Capstone Project Submission

Team Member's Name, Email and Contribution:

Name – Baishnavee Mahato
Email. mbaishnavee@gmail.com

Data Understanding

- Data Analysis
- Data preprocessing

Data Visualization

- Bar plot
- Count plot
- Line plot
- WordCloud

Data Preprocessing

Classification Algorithm Implementation

- Logistic Regression
- Random Forest Classifier
- Catboost model
- Naive Bayes Classifier

Conclusion

Technical Documentation

GitHub Repo link.

GitHub Link:
https://github.com/mbaishnavee05/Coronavirus_tweet_sentiment_analysis

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Problem Statement

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the
cause of an outbreak of respiratory illness first detected in Wuhan, China. The virus was declared a
pandemic by World Health Organization on 11th March 2020. Some people lost their lives, but
many of us successfully defeated this new strain i.e. Covid-19.
Sentiment Analysis is the process of computationally identifying and categorizing opinions
expressed in a piece of text, especially in order to determine whether the writer's attitude towards a
particular topic is Positive, Negative, or Neutral.
In this project, we are going to predict the Sentiments of COVID-19 tweets. The data gathered from
the Tweeter and I'm going to use Python environment to implement this project.

Approach

In first step, imported the data set to carry out the descriptive analysis over the data set to understand the information of data available.

Checked for missing values in the data set provided.

Performed EDA and Data Preprocessing.

Applied different ML Algorithm:
Logistic Regression
Random Forest Classifier
Catboost model
Naïve Bayes Classifier

Conclusion

1. **Inferences from EDA**

1. The maximum tweets came from London.

2. From the ‘Sentiment’ column, I came to know that most of the peoples are having positive sentiments about various issues shows us their optimism during pandemic times. Very few people are having extremely negatives thoughts about Covid-19.
3. Columns such as “UserName” and “ScreenName” do not give any meaningful insights for our analysis. Hence we are not using these features for model building.
4. From AUC Curve the AUC Score is 0.63 for the Naive Bayes Classifier which is not good but acceptable.
5. For multiclass classification, the best model for this dataset is CatBoost with Testing Accuracy Score = 0.62329.

2. For multiclass classification, the best model for this dataset is CatBoost with Testing Accuracy Score = 0.62329.