Machine Learning Assignment: Sentiment Analysis using Naïve bayes!

This project was done in R Studio.
We were given textual data with sentiment of 3 datasets namely, amazon, IMdb and Yelp. We had to calculate the posterior probability after stratified cross validation on the dataset. We had to display our findings graphically for each of the dataset. I did the graphical work in R Studio. All of the work was done ONLY using Base R functions and no foreign libraries were imported for the tasks.

1) First of all, I did some textual preprocessing and cleaning before feeding the data for cross validation and modelling. This is a glimpse of the function that I created for it which uses some normal regex knowledge.

```
#Cleaning the textual data to remove extra noise
data_cleaning<-function(x){
  x[,1]<-tolower(x[,1]) #Convert to lowercase
  x[,1]<-gsub(pattern = "\\W", replacement = " ",  x[,1], ignore.case = TRUE) #Removed ext
  x[,1]<-gsub(pattern = "\\d", replacement = " ",  x[,1], ignore.case = TRUE) #Removed num
  x[,1]<-gsub(pattern = "\\b[A-z]\\b{1}", replacement = " ",  x[,1], ignore.case = TRUE) #
  x[,1]<-gsub(pattern = "\\s", replacement = " ",  x[,1], ignore.case = TRUE) #Removing WH
  return(x)
}
```

2) Amazon, Yelp and IMdb

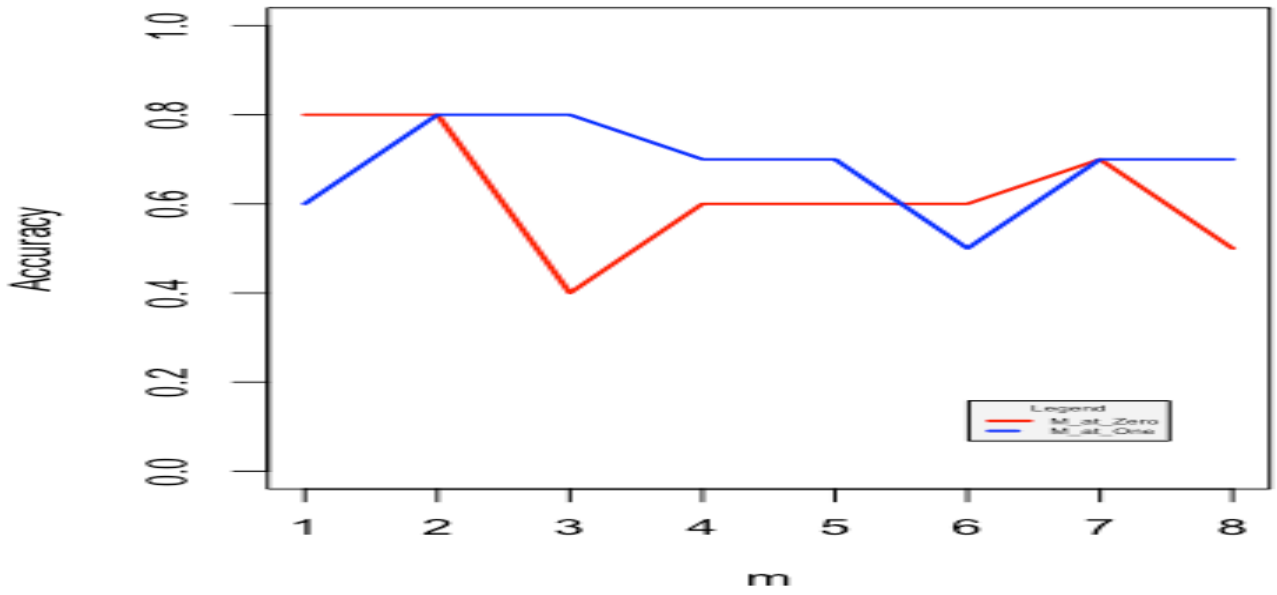
A) This is a glimpse of the Probability(Word/+) and Probability(Word/-) that I calculated during this experiment for all the 3 datasets.

