Assignment 2 Report

Anoop (2015CS10265)

March 10, 2018

Text Classification

Part a

Text Classification using Multinomial Naive Bayes
Train Accuracy = 0.68448
Test Accuracy = 0.38752

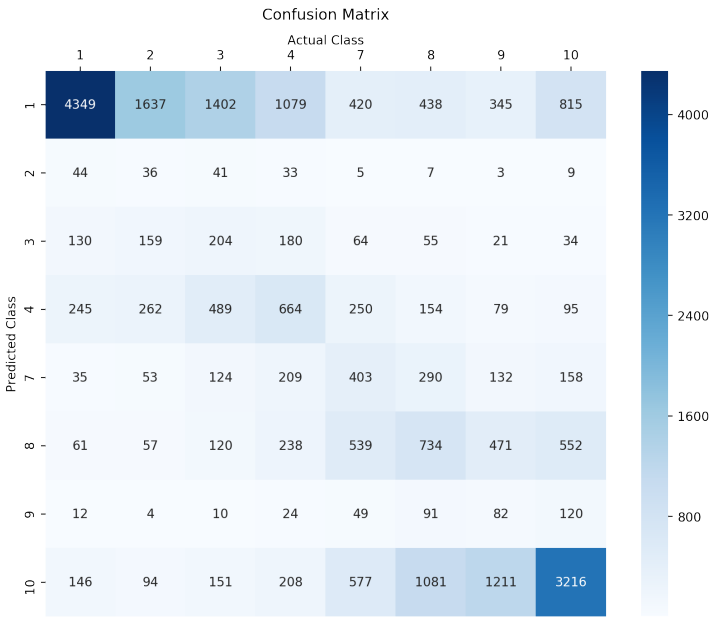
Part b

Text Classification using Random guessing
Test Accuracy = 0.07264

Text Classification using Majority Class
Test Accuracy = 0.20088

Part c

Text Classification using Multinomial Naive Bayes - Confusion Matrix



Micro F1 score = 0.2793
Macro F1 score = 0.7800
Class with highest diagonal entry = 1
From the low Micro F1 score, we can infer that there is class imbalance. Also from the Confusion Matrix we can observe that there are more predictions of class 1 and 10. This means that the model is influenced a lot by the class imbalance.

Part d

Text Classification using Multinomial Naive Bayes - With Stemming

Train Accuracy = 0.6798

Test Accuracy = 0.38684

There is not much increase in accuracy as compared with the model with no stopword removal and stemming. This is because of class imbalance that we observed in from the confusion matrix previously.