### (Questions & Answers)

### - Which techniques you have used while cleaning the data if you have

### cleaned it?

* I used regular expression to filter data from numbers and special characters.
* I used porter stemmer from nltk library to remove the commoner morphological and inflexional endings from words in english.

- Why have you chosen this classifier ?

- Extra Trees is like Random Forest, in that it builds multiple trees and splits nodes

using random subsets of features, but with two key differences: it does not

bootstrap observations (meaning it samples without replacement), and nodes are

split on random splits, not best splits.

- How do you deal with (Imbalance learning)?

- I used stratify in splitting the dataset to make the dataset

balanced after splitting,I did not use smote technique or

oversampling until we do not increase the noise, but it was enough

to keep partial duplicates as a benefit for this.

- I did not use stratified kfold for imbalanced because the

algorithm that I used deal with the features of data and does not

take it instance by instance.

- How can you extend the model to have better performance?

* We can use techniques such as CounterVectorizer or build use another algorithm to predict classes that the used algorithm can not predict it correctly and use both in parallel as the fisrt as the main model and the other is the sub model for voting for overlapped class.
* We can use gridsearch for tuning to define the best parameters for the model.

- How do you evaluate your model? (i.e. accuracy, F1 score,

Recall)?

* We used classification report and confusion matrix to check precision,recall and F1 score, we noticed that there is not imbalanced scores,Macro AVG is greater than 0.9 and F1 score too, in addition, recall for class Accountancy is 0.79 because count of this class in dataset is limit.

- What are the limitations of your methodology or Where does your

approach fail?

* There is an overlap in IT class and Marketing class from confusion matrix, there are some titles are confused for model to classify it as IT class or Marketing class.