Multi-Text Text Classification Task

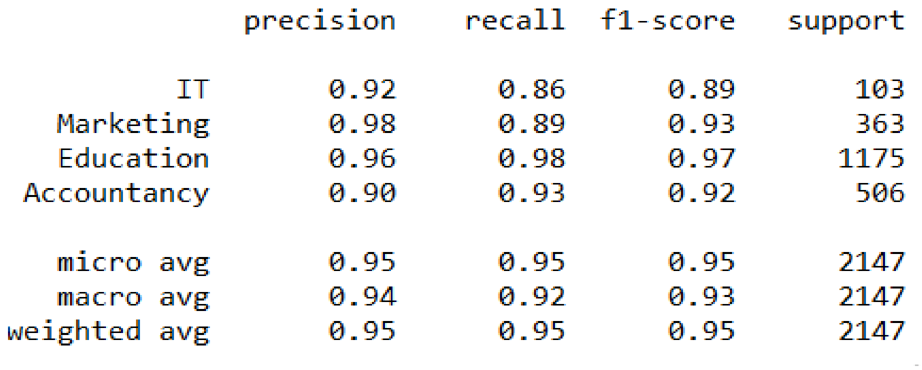
- I used Regular Expression to clean the Data set from the Numbers , Special Characters and Punctuation . by using ‘re’ library with this Pattern re.sub("[^A-Za-z]+"," ",job) to substitute any character not Alphabet with space .

2- I used Linear Support Vector Machine because that classifier used for solving Multiclass Classification Problems from Ultra large Data sets and also gives high prediction accuracy over the majority class while maintaining a reasonable accuracy for the minority classes .

3- There are more than one Technique to deal with imbalance Data set . One of them is to Resample the training set . I used Under-sampling by reducing the size of the abundant class .This method is used when quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modeling .

4- To improve the performance of The Model before the Texts are classified i used TFIDF model and also made a Feature Vector with a fixed size using Bag of Words Model (BOW) as the Classifier can’t directly process the Text documents in their Original Form.

5- To evaluate My Model I import metrics from sklearn library and call metrics.classification_report() and passing the Y_test , Y_predict and target_names as Parameters and it gives me back ..



```
              precision    recall  f1-score   support

          IT       0.92      0.86      0.89       103
   Marketing       0.98      0.89      0.93       363
   Education       0.96      0.98      0.97      1175
 Accountancy       0.90      0.93      0.92       506

   micro avg       0.95      0.95      0.95      2147
   macro avg       0.94      0.92      0.93      2147
weighted avg       0.95      0.95      0.95      2147
```

6- I tried to handle the imbalanced data and the model gives a reasonable prediction accuracy but there's some predictions are biased to the abundant class .