1. I used a function to clean the data using regex to remove some punctuation , links if the there any , numbers ..etc.
2. SGD Classifier did better than other models especially in recall , it’s a simple yet efficient approach , also it performs well with sparse data.
3. The data was apparently Imbalanced , some solutions are k-folds cross validation or using oversampling.
4. I based my evaluation on Recall , Precision and Accuracy .
5. There is a lot of duplicated data and at the same time there isn’t enough examples for some classes especially the Accountancy.