Introduction-

We have used the Amazon dataset in the second question. The files given are amazon_baby_test.csv and amazon_baby_train.csv. Systematic experiment has been done by varying parameters to improve the generalization of performance.

a. What is the number of attributes in each dataset?

There are three attributes in this dataset they are- Product, Review and Rating.

b. What is the number of observations?

The observations are the ratings possible that is 1-5.

c. What is the mean and standard deviation of each attribute?

The **mean** and **standard deviation** for the ratings attribute is **4.12 and 1.29** respectively.

Design-

The amazon data set we need to analyze the reviews based on the ratings given.For this dataset,we will use the bag-of-words model to do Sentiment Analysis. The bag-of-words model can perform quiet well at Topic Classification, but is inaccurate when it comes to Sentiment Classification. The reason to still make a bag-of-words model is that it gives us a better understanding of the content of the text and we can use this to select the features for the classifiers. For the training dataset,we have **36,707** records and have reviews ranging from 1 till 5.As the next step, we are going to divide the corpus of reviews into a training set and a test set. In order to also take into account the effects of the training set size on the accuracy of our model, we will vary the training set size from 1.000 up to 20.000.The bag-of-words model is one of the simplest language models used in NLP. It makes an unigram model of the text by keeping track of the number of occurrences of each word. A simple improvement on using unigrams would be to use unigrams + bigrams. That is, not split a sentence after words like “not”,”no”,”very”,“just” etc. It is easy to implement but can give significant improvement to the accuracy. The sentence “This book is not good” will be interpreted as a positive sentence, unless such a construct is implemented. Another example is that the sentences “This book is very good” and “This book is good” will have the same score with a unigram model of the text, but not with an unigram + bigram model.

Implementation:

Initially we are extracting all the reviews based on ratings which are more than 3.If the rating is more than 3, then it is a positive rating otherwise, the rating is negative. Then we group by the rating and reviews for the movies having positive reviews and plot it through MATLIB. Then we preprocess the data and use NTLK library to find all the stop words and apply it to remove them from the dataset. We apply this procedure for both the positive as well as negative reviews. Then we get all the common 5000 words that are used for our reviews and the number of times it occurred in our dataset. Neural Learning

algorithm was applied using the MLPClassifier from the sklearn.tree library. Multi-layer Perceptron classifier model optimizes the log-loss function using LBFGS or stochastic gradient descent. The accuracy for our training dataset is found out along with prediction and fscore.
The accuracy for our training dataset is found out along with prediction and fscore and is found to be **99.992** percent. The accuracy for the same is repeated for our testing dataset and the accuracy is **60.188** percent for the testing dataset without any pruning