Programming Assignment 1

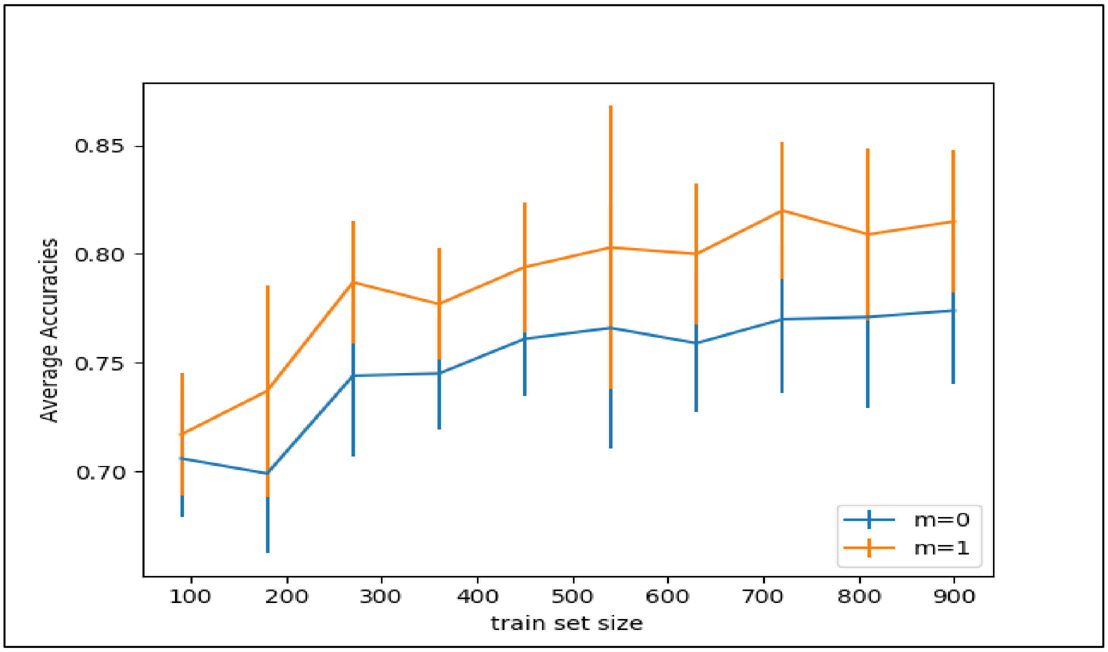
Machine Learning

Experiment 1:

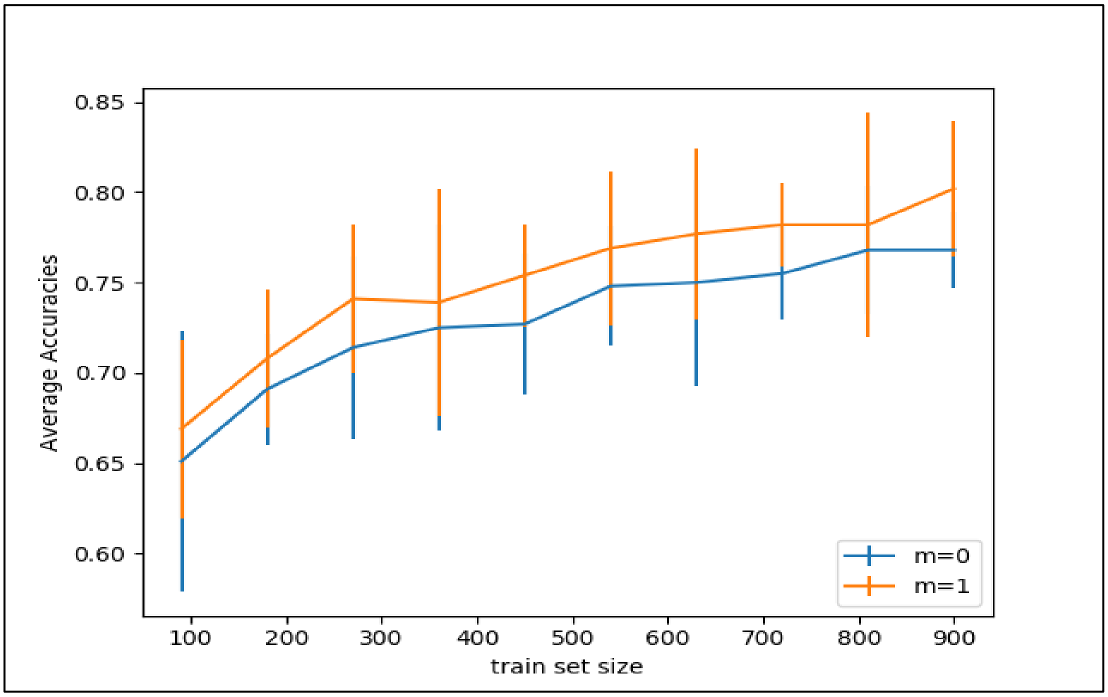
Results:

