Nykole Edward

Information Systems and Applications

Data Mining

Lab 2- Kaggle Competition Report

For the purpose of the Kaggle Competition, a Twitter data set provided by the IDEA Lab was crawled with tweets containing specific hashtags. The dataset contains eight classes namely anger, joy, sadness, anticipation, disgust, surprise, trust and fear. These classes were then renamed as “emotions”. The end result of this competition aimed to allow students to perform preprocessing on the dataset in order to clean the data and then to train a model learnt in Data Mining to be able to predict the emotion behind every tweet.

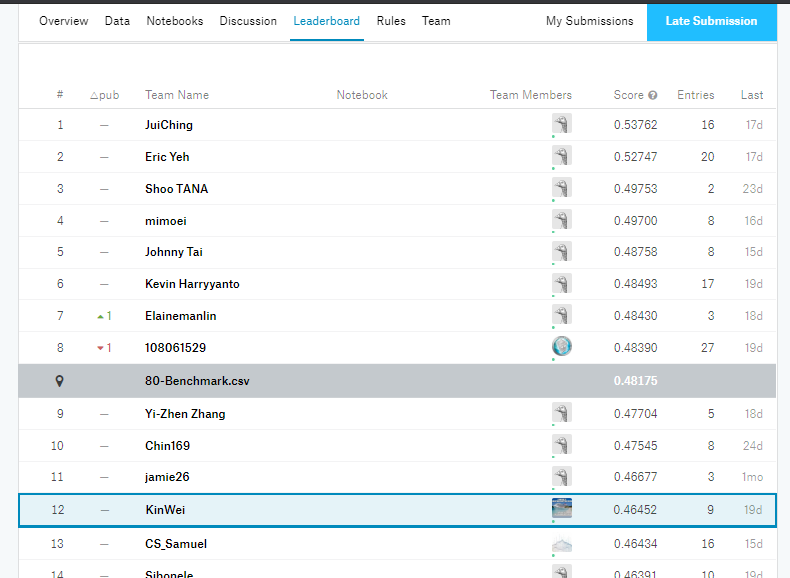
I proceeded to import the necessary libraries needed to perform the tasks. I then set file paths to separate the dataset by emotion, ID, and tweets. For the stage of preprocessing of the data, I proceeded to import the datasets and to make it easier to visualize I decided to merge the data by a specific column. This specific column was common to all three datasets. I chose the column “tweet\_id”. After that I merged the tweet\_id column with data\_id then performed a merge of that combination with emotion\_id. After that was done I proceeded to separate the training data from the testing data.

I proceeded to pickle the train and test data files in order to be able to save them on disk. After loading the pickle file into the workbook, I then decided to make the train set much smaller in order to be able to work with it more comfortably. I then checked the size of the reduced dataset to ensure that it was adequate. I then decided to observe the combined dataset to ensure that the headings were correct and to view a sample of the data in the table.

I then went on to perform some exploratory analysis. I first tried a decision tree and tested it out I was able to get 0.35 accuracy so I then proceeded to test out the decision tree by increasing the number to about 30000 the accuracy went up but I still did not meet the benchmark. So, I decided to try the Naïve Bayes Model. I did a frequency chart just to get an idea of how many times the emotions listed above showed up in our dataset. The next step for me was to start my feature engineering. I decided to include some dummy values into the dataset to test the data and the accuracy of it. I went further to add a tokenizer for tweets and then onto using TF-IDF to test the features of the model I would be working with. I then fit some predictions for the model and assigned some predictions and outcome variables. The final steps I took were to perform model fitting where I trained my model using logistic regression. I then fit some predictions and also predicted test values. For the output I assigned results to data frame and output my predictions as a csv file and uploaded to the competition board.

This assignment made me notice that the decision tree while it was easier to build the accuracy would not change much even if I increase the number of the sample. It became clear when I utilized the naïve bayes just how much more accurate the results started to be. This method took the most time to figure out but then I was able to increase my accuracy. I wanted to try another model just to see what would happen however I still was having some trouble getting any of the codes to work so I stuck with my naïve bayes as is.

**Leaderboard Standing as of 15th December 2019**

**1st Competition Leaderboard**

2nd Competition Leaderboard

