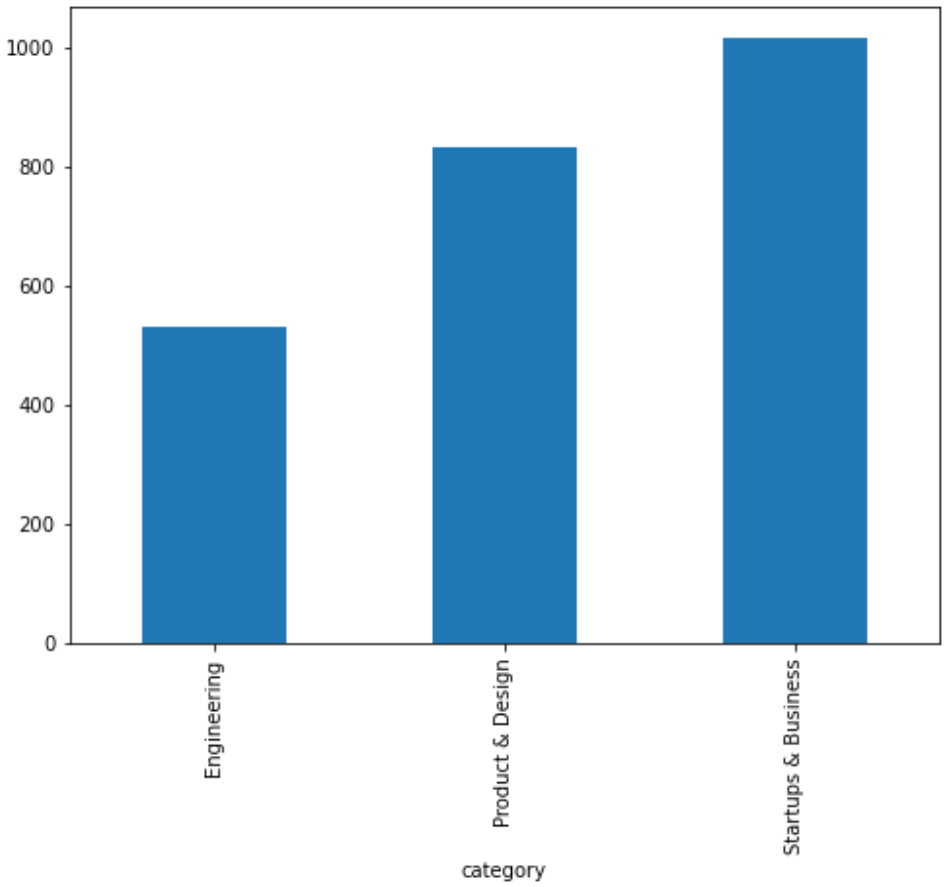
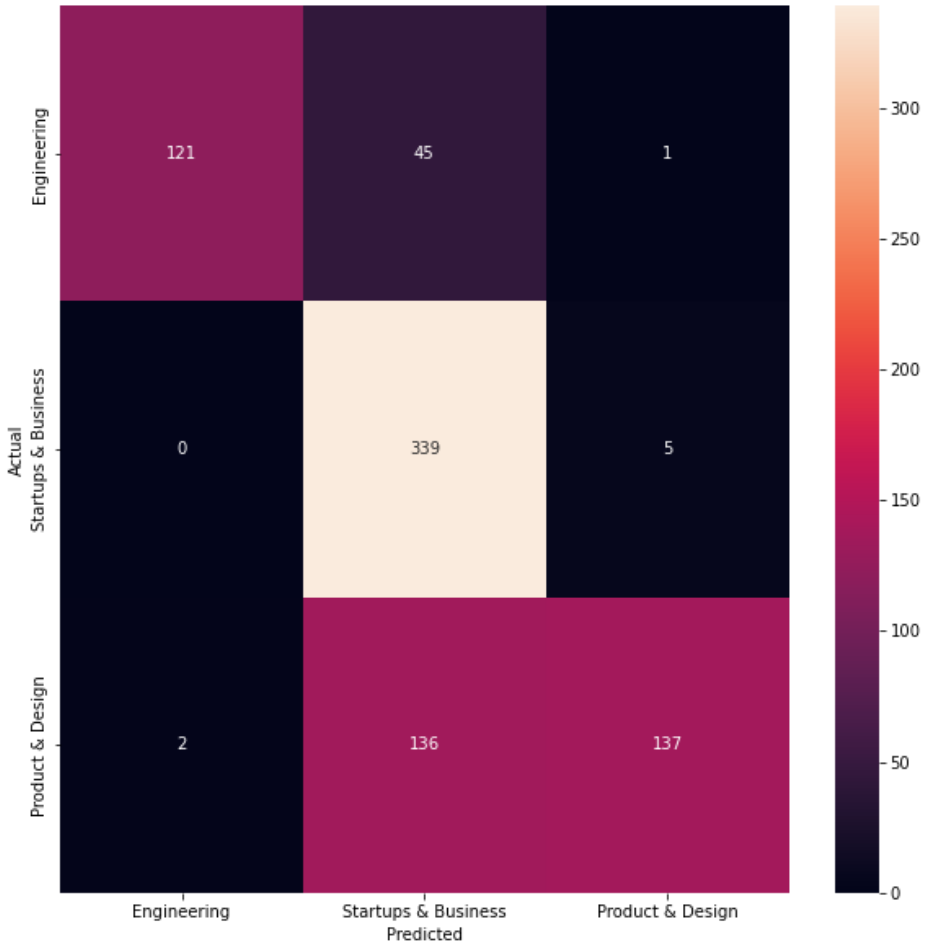
Naive Bayes and Support Vector Machine Classifiers

