Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. **Mohd Taufique**: taufiquemohd2@gmail.com • Data Exploration & Variables Identification. • Text Preprocessing – Stemming, removed Stopwords & Punctuations marks. • Performed Exploratory Data Analysis (EDA). • Performed Tokenization and Vectorization • Feature Engineering • ML Modelling and Conclusions • Presentation, Technical Documentation, Project Summary 2. **Sonica Sinha**: sonicasinha2012@gmail.com • Data Understanding- How data looks, how big the data is. • Data Wrangling- tackled null values. • Text Cleaning- Removed URL, User handle and special characters. • Made Word-cloud for Tweets of different sentiments. • Categorical Variable Analysis • ML Modelling and Conclusions • Presentation, Technical Documentation, Project Summary
Please paste the GitHub Repo link.
Mohd Taufique GitHub Link: - https://github.com/MOHD-TAUFIQUE/Covid-19-Tweets-Sentiment-Analysis_Capstone-Project-3 **Sonica Sinha GitHub Link: - https://github.com/Soni-Test/Sentiment-Analysis-Supervised-ML-Classification-Predicting-sentiment-of-COVID-19-tweets**

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

The world encountered its first corona virus case on December 31, 2019 in China. People all around the world in a state of confusion and fear started sharing their concerns on twitter. COVID-19 originally known as Coronavirus Disease of 2019, has been declared as a pandemic by World Health Organization (WHO) on 11th March 2020.

We were provided with Coronavirus_Tweets.csv dataset.

- Shape of the dataset is (41157,6).
- The Sentiment column is the dependent variable which consists of 5 different labels which are positive, negative, neutral, extremely positive and extremely negative.

The challenge is 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 first step in the analysis involved understanding the data, exploring the data, identifying the variables, and then performed EDA & text preprocessing like looking for any null values and tackling them, Removing stop words, user handle, special character and punctuations. Applying a Stemming function on the given text which reduces the word to its base word. For this purpose, we import necessary python libraries, load the datasets, and used library like pandas, NLTK, porter stammer and NumPy.

The second step involved analyzing the necessary columns and show the analyzed result using different visualization charts like bar graph, clustered bar chart, count plot, pie chart, word-cloud etc. For this purpose, we used data visualization libraries – seaborn, plotly and matplotlib.

The third step involved feature engineering – dealing with categorical variables, encoding the sentiments, performed tokenization using count Vectorizer and TF-IDF. At last after all these Data preprocessing steps, we did train-test split and go for applying the Machine Learning algorithm to build a model that will predict the sentiment for the test data. For this purpose, we used Supervised classification ML models like Logistic Regression, Naïve Bayes, SVM, SGD, Catboost and Random Forest and calculated evaluation metrics to check the performance.

The Final step involved is summing up the key observations and insights developed during the analysis and ML Model selection. Some key takeaways were: Getting a plot on the number of tweets in different months, Plotting the number of tweets on a particular day of the month, Plotting the number of positive, negative and neutral sentiments, plotting of top 15 locations with maximum number of tweets. SGD proved to be the best models among all the models used.