AMAZON:

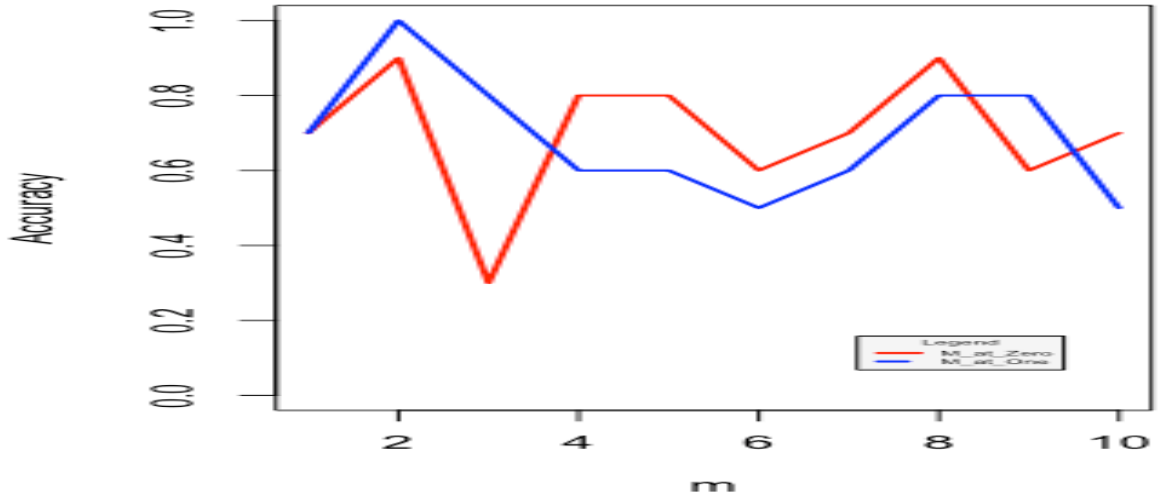
	Var1	Freq+	p(w/+)	Freq-	p(w/-)
1	about	0	0.000000000	3	0.006802721
2	above	0	0.000000000	1	0.002267574
3	ac	1	0.002375297	0	0.000000000
4	acceptable	0	0.000000000	1	0.002267574
5	according	1	0.002375297	0	0.000000000
6	advertised	1	0.002375297	0	0.000000000
7	after	0	0.000000000	2	0.004535147
8	ago	0	0.000000000	1	0.002267574
9	all	0	0.000000000	2	0.004535147
10	almost	0	0.000000000	1	0.002267574
11	also	3	0.007125891	0	0.000000000

B) Then We calculated the posterior probability and then the accuracy for Smoothening Parameter m=0 & m=1

