Problem statement

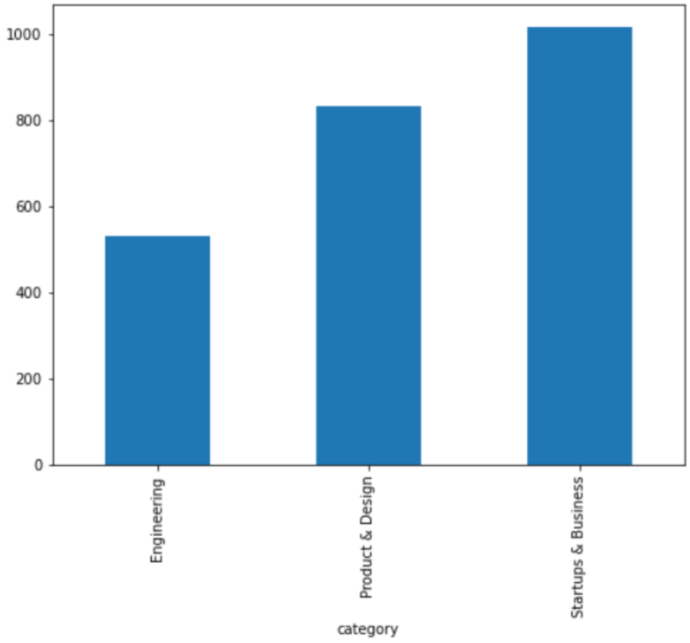
Text Classification problem using two classifiers :Naive Bayes and Support Vector Machine Classifiers
We have worked on JSON file for a group of categorized articles ,the goal is to measure the accuracy of each algorithm and suggest the best classifier

Data preparation

- Reading Data
  - Pandas - Dataframe is used to hold the json data
- Cleaning Data
  - Remove any duplicate in rows , filtered on Body's content and make sure no empty fields

Assumption

- The model will be trained on Body not title as it is more informative
- We detected the 3 categories and change them to numeric value
- TfidfVectorizer is used to change body content to numeric values
  - analyzer='word' , and give english stop word list for better features •
  - Use train_test_split with stratify=category to make sure that the data is not baised to certain category and be balanced , no need to have validation dataset
  - Plot data to check if it is imbalanced or not , the data can be considered to be unbalanced , as startup and business articles are much more than engineering and product and design



NB classifier Vs SVM

classifier	NB	SVM
accuracy	0.7595419847328244	**0.8854961832061069**

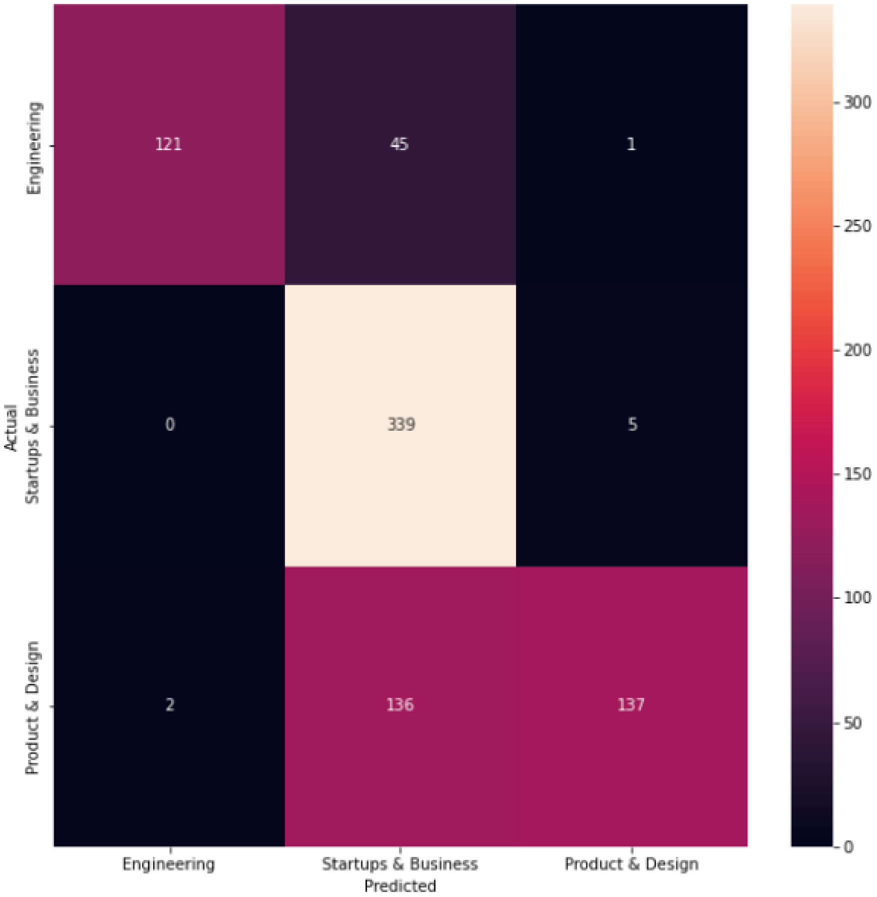
Conclusion

SVM is much better with text data as we see below The vast majority of the predictions end up on the diagonal (predicted label = actual label)

