Test analysis using Naïve Bayes Classifier

Introduction:

A Bayes classifier is a simple probabilistic classifier, which is based on applying Bayes' theorem. The feature model used by a naive Bayes classifier makes strong independence assumptions. This means that the existence of a feature of a class is independent or unrelated to the existence of every other feature and hence it is Naïve in nature.

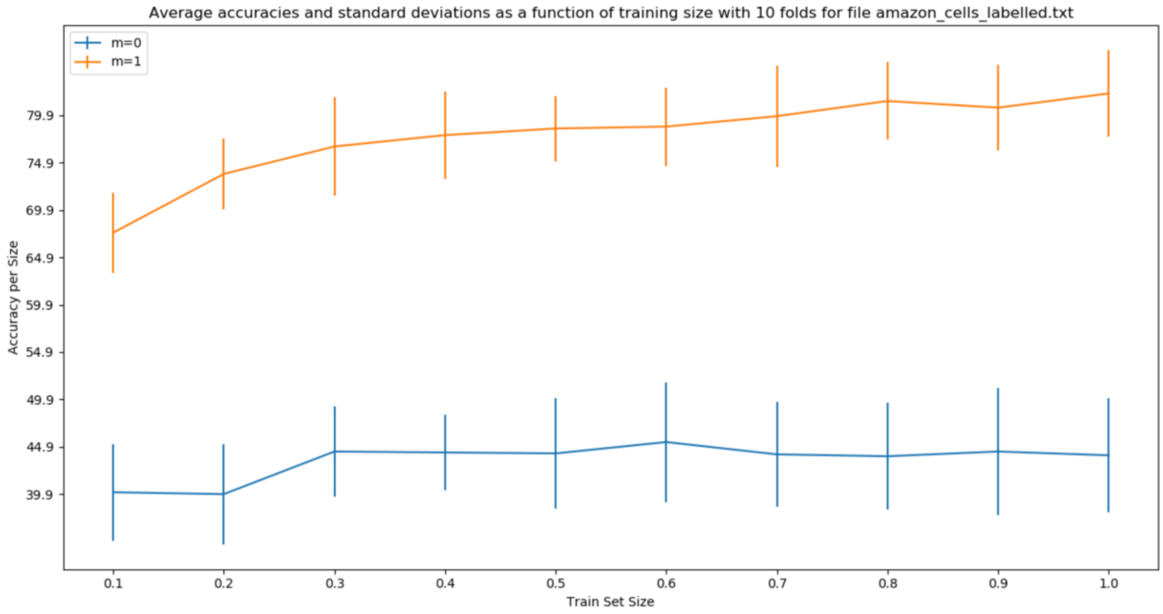
Here in our experiment, we are considering 3 datasets as input:

