ML – Classification

CORONAVIRUS TWEET SENTIMENT ANALYSIS

Almabetter, Bengaluru

Team Members	
Varsha Rani	**Vivek Chandrakant Pawar**
Rabista Parween	**Tushar Gaikwad**

1. Problem Statement:

This challenge asks us 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.

2. Introduction:

We are given the following information:

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

3. Objective:

The objective of this project is to build a classification model to predict the sentiment of COVID-19 tweets.

4. Steps Involved:

Importing Libraries and Data Inspection

We have used the following Libraries:

- Pandas – Manipulation of tabular data in Dataframes
- Numpy – Mathematical operations on arrays
- Matplotlib – Visualization
- Seaborn – Visualization
- Sklearn – Data Modeling
- Nltk – Pre Processing / Feature Engineering
- WordCloud – Visualization

We have done data inspection. The original Dataset contains 6 columns and 41157 rows.

Feature Engineering

Step 1 : Converted all characters to lowercase.

Step 2 : Removed Punctuation.

Step 3 : Removed stop words.

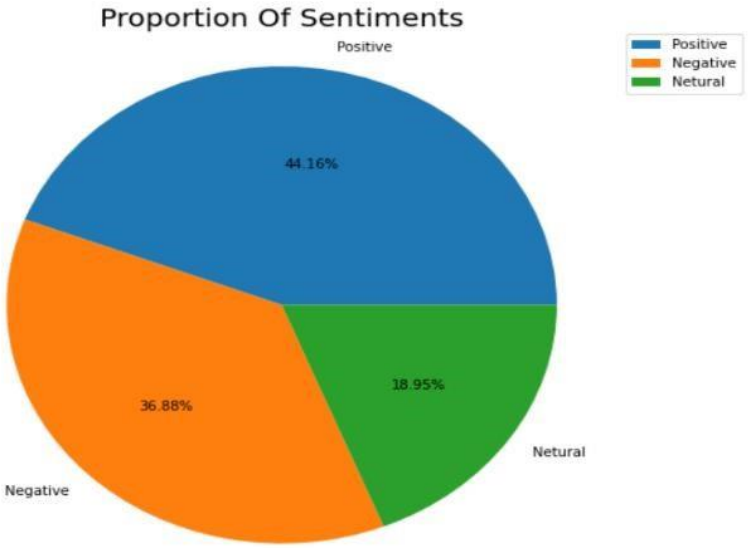
After dropping the null values and adding a new column ‘clean_tweets’, now we have 7 columns and 32567 rows.

Exploratory Data Analysis

- Proportion of Sentiments
- Proportion of new Sentiments

  As there was five types of sentiments – Positive Sentiment, Extremely Positive Sentiment, Negative Sentiment, Extremely Negative Sentiment and Neutral Sentiment. So, we have replaced Extremely Positive Sentiment by Positive Sentiment and Extremely Negative Sentiment by Negative Sentiment. Now we have three types of sentiments – Positive Sentiment, Negative Sentiment and Neutral Sentiment.

  Positive Sentiments are having higher proportion among all.



- Unique values in each column
- Original Tweet according to TweetAt column
- Histogram plot of original tweet
- Top 10 Locations
- Most common words in tweets
- Word Cloud

Model Training

- Linear Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier

Performance Metrics and Accuracy

Performance of Logistic Regression Model

Accuracy : 0.7710636207320068

Precision : 0.7857896154337569

Recall : 0.7710636207320068

Performance of Decision Tree Classifier

Accuracy : 0.6129943502824858

Precision : 0.6109013482361811

Recall : 0.6129943502824858

Performance of Random Forest Classifier

Accuracy : 0.7248833210513388

Precision : 0.7282331957151178

Recall : 0.7248833210513388

Performance of Gradient Boosting Classifier

Accuracy : 0.6545074920167036

Precision : 0.6962747491301503

Recall : 0.6545074920167036

Confusion Matrix

A **Confusion matrix** is an N x N matrix **used for** evaluating the performance of a classification model, where N is the number of target classes.

Conclusion

Conclusion on EDA:

- Original Dataset contains 6 columns and 41157 rows.
- Location column contains null values. So, we have dropped the null values.
- And we added a new column "clean_tweets" after cleaning the tweets.
- After dropping and adding a new column, now we have 7 columns and 32567 rows.
- In order to analyze the data we required only two columns "OriginalTweet" and "Sentiment".
- The columns such as "UserName" and "ScreenName" does not give any meaningful insights for our analysis.
- There are five types of sentiments - Extremely Positive, Positive, Extremely Negative, Negative and Neutral.
- We have renamed the Extremely Positive and Extremely Negative sentiments to Positive and Negative respectively. And we are left with three types of sentiments - Positive, Negative and Neutral.
- The pie chart shows the proportion of sentiments.
- Bar plot for unique values shows us the number of unique values in each column.
- The graphical representation of top 10 locations shows us that most of the tweets came from London followed by United States.

Conclusion on Model Training:

- At the end we conclude our classification project with four models namely

  - Logistic Regression Model, Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier.

- We are getting the highest accuracy of about 77% with Logistic Regression.