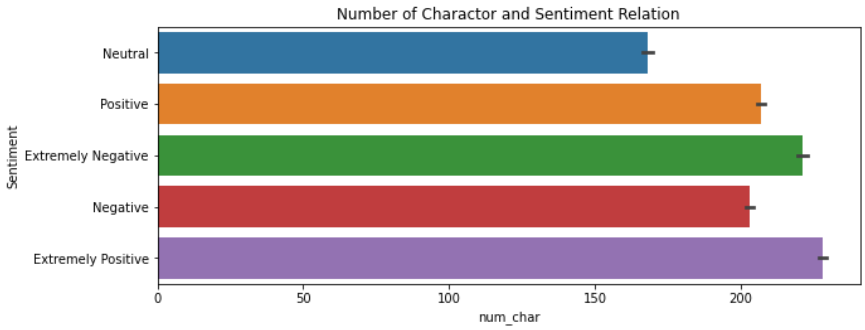
- **What are the Sentiment types and its Distribution?**

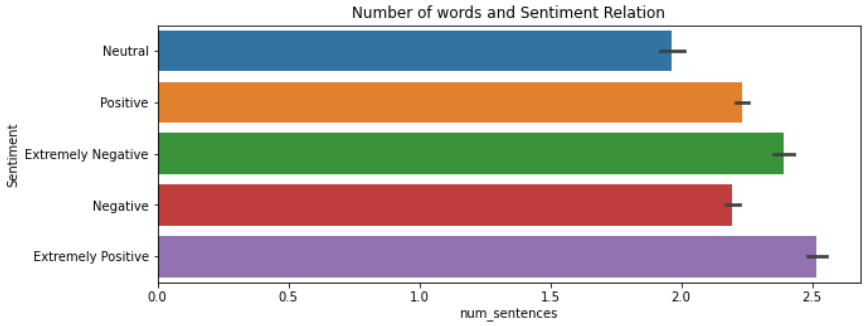


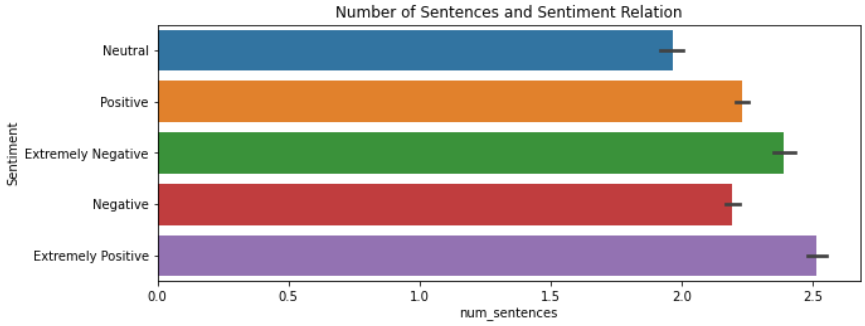
Most of the people (28%) were having Positive sentiment about covid followed by Negative (24%), Neutral(19%), Extremely Positive (16%) and Extremely Negative(13%).

Our Target variable is not Unbalanced as all Categories are not having much differences between them.

- **What is the relationship between the number of words, characters, sentences and different types of Sentiments?**







We can see that if the number of words, characters and sentences are more than sentiment of the tweet is more positive.

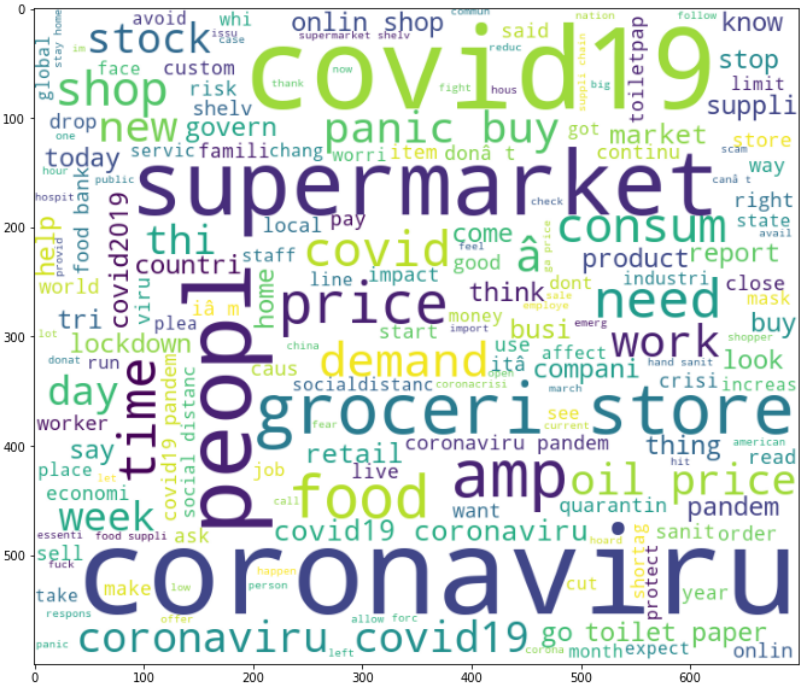
Neutral Sentiment Tweets consist of fewer words, characters and sentences.

