**Stumbleuponchallenge Approach**

**Name-Amit Kumar Jena**

**Steps**

**1.Imported the requisite libraries**

**2.Imported Java and h2o library as my strategy was to use AutoML for the classification problem to determine the best model**

**3.imported the files into my colab notebook(used colab because it was easier to get java and use h2o library) rather than installing java in my desktop**

**4.checked the imbalance in data. found out that the data imbalance can be ignored since it is marginal**

**5.checked for missing data .No missing data was present in the dataset**

**6.Converted data of both the train and test dataset using regex and used labelencoders and stemming to get the rootwords**

**7.converted “boilerplate” column to tfidv vectors**

**8.used h2O to find the best model for training**

**9.used GridsearchCV to find the n\_estimators and learning\_rate(hyperparameter tuning)**

**10.fitted the model and checked for roc\_auc and accuracy score with f1 score for each class**

**11.used test data after converting the test data to tfidf vectors to find the predictions for test data**

**12.dumped the results.**