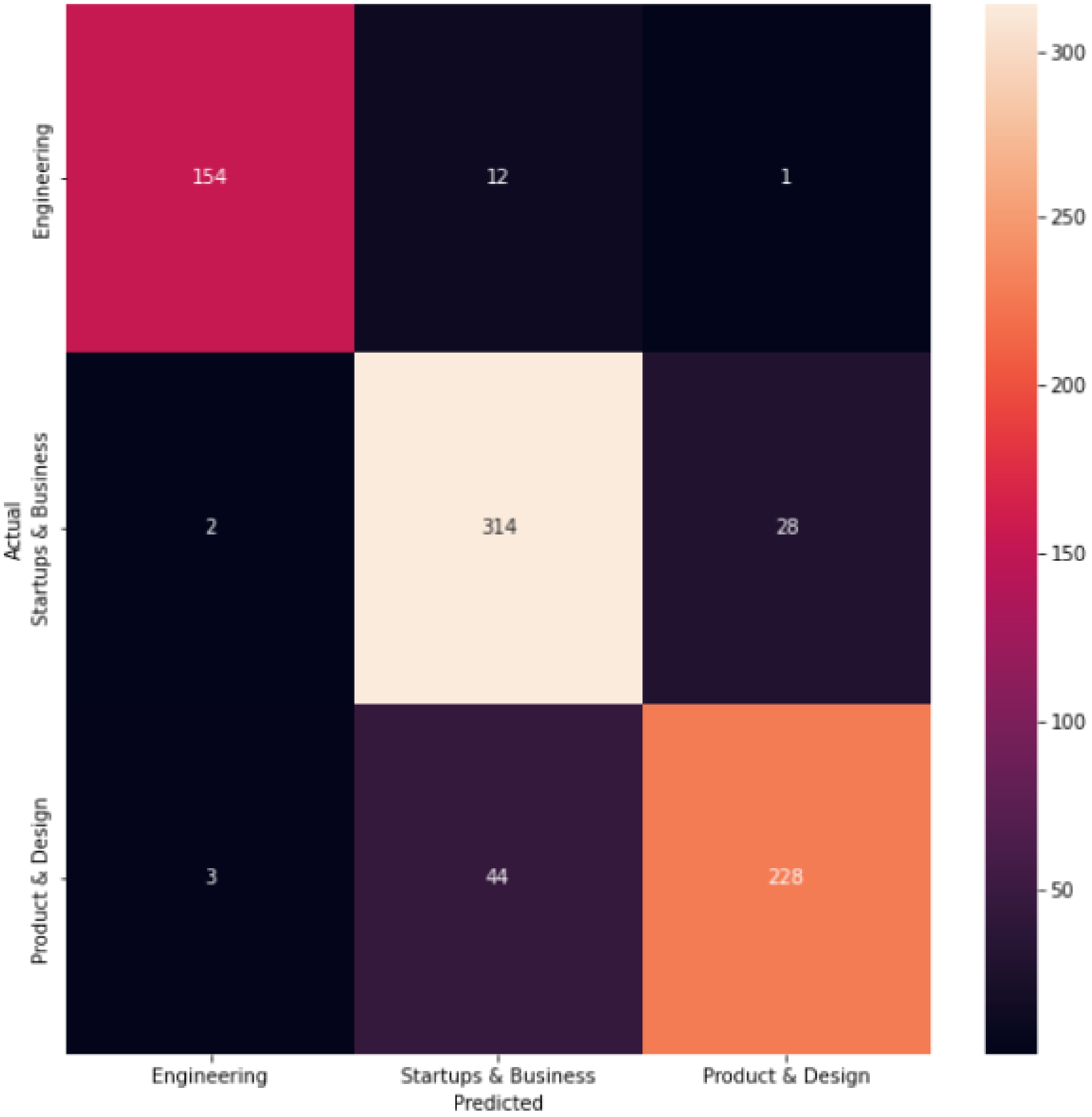
Result analysis

- confusion matrix to evaluate the accuracy ,The daiagnoal shows the correctly classified articles

NB model matrix



SVM Model matrix



Actual \ Predicted	Engineering	Startups & Business	Product & Design
Engineering	154	12	1
Startups & Business	2	314	28
Product & Design	3	44	228

As shows above by numbers , svm model is better for text classification