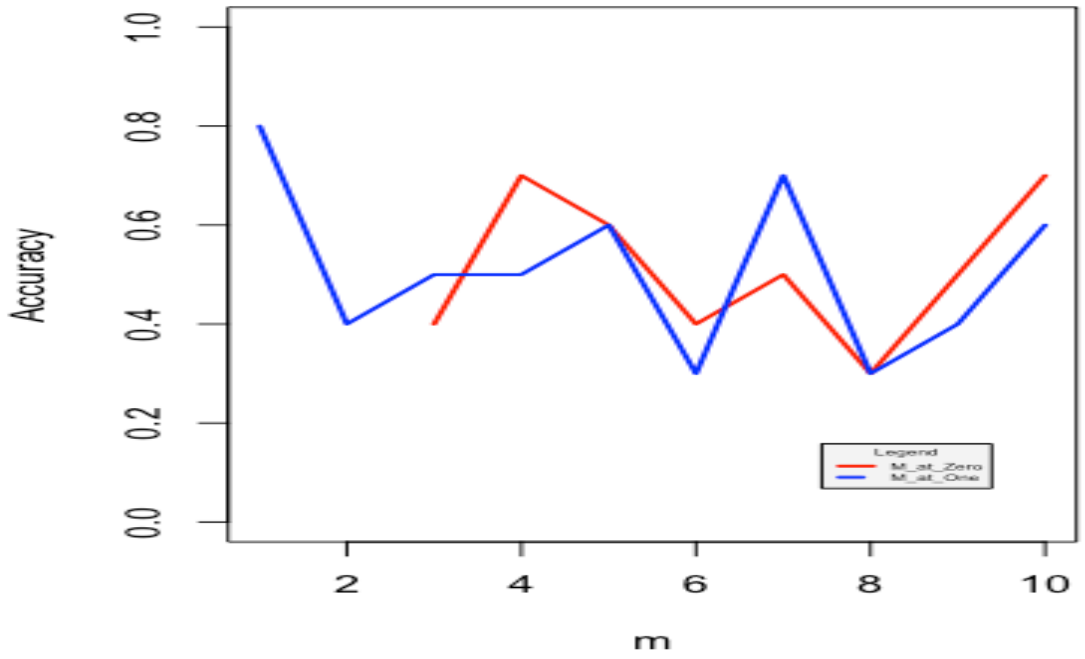
AMAZON



YELP



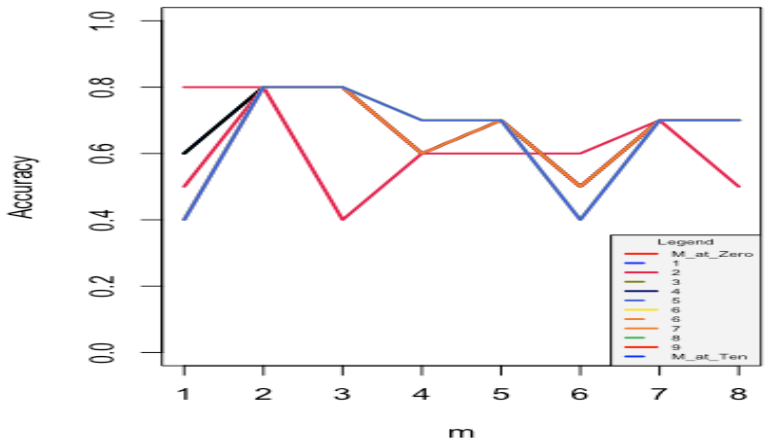
IMDB



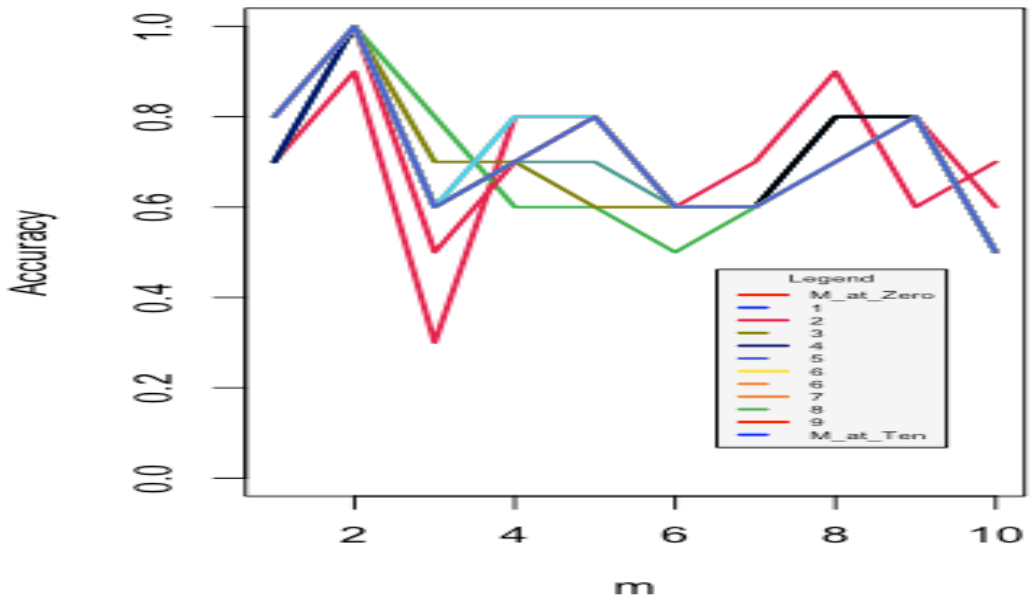
Here m is the train size. The red line is the accuracy plotted for M=0 and blue is for M=1. We can clearly and there is a clear trend that the accuracy is better for M=1 as compared to M=0

C) Then we plotted the plots of Accuracy vs the parameters for a number of parameters of m(1-10). The blue plot is for m=10 and red plot is for m=0. For most of the datasets blue line is considerably over the red line.