We have plotted error bar plots of average accuracy subsamples and standard deviation as a function of train set size for 3 data sets. The plots for the respective data sets are as below:

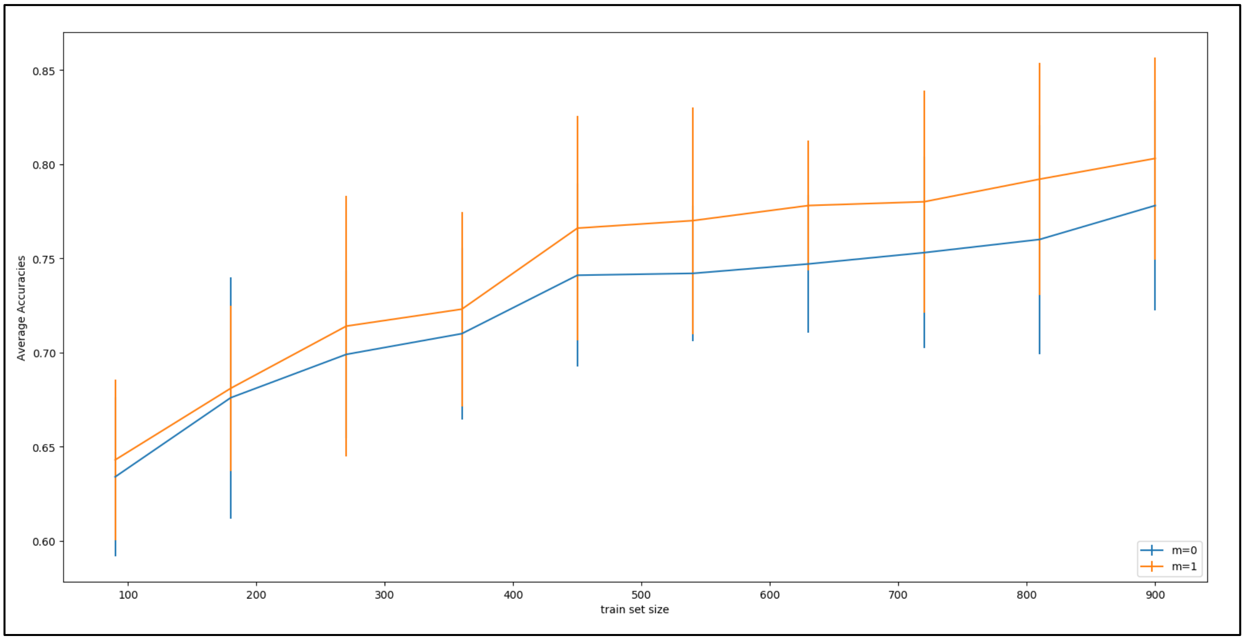
1. Amazon_cells_labelled



2. Yelp_labelled



3. Imdb_labelled



Observations:

