Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1). Kunal Gawande

E-mail: gkunal8019@gmail.com

- Data preprocessing
- Perform story generation from visualization
- Naive Bayes Classifier for multiclass Classification
- Stochastic Gradient Descent-SGD Classifier(MULTICLASS CLASSIFICATION)
- RANDOM FOREST CLASSIFIER (For Multiclass Classification)
- Extreme Gradient Boosting (For Multiclass Classification)
- Support vector machine(For Multiclass Classification)
- Logistic Regression(For Multiclass Classification)

2). Bipasha Zade

- Tree Based Model Selection
- Model Deployment
- Feature Importance
- Shapley Additive explanations

3). Deepali Mahajan

E-mail: deepali2062@gmail.com

- Debugging Error
- Data Sorting
- Technical Documentation
- ppt Presentation
- Approach Towards Plan
- Seaborn, matplotlib
- Heatmap
- Linear Model Selection
- Evaluation Matrix

4). Chinmay Rojatkar

E-mail: chinmayrojatkar4@gmail.com

- Data Sorting
- Matplotlib
- ppt Presentation
- Data Visualization
- Technical Documentation
- Approach toward Plan
- Line Plot, Bar plot , Histogram
- Heatmap
- Linear Model Selection
- Data Preparation

5). Nikhil Aggarwal

E-mail: nickagg30899@gmail.com

- Data Cleaning
- Data Analysis
- Error Handling

Please paste the GitHub Repo link.

Kunal Gawande Link:- https://github.com/gkunal8019

Chinmay Rojatkar Link:- https://github.com/ChinmayRojatkar

Bipasha Zade Link:- https://github.com/Bipashazade

Nikhil Aggarwal Link:- https://github.com/Nikhil8815

Deepali Mahajan Link::- https://github.com/deepali2062

Introduction:

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether people attitude towards a particular topic is positive, negative, or neutral. On 31st December, 2019 the Covid-19 outbreak was first reported in the Wuhan, Hubei Province, China and it started spreading rapidly all over the world. Finally, WHO announced Covid-19 outbreak as pandemic on 11th March, 2020, when the virus continues to spread. We are analyzing data during pandemic time to gather correct information for making policy for further use.

Fitting Different Models

- Support Vector Machine
- Naive Bayes Classifier
- Stochastic Gradient Descent
- Random Forest Classifier
- Extreme Gradient Boosting
- Logistic Regression

Please paste the GitHub Repo link.

GitHub Link:-

https://github.com/gkunal8019/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

Problem Statement:

This challenge asks you to build a classification model to predict thesentiment 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. You are given the following information:

Conclusion:

We focused on sentiment analysis for sentence labelling. We described the preprocessing steps, and pipeline steps within which text normalization and model cross-validation is included, and performance has been measured using balanced accuracy, f1 score etc. We used “Stemming” instead of Lemmatization to reduce dimensions, for the same reason we haven’t tried tf-idf or term frequency vectorizer. We concentrated on feeding our model with word count information. We assume, in the case of binary classification, we can further improve this score.

In this way, we can explore more from various textual data and tweets. Our models will try to predict the various sentiments correctly. I have used various models for training our dataset but some models show greater accuracy while some do not. The best model for this dataset would be Stochastic Gradient Descent.