- Reading Data
- Pandas - Dataframe is used to hold the json data
- Remove any duplicate in rows , filtered on Body's content
- The model will be trained on Body not title as it is more informative
- Detect 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

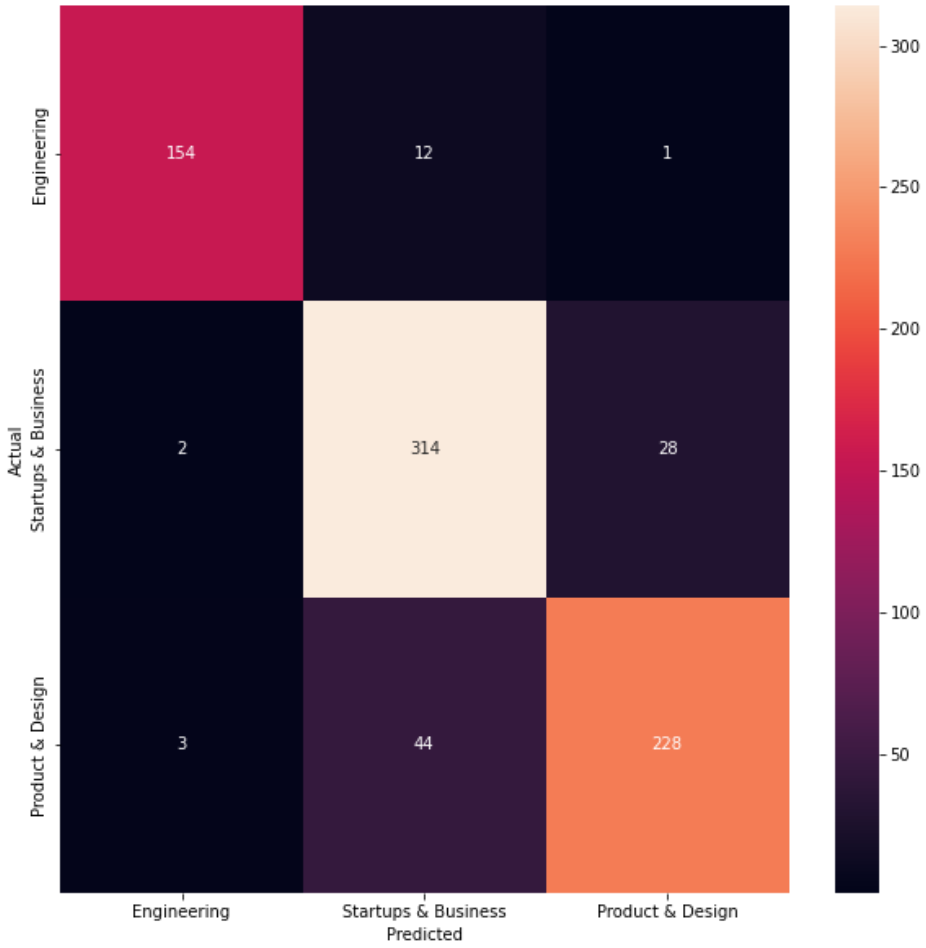


NB classifier Vs SVM

- NB accuracy 0.7595419847328244
- SVM accuracy 0.8854961832061069
- 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)
- NB



- SVM



- So svm is better in text classification