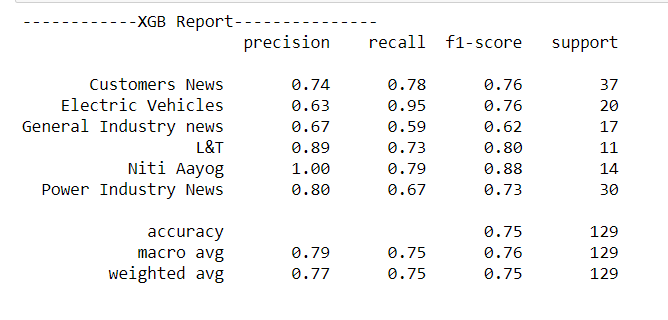
**News article classification-**

Files included-

1. doc2vec.model – Doc2Vec trained model
2. DOC2VEC+ML- Ipython notebook with code
3. Sample- XML dataset

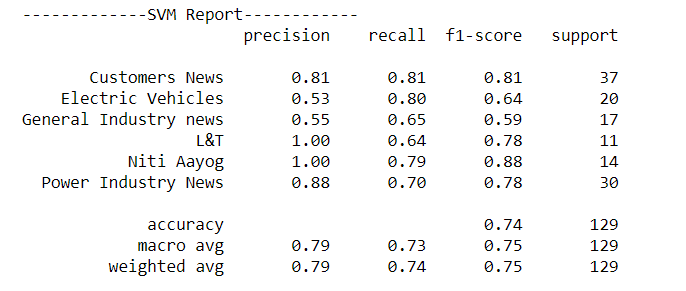
Key highlights -

* Minority classes were clubbed to improve class balance
* For the purpose of this assignment the model classifies the major tag of a news article
* Doc2Vec is used for creating document embedding for each news articles.
* Classifier is used in a cascaded manner to classify the news articles.
* As dataset size was small a 70-30 split was taken between train test.
* Two models were trained using Grid Search for parameter tuning and cross validation was used .Following which performance were compared-
  + XGBoost



A combined accuracy of 75 percent was achieved on test data as well as satisfactory metrics(precision, recall, F1) achieved for all the classes.

* SVM



Marginally lower accuracy of 74 percent. However the precision of L&T was higher compared to XGBoost

Next steps-

1. More ML models could be tried and further hyper parameter tuning for better results.
2. Models like ULMFit could be tried to achieve higher performance by using the power of transfer learning for such tasks.