Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Team Members	Email_id
Vivek Pawar	Vivek16pawar@gmail.com
Varsha Rani	e.varsharani@gmail.com
Tushar Gaikwad	gaikwadtushar140@gmail.com
Rabista Parween	2001rabista@gmail.com

Contribution	Responsibilities
Coding	Varsha Rani And Vivek Pawar
Presentation	Varsha Rani And Rabista Parween
Technical Documentation	Varsha Rani
Summary	Rabista Parween
Presentation Video	Individual

Git Hub Repository and Google Drive Link

Github Link:- https://github.com/2001rabista/Classification_ML_Capstone

Summary

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.

Objective:

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

Steps Involved:

- Libraries Importing
- Data Loading And Data Inspection
- Data Manipulation
- EDA
- Feature Engineering
    1. Lower Cap Text
    2. Remove Punctuation
    3. Remove Stop words
- Training Model
    1. Logistic Regression
    2. Decision Tree Classifier
    3. Random Forest Classifer
    4. Gradient Boosting
- Performance Metrics and Accuracy
- Confusion Matrix

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 Data Modeling:

- At the end we conclude our classification project with five 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.