Capstone Project Submission

My Name, Email and Contribution:

Gaurav Yogeshwar (yogesh.grv9@gmail.com)

- Upload dataset to Google colab.
- Examining Null Values
- Data cleaning
- Correction of data types
- Data wrangling
- Data Visualizations
- Fitting different models
- Tuning the Hyperparameters
- Evaluation Metrics
- Lime explanation
- Model Persistence
- Technical Write up
- PowerPoint presentation
- video explanation
- Project summary

Github Link-

https://github.com/Gaurav2912/Cronavirus-Tweet-Sentiment-Analysis.git

Goal -

This project addresses the problem of sentiment analysis and to build a classification model to predict the sentiment of COVID-19.
That is, classifying tweets according to the sentiment expressed in them: extremely positive, positive, extremely negative, negative or neutral.

Short summary of Capstone project and its components.

The worldwide coronavirus pandemic has led to the establishment of worldwide curfews, quarantines and lockdown to mitigate further spread of the virus. During this time, it can be helpful to track the public's responses to these changes.

In this EDA project we were provided datasets, which have
Location
Tweet At
Original Tweet
Label
Features

At first, we break down the datasets by importing necessary library classes, followed by checking missing and unique values, getting a statistical summary of the numeric columns, converting the data types to similar objects, doing feature engineering and making the entire dataset ready for analyzing & plotting actionable insights.

After examining null & duplicate values from the dataset we directly went deep into the visualization steps. Machine learning models cannot take data in text format directly, to pass data to model first I need to convert it into a bag of words.

Then training and interpreting with multiple models and hyperparameter tuning with the best model, then predicting and computing evaluation metrics for the model such as accuracy, precision, recall, F1 score and area under the ROC curve.

eventually
I make predictions on a single input and this is likely to happen LIME
Then saving the model for later use.

Overall Conclusion

1. Exploratory Data Analysis

- Length of the tweets is negatively skewed.
- Length of neutral sentiment tweet is positively skewed.
- As the location suggests, most of the places are from English speaking countries or countries where people understand English, such as the UK, USA, India, Canada, Australia etc., and among these most of them are also from the US and UK.
- The first quarter has the highest percentage of overall negative sentiment tweets as compared to the other.
- Over time, the trend of tweets with negative sentiment is decreasing except in the last two months.
- By analyzing hashtags
  - Most of them are about coronavirus outbreak and pandemic, Social distancing, lockdown , staying at home etc..
  - Due to the lockdown, people are also facing problems due to the closure of supermarkets, shortage of food, and running out of toilet paper.

2. **Natural Language Processing and Machine Learning**

- Even after passing through different machine learning models, we noticed that there is not much improvement in accuracy, so I changed the Sentiment feature to Binary class and the accuracy improved from 61 to 86 percent.
- Among all models Stochastic Gradient Descent Classifier gave best performance so I tuned its hyperparameters using grid search with five fold stratified cross validation, and accuracy has increased to 87.5 percent.
- Area under ROC curve is 0.92 , area under precision-recall curve is 0.88.