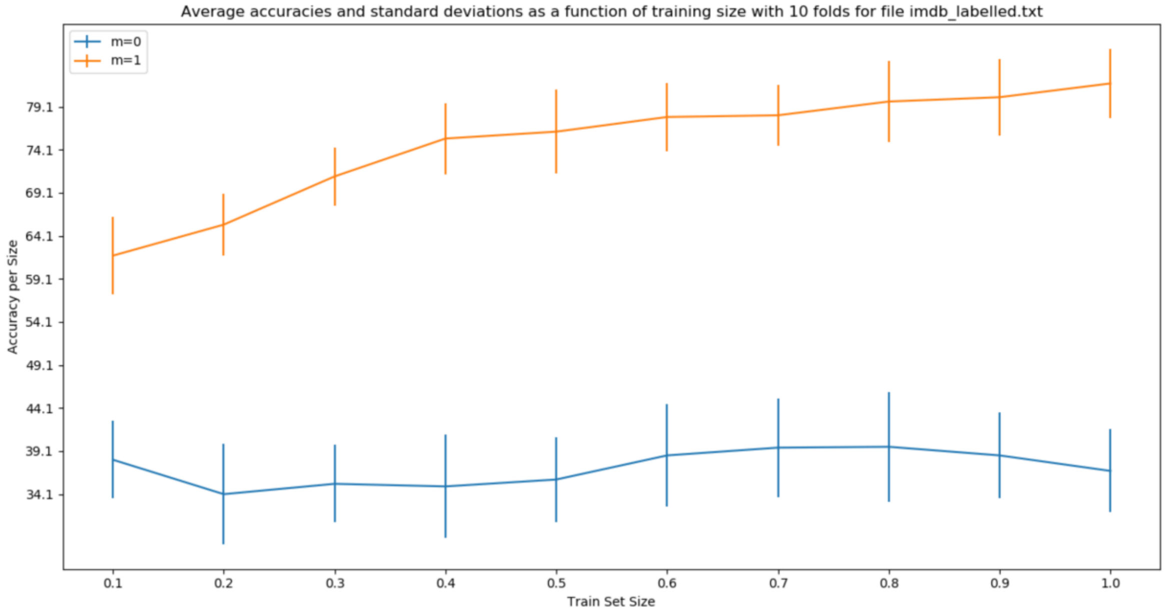
1) Amazon
2) Yelp
3) IMDB

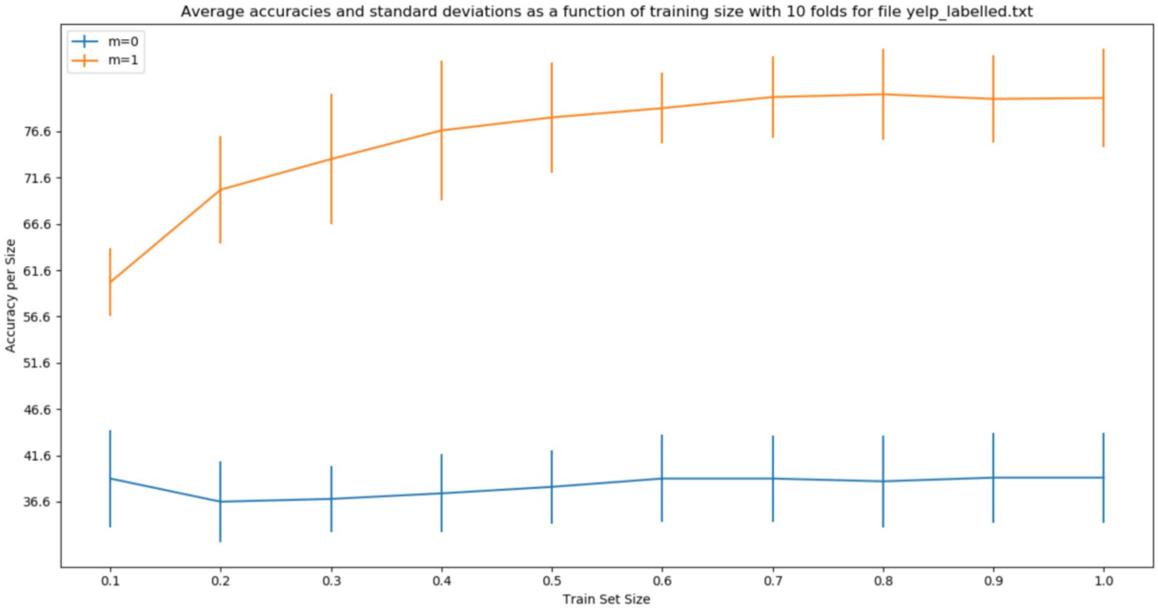
Experiment 1:

In this experiment, we have used stratified cross validation to generate learning curves for MLE and MAP where we used the smoothing parameter as 1. We divided our dataset based on the class value and permuted the documents in the set to avoid any skewness of data. Later, we divided our dataset into k folds, used 10 in the experiment and merged the folds for each class. After generating the folds, we used subsamples of randomized data of sized 0.N,0.2N,0.3N….N and trained on these sets train and evaluated predictions on the test data test(i).

Below graphs have been captured for each dataset, and these graphs are plotted for Average accuracy (y axis) against training set 0.1N to N (x-axis) and standard deviation over the error bar for m=0 and m=1


Average accuracies and standard deviations as a function of training size with 10 folds for file amazon_cells_labelled.txt
m=0
m=1
Accuracy per Size
79.9
74.9
69.9
64.9
59.9
54.9
49.9
44.9
39.9
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Train Set Size


Average accuracies and standard deviations as a function of training size with 10 folds for file imdb_labelled.txt
m=0
m=1
Accuracy per Size
79.1
74.1
69.1
64.1
59.1
54.1
49.1
44.1
39.1
34.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Train Set Size



Observation:

1) It can be clearly seen that the accuracy of the classifier improves when smoothing factor is used(m=1). In m=0, the probabilities of the words not present in the class comes to 0 reducing the accuracy of the model.
2) When the training data is less than the test data, the accuracy of the model is very less compared to the case where the training data is more than sufficient for the test data.