- **What are the most frequent Extremely Positive Sentiment Words?**

- **What are the most frequent Positive Sentiment Words?**

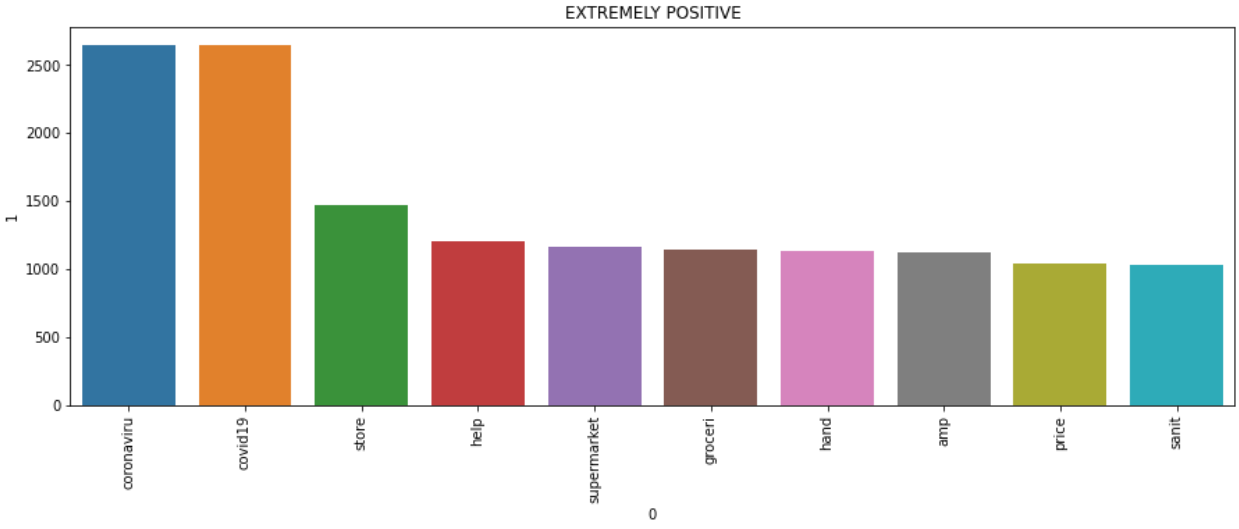
- **What are the most frequent Extremely Negative Sentiment Words?**

