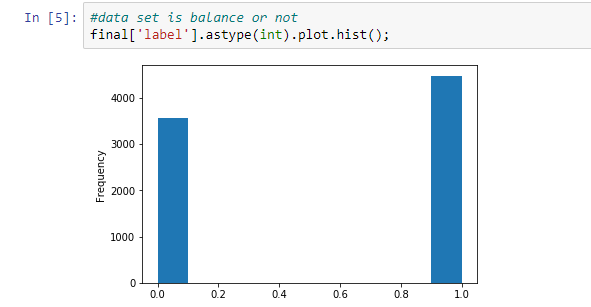
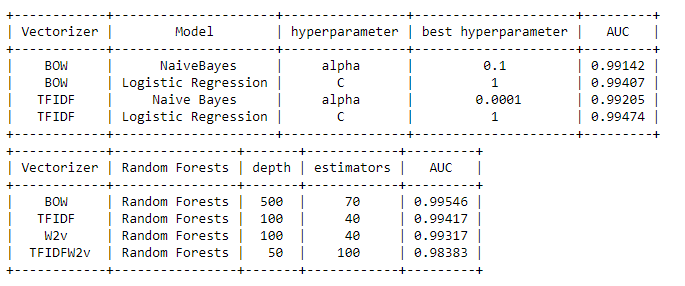
**Report**

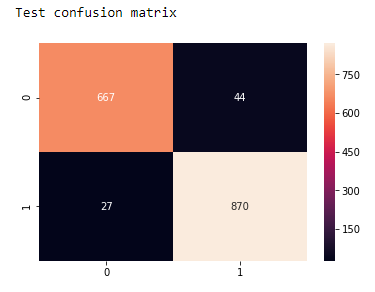
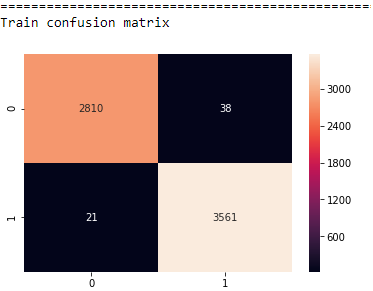
1. Understanding the Data**:** Before we start cleaning data for a machine learning project, it is vital to understand what the data is, and what we want to achieve. Without that understanding, we have no basis from which to make decisions about what data is relevant as we clean and prepare our data.By seeing data it is binary classification problem. First read a both text file and convert into dataframe for data cleaning.
2. Data cleaning: remove urls and all tag, word with number, special character from text. Remove stop words.
3. Check data set is balance or not, in this case study dataset is balanced.

****

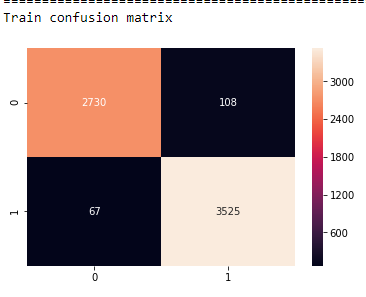
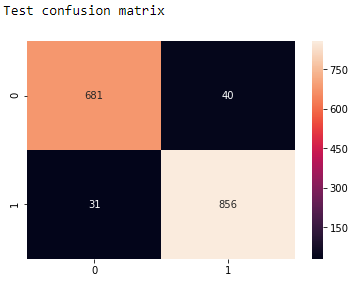
1. Featurization : for text data we can use featurization like Bag of words, TfIdf, AvgW2v and TfIdfW2v.
2. Diving dataset into test and train data by 20:80 ratio.
3. In each model for finding best hyper parameter i used grid search method.
4. For classification I have try 3 classification model. NB is benchmark for text classification, so first find NB classification using Bow it give good result **AUC is 0.99142**. below table shows the AUC for each model I have try.as we can see best AUC for random forests for BoW is **0.99546** among all the model but just looking at AUC we cannot decide that Random forest give best result because all AUC value are closed to each other.

****

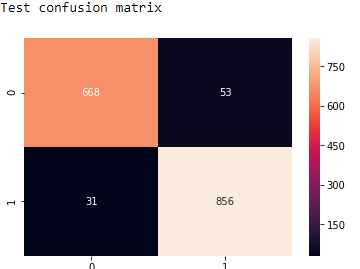
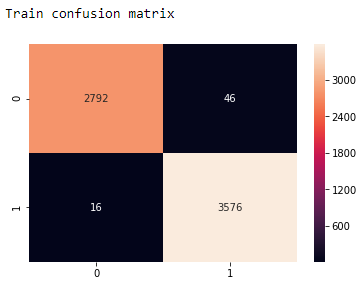
So we see the confusion matrix for NB and logistic regression and random forests . NB with BOW featurization , As we can see FN and FP are small.



Logistic regression with TfIdf, as we can see confusion matrix for NB with BOW is better than this confusion matrix.

Random forests with TfIdf.



So i select NB with BOW as classfier for this case study.

1. Apply model on 3 text file for predition, then save data into csv file.