1.First I have choosen different techniwes to clean my data, after removing duplicates I started preparing the text data before I entered it to my model to get the highest accuracy, I removed punctuation marks, removed stopwords like in,out ,of ,etc, then applied tokenization to split each sentence into several words. Then performed stemming so that each word returns to its core original, then limmitization so that the core is meaningful.

2. I have tried multiple classifiers like stochastic gradient decent , logistic regression and naïve bayes

But I have chosen logistic regression because its balance results since I measure the score on both train and test data to make sure my model doesn't overfit and logistic regression had the best score on test data and was the the least overfit model since the score on train data was so close to that on test data,

3. There are several ways to deal with Imbalance learning such as upsampling, downsampling and adding class weights into consideration when training.

4.my model can have better performance if I had more data since half of the data was duplicated and by also tuning the

hyper parameters of the model I can obtain better performance.

5. I did evaluate my model first by using simple score function to get intuition then using precision ,f1 score and recall since score function only is not sufficient to measure a classifier performance, the best score is 1 and the closer you get to 1 the more correct predictions you had.

6.i think my model limitations are 2 things first it tends to overfit a lilttle bit ,second it is a little bit bias to the most frequent class which is IT since the data is unbalanced and accountancy class occurance is very low compared to IT class.