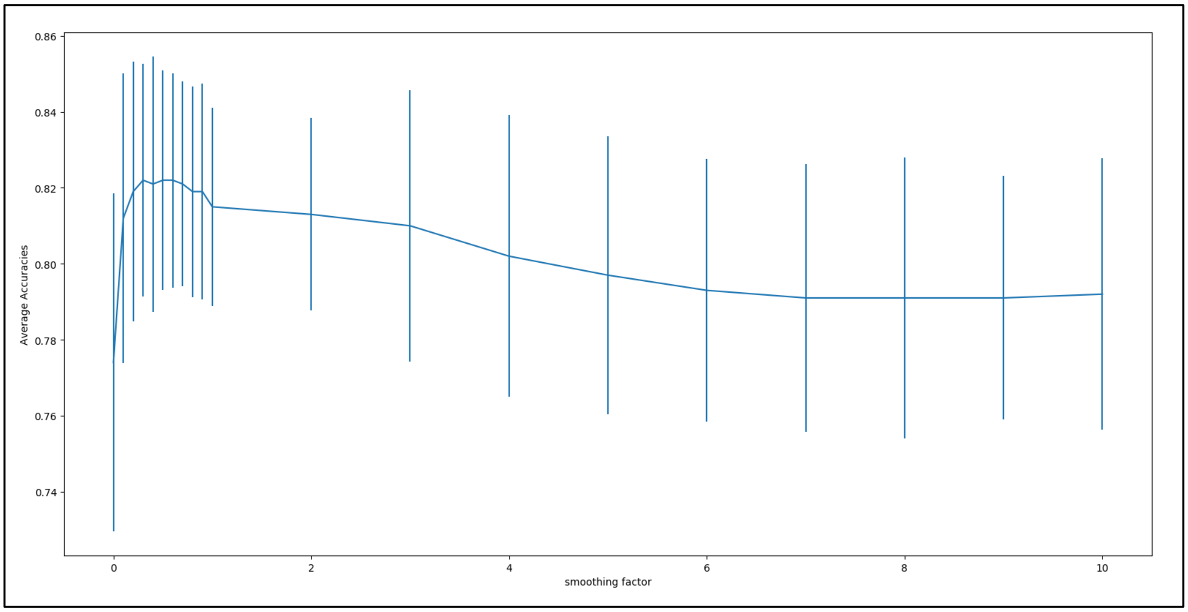
It is observed that as our training data size increases, the accuracy also increases. This can be attributed to the increase in no. of words in training data set, and thus the model is trained well giving better results with an increasing learning curve.

Experiment 2:

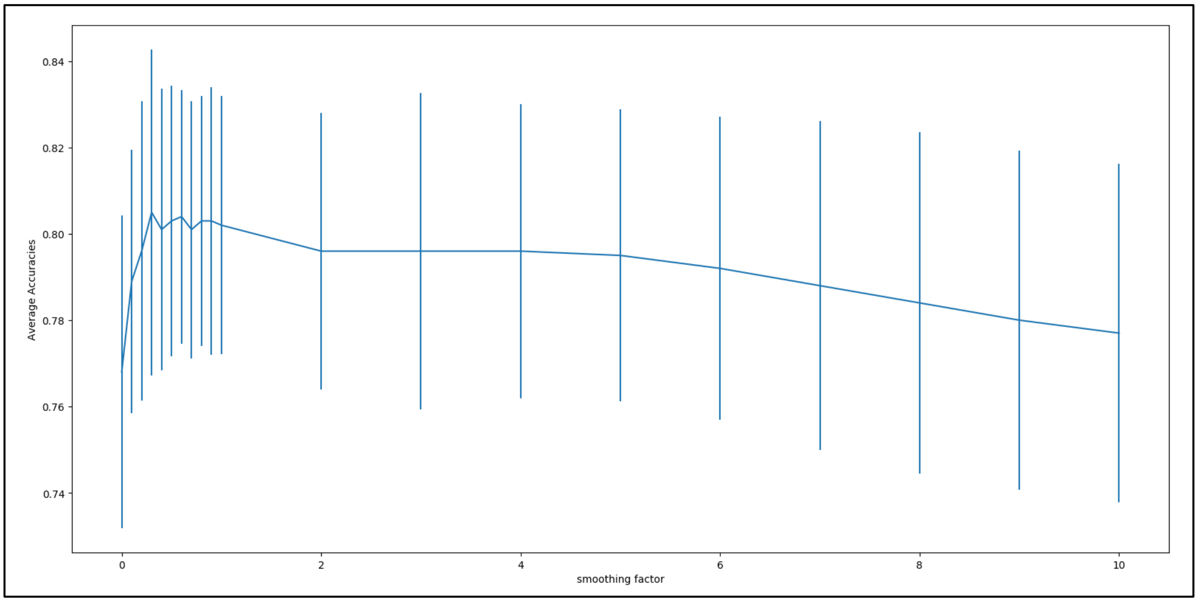
Results:

