Machine Learning Intern.

Job title Classification by industry

(Multi-text Text Classification Task)

Name: **Mariam Khaled Mohammed Kamel Elwy**

Group: **8**

- Which techniques you have used while cleaning the data if you have cleaned it?

- Removing special characters using regular expressions, tokenization, removing stop words, standardizing by converting all tokens to lowercase.
- Using Count Vectorizer to create a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix.
- Using TF-IDF to calculate relevance of words.

- Why have you chosen this classifier? (E.g. I used Multinomial Naive Bayes because it is easy to interpret with text data and there are more than two outcomes).

- I've used Linear SVC classifier as it's regarded as one of the best text classification algorithms and gave the best scoring metrics among other candidate classifiers.

- How do you deal with (Imbalance learning)?

- Through resampling or giving weights to training instancesI chose to balance weights of the data instances as it achieved better results after trying both methods in this problem.

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

- Hyper parameter tuning using methods like Grid search, randomized search and Bayesian optimization.

- Using Ensemble methods like bagging and boosting.

- Further Data processing (e.g. stemming).

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

- I choose F1-score as it’s a harmony between recall and precision and is better suited for imbalanced data problems.

- What are the limitations of your methodology or Where does your approach fail? (e.g. your predictions are biased because you do not have enough data for a certain class)

- It would’ve been better if we had more data instances from the “Accountant” class as its representation in the dataset was not sufficient.