**Job title Classification by industry**

**(Multi-text Text Classification Task)**

**Report**

**Description**:

You can think of the job industry as the category or general field in which

you work. On a job application, "industry" refers to a broad category under

which a number of job titles can fall. For example, sales is an industry; job

titles under this category can include sales associate, sales manager,

manufacturing sales rep, pharmaceutical sales and so on.

**Details:**

1- We have a dataset that has two variables (Job title & Industry) in a csv

format of more than 8,500 samples.

This dataset is imbalanced (Imbalance means that the number of data

points available for different classes is different) as follows:

* IT 4746
* Marketing 2031
* Education 1435
* Accountancy 374

**Answer the following questions:**

After I finished the project, I can now answer these questions

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

**cleaned it?**

I made function to remove from title stop word and any special characters , and made all title in lower case

**- Why have you chosen this classifier?**

After I tried different classifier

(GaussianNB, MultinomialNB, ExtraTreesClassifier)at first before we solve imbalance data , each model made different score (F1 score, Recall ) ,like there are model predict class right than other , So I used Voting Classifier method , it’s about use different models to get best result by majority BUT the best model was Multinomial Naive Bayes

because it is easy to interpret with text data and there are more than two

outcomes, but mostly all models work well after solve imbalance data

**- How do you deal with (Imbalance learning)?**

Handling imbalanced data using Oversampling technique on data

It's generating synthetic data that tries to randomly generate a sample of the attributes from observations in the minority class.

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

1) get more data sample and feature

## 2) I've read article about word embedding is more suitable for our case is about convert word to vector

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

Due to our data is Imbalance so it was wrong to depend on accuracy

So, I used F1 score and recall and precision to evaluate the model

- **What are the limitations of your methodology or where does your**

**approach fails?**

I think the model work well, but maybe fail if wrote one word to describe the title job , so it can confuse for the model to predict the industry like if wrote “Manger” is he/she manger of school so can predict education or manger of company or etc.

**References**

<https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794>

<https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794>

<https://www.youtube.com/playlist?list=PLQpVVU8sLpqx0t7aFOfXJzJMrtxX5yvKU> (youtube channel)