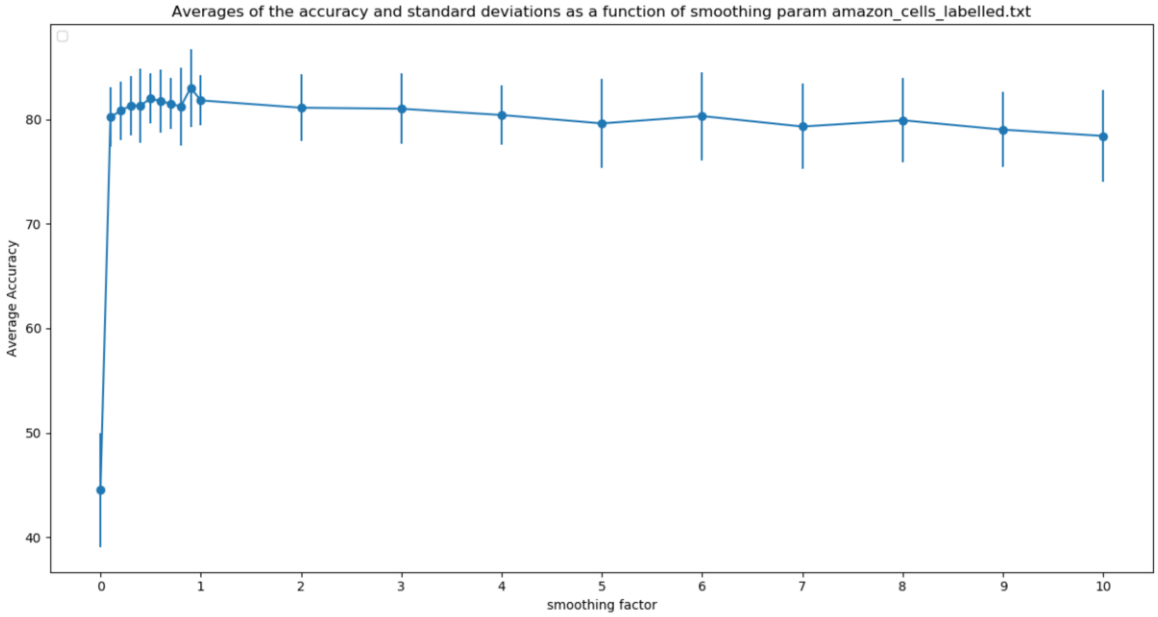
Experiment 2:

In this experiment, we ran stratified cross validation for Naive Bayes with smoothing parameter taking values as m = 0,0.1,0.2,...,0.9 and 1,2,3, ...,10 (i.e., 20 values overall).

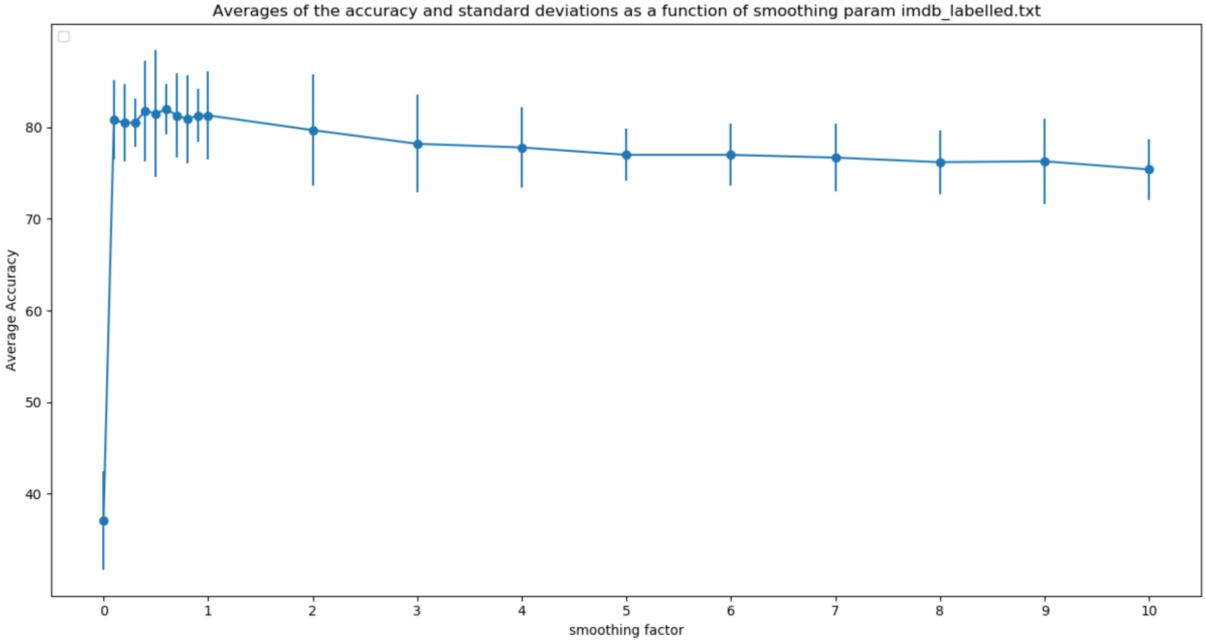
This was run against each below dataset:

1) Amazon
2) IMDB
3) Yelp

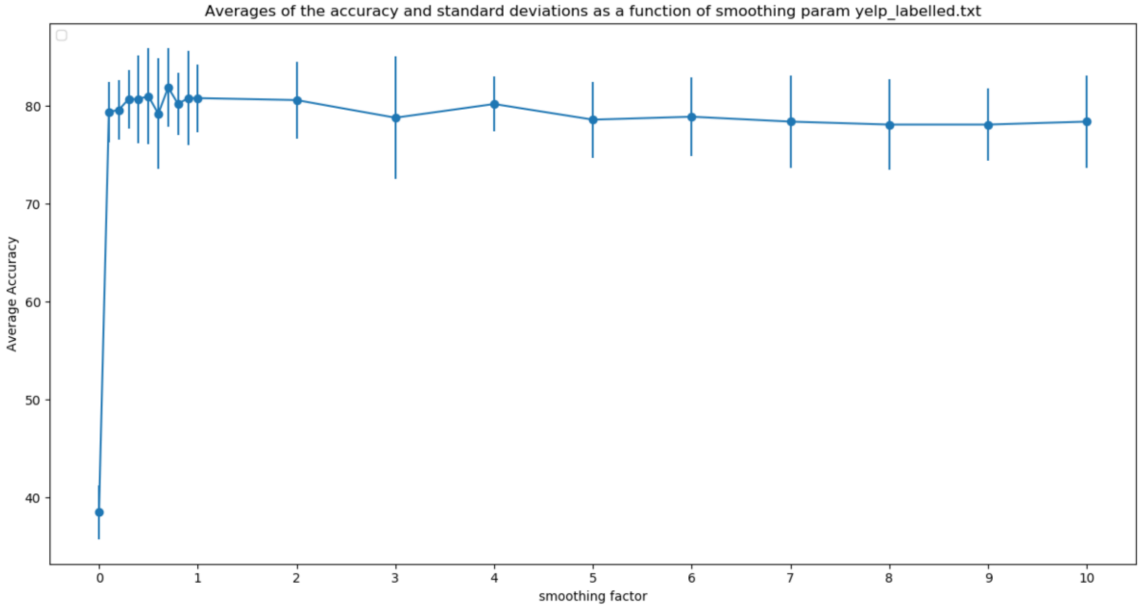
Below graphs have been captured for each dataset, and these graphs are plotted for Average accuracy(y axis) against smoothing factor (x-axis) and standard deviation over the error bar.



Averages of the accuracy and standard deviations as a function of smoothing param amazon_cells_labelled.txt



Averages of the accuracy and standard deviations as a function of smoothing param imdb_labelled.txt



Observations:

1) As the smoothing parameter increases, there is not much change in the accuracy of the classifier.
2) Accuracy of the model fluctuates in a small range when the smoothing factor is less and between 0 to 1.

Conclusion:

We have delved into building and understanding a Naive Bayesian Classifier in this assignment. This algorithm is well suited for data that can be asserted to be independent. Being a probablistic model, it works well for classifying data into multiple directions given the underlying score. This supervised learning method is useful for fraud detection, spam filtering, and any other problem that has these types of features.