- **What are the most frequent Negative Sentiment Words?**

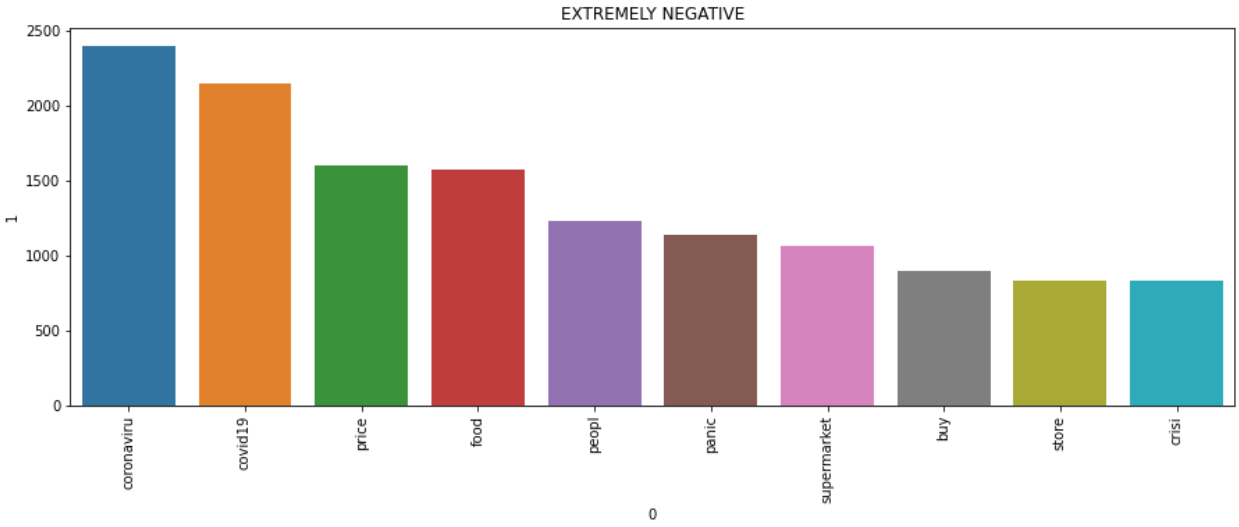


- **What are the most frequent Neutral Sentiment Words?**

- **Finding out Most frequent top 10 words based on the Sentiment**



Number of words in Extremely Positive =127343



Number of words in Extremely Negative =105038

Number of words in Positive =196220

Number of words in Negative =168977

Number of words in Neutral =102356

7. Model Development

- **Feature selection**

Feature selection is the process of selecting what we think is worthwhile in our documents, and what can be ignored. This will likely include removing punctuation and stopwords, modifying words by making them lower case, choosing what to do with typos or grammar features, and choosing whether to do stemming.

Feature selection is the process of selecting a subset of the terms occurring in the training set and using only these subset features in text classification. Feature selection serves two main purposes. First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary. This is of particular importance for classifiers that, unlike NB, are expensive to train. Second, feature selection often increases classification accuracy by eliminating noise features. A *noise feature* is one that, when added to the document representation, increases the classification error on new data.

1. As we want the sentiment analysis of the tweet, hence 'UserName', 'ScreenName', 'Location', 'TweetAt' are not playing an important role for the determination of sentiment of the tweet.

2. Hence we select important features as 'Sentiment', 'num_char', 'num_words', 'num_sentences', 'Tweet' for our analysis.

- **Train Test Split**

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications. This method is a fast and easy procedure to perform such that we can compare our own machine learning model results to machine results. 80 % training- 20% testing is taken.

train shape : (32925, 5)
test shape : (8232, 5)

- **Feature Extraction**

Word Vectorization- Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which is used to find word predictions, word similarities/semantics.

TfidfVectorizer - Transforms text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index.

- **Classification Model Selection**

➔ **Naive Bayes classifier for multinomial models**

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts.

training accuracy Score : 0.6468033409263477

Validation accuracy Score : 0.4870019436345967

➔ **SGDClassifier**

The class SGDClassifier implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties for classification. Below is the

decision boundary of a SGDClassifier trained with the hinge loss, equivalent to a linear SVM.

Training accuracy Score : 0.774791192103265

Validation accuracy Score : 0.5640184645286687

➔ RandomForestClassifier

The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

Training accuracy Score : 0.9995140470766894

Validation accuracy Score : 0.5788386783284742

➔ XGBoost classifier

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Training accuracy Score : 0.4946089597570235

Validation accuracy Score : 0.47351797862001943

➔ Support vector machine classifier

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text

Training accuracy Score : 0.8969779802581624

Validation accuracy Score : 0.6076287657920311

➔ CatBoost classifier

CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters.

Training accuracy Score : 0.6689445709946849

Validation accuracy Score : 0.619290573372206

Table: Multiclass Classification Comparative Results

	Model	Test accuracy
5	CatBoost	0.619291
0	Support Vector Machines	0.607629
1	Random Forest	0.578839
3	Stochastic Gradient Decent	0.564018
2	Naive Bayes	0.487002
4	XGBoost	0.473518