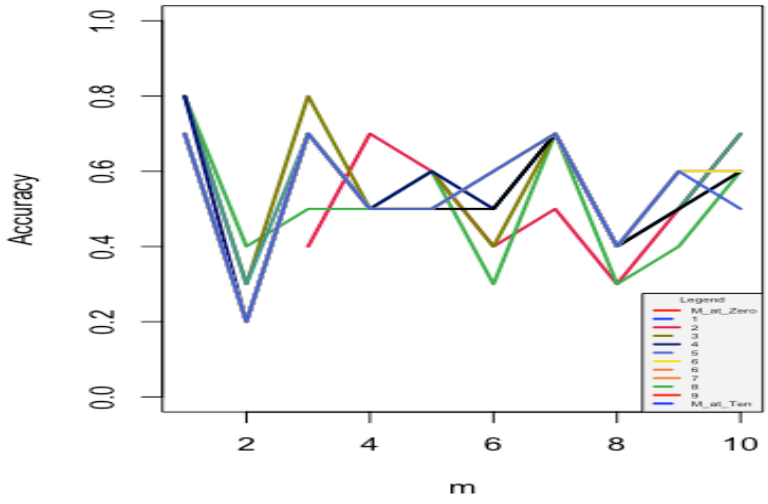
AMAZON



YELP

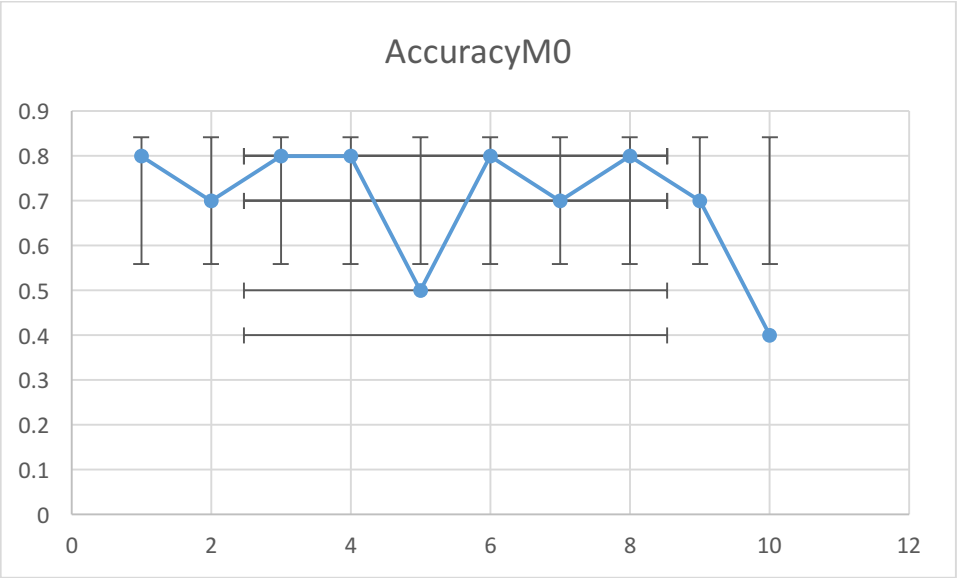


IMdb



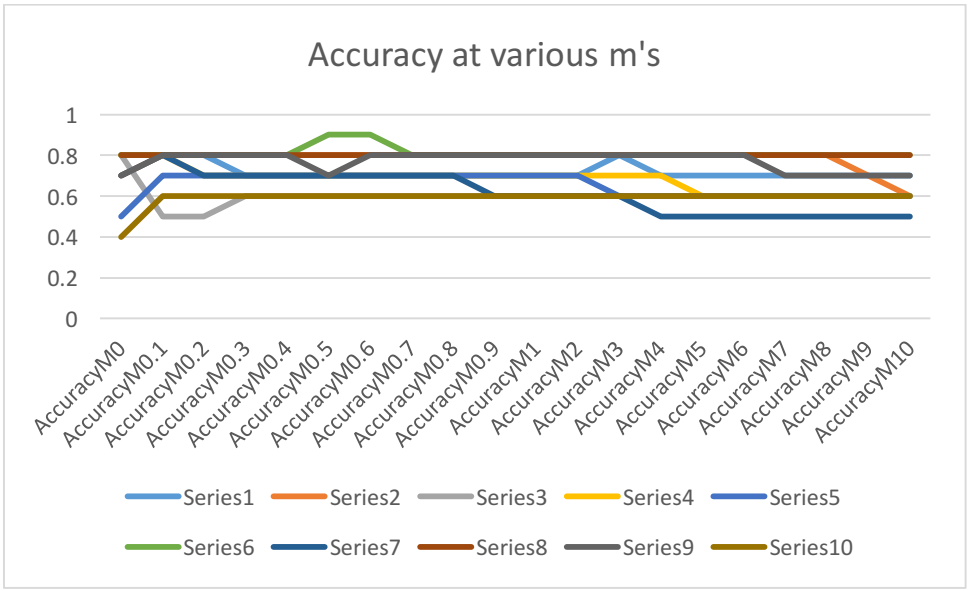
D) Then I calculated the standard deviation of the model for various accuracies. For this I imported the consolidated accuracies dataframe in excel and made the error bars there due to lack of flexibility in graphs in R for standard error.

AMAZON



This is the standard Error for M=0 for various train sizes which is SD/sq.rt (N)

This is the accuracies graph for m=0-1 and then 1:10. Personally I think this is too much information to show on one graph but we can see minor peaks and and troughs showing that that the deviation from the mean is not too much which is good news!!



The Accuracy of my model was close to 60%