We have plotted error bar plots of cross validation accuracy and standard deviation as a function of smoothing parameter for 3 data sets. The plots for the respective data sets are as below:

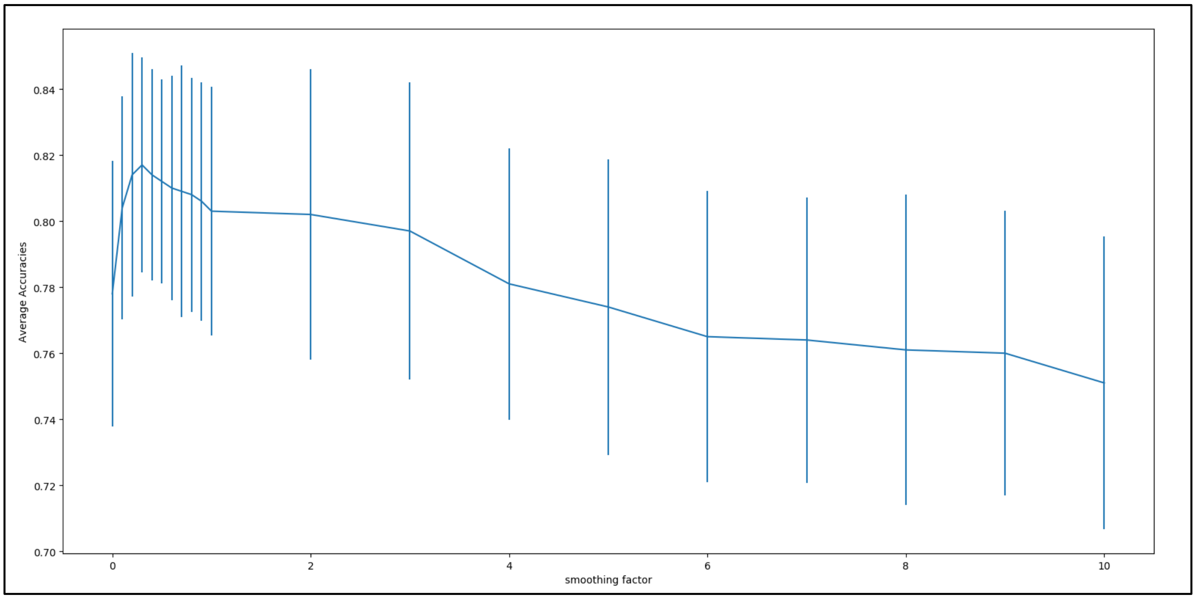
1. Amazon_cells_labelled



2. Yelp_labelled



3. Imdb_labelled



Observations:

One thing we can observe from the graphs is that at m = 0 has the lowest accuracy. This is because m = 0 mean no smoothing and thus many probability **word given class** will be zero resulting in loss of information, resulting in poor results. It is seen that we have better accuracy between 0 and 1, and as the smoothing factor increases the graph becomes stagnant.