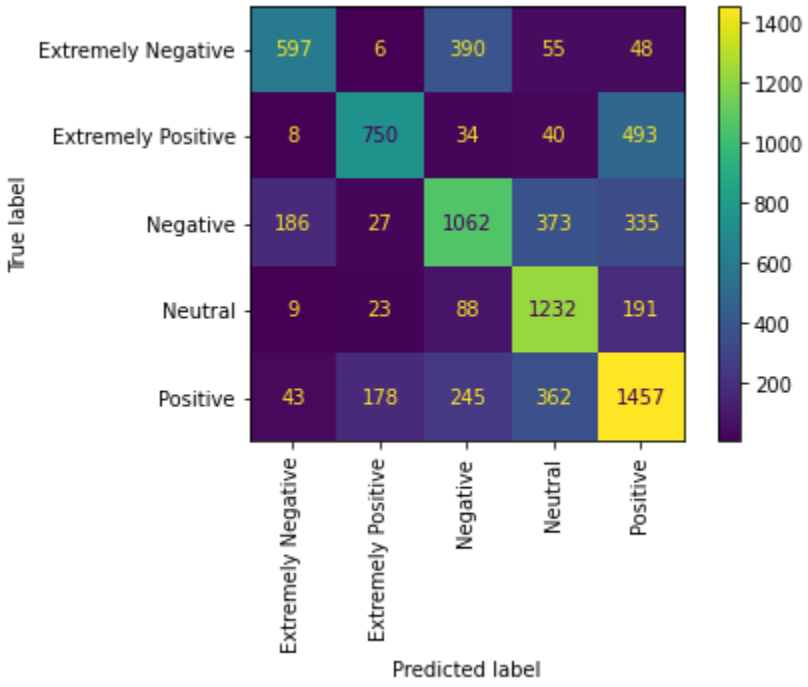
- Best Multiclass Classification model- CatBoost Classifier

```
Training accuracy Score    :  0.6689445709946849
Validation accuracy Score :  0.619290573372206
                   precision    recall  f1-score   support

Extremely Negative       0.54      0.71      0.62       843
Extremely Positive       0.57      0.76      0.65       984
          Negative       0.54      0.58      0.56      1819
           Neutral       0.80      0.60      0.68      2062
          Positive       0.64      0.58      0.61      2524

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

- Confusion Matrix of Multiclass CatBoost Classifier



- **CONVERTING OUR MULTICLASS CLASSIFICATION INTO BINARY CLASSIFICATION**

Let us create binary class as Positive, Extremely Positive, Neutral as (positive) and Negative and Extremely Negative Sentiment as (negative)

➔ Voting Classifier

A voting classifier is a machine learning estimator that trains various base models or estimators and predicts on the basis of aggregating the findings of each base estimator. The aggregating criteria can be combined with a decision of voting for each estimator output.

Training accuracy Score : 0.9354290053151101

Validation accuracy Score : 0.863824101068999

➔ Stacking Classifier Algorithms

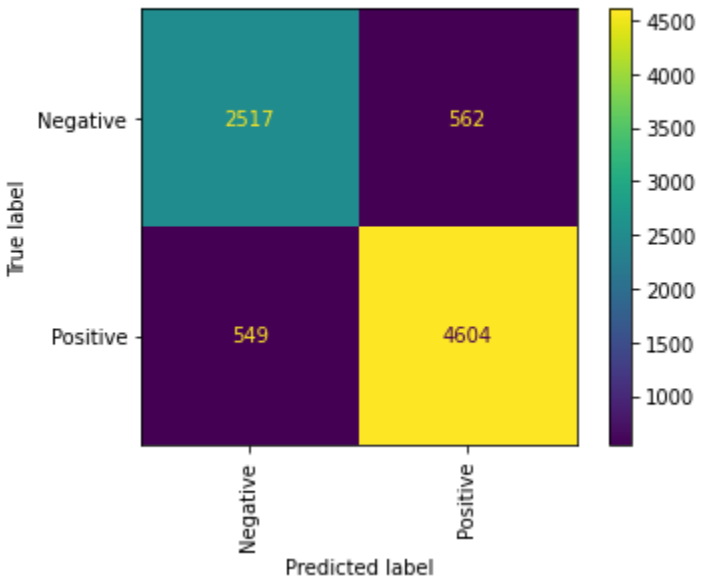
Stacking is a way of ensembling classification or regression models; it consists of two-layer estimators. The first layer consists of all the baseline models that are used to predict the outputs on the test datasets

Training accuracy Score : 0.994290053151101
Validation accuracy Score : 0.8650388726919339

classification_report

	precision	recall	f1-score	support
Negative	0.82	0.82	0.82	3066
Positive	0.89	0.89	0.89	5166
accuracy			0.87	8232
macro avg	0.86	0.86	0.86	8232
weighted avg	0.87	0.87	0.87	8232

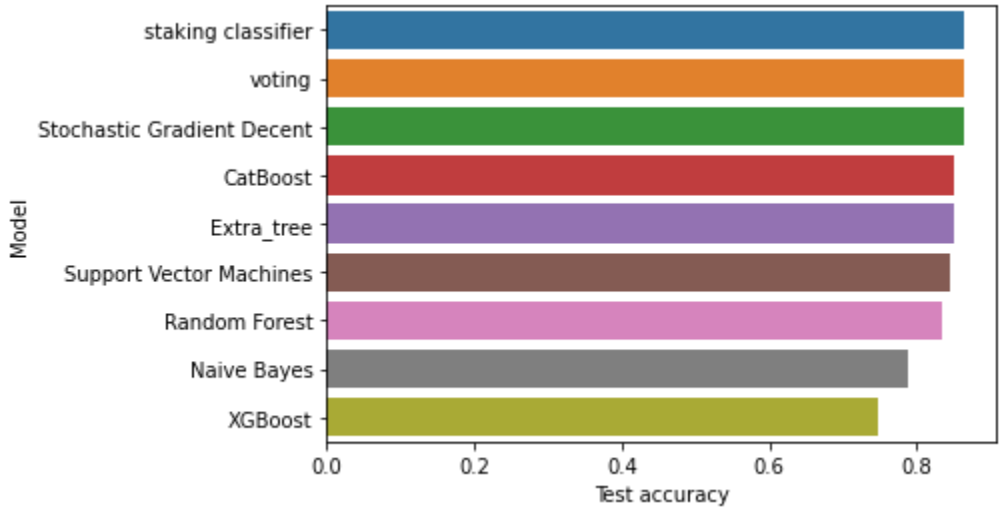
confusion_matrix



8. Result

Table: Binary Classification Comparative Results

	Model	Test accuracy
8	staking classifier	0.865039
7	voting	0.863824
3	Stochastic Gradient Decent	0.863460
5	CatBoost	0.851919
6	Extra_tree	0.849854
0	Support Vector Machines	0.844752
1	Random Forest	0.834913
2	Naive Bayes	0.788630
4	XGBoost	0.746842



Stacking Classifier is the best classification model for our problem with a 0.86 accuracy score.

9. Conclusion

1. We have tried Stacking classifier, Stochastic Gradient Descent Classifier, voting, CatBoost, Extra Tree, Support Vector Classifier, Random Forest Classifier, Multinomial Naive Bayes and XGboost classifier.
2. We have Extracted new Features as number of words, characters and sentences to check sentiment of the Tweet.We can see that if the number of words, characters and sentences are more than sentiment of the tweet is more positive. Neutral Sentiment Tweets consist of fewer words, characters and sentences. After evaluating models with these features we have found that these features are not contributing much in the model performance. Hence we have decided to drop them in Binary classification models
3. Firstly we evaluate the multi class models as categories -Positive, Extremely Positive, Neutral, Negative, Extremely Negative. Most of the people (28%) were having Positive sentiment about covid followed by Negative (24%), Neutral(19%), Extremely Positive (16%) and Extremely Negative(13%). Our Target variable is not Unbalanced as all Categories are not having much differences between them.
4. The performance of the Multiclass Models is not satisfactory then we convert the problem into binary class. We have kept only two Sentiments as Positive and Negative.
5. The Binary Stacking Classifier has the best performance of accuray 0.8650 followed by CatBoost and Extra Tree classifier.
6. We have checked for Voting classifiers and Stacking Classifiers with hyperparameter tuning.
7. For vectorizing we have tried all kinds of vectorizing methods i.e. tfidfvectorizer ,countvectorizer, ngrams, after doing hyper parameter tuning, we have decided to use countvectorizer for vectorization.
8. Sentiment Analysis is done based on Positive and Negative Sentiment.
9. Feature Importance is calculated based on Most Frequent words in each class.
10. We can see London, United States, New York, Washington DC, United Kingdom ,India, Australia, USA are the top locations as far as count of tweets are concerned.
11. We can see that maximum tweets were done on 20 march 2020, when the first lockdown was declared. People were more active on Twitter in the month of March because it was the early stage of Corona virus Pandemics and people wanted to know more about this disease.

10. Challenges Faced

1. As Data is related to natural language processing sentiment analysis, Time required for data pre processing is high.
2. Feature Extraction and Computation Cost is high.
3. There is a wide scope for Vectorization techniques.
4. Algorithms like SVC, XGBoost, CatBoost have high computation cost.
5. Time required to run Stacking and Voting Algorithm more than other algorithms

11. Scope

As we are dealing with sentiment analysis of coronavirus tweets, It's very important to classify sentiment as either positive or negative to use it as reference for different stakeholders. Governments can make use of this information in policymaking as they are able to know how people are reacting to this new strain, what all challenges they are facing such as food scarcity, panic attacks, etc. Various profit organizations can make a profit by analyzing various sentiments as one of the tweets tells us about the scarcity of masks and toilet papers. These organizations are able to start the production of essential items thereby making profits. Various NGOs can decide their strategy of how to rehabilitate people by using pertinent facts and information.We could do the analysis with three classes as positive, negative and neutral.

Author- Pankaj R. Beldar

Email id- pankajrbell@gmail.com

Thank You.....!