**Lab 4 Report**

**Introduction**

The purpose of this lab is to implement two text classifiers with Naïve Bayes and Logistic regression models. The performance of the models was then analyzed by looking at a combination of metrics such as accuracy, precision and recall. The effect of text normalization such as lemmatization and stemming were also analyzed to study their effect on the performance of each model.

**Methodology**

The models were trained, using python libraries, on a dataset of labelled reviews from Amazon, Yelp and IMDB. There were three sets of data, from each company, with each data set containing 1,000 documents. This makes a total of 3,000 documents for training and testing.

The data was randomly separated such that 80% of the data was used to train the models and 20% was used to for testing. After evaluation the performance of the models, they were trained one more time with all the data and the models were persisted using python’s pickle library.

**Libraries**

The scikit-learn library was used to implement the models because they it has all the features required for this lab. Also, it is relatively popular and easy to use and the creators of the library provide a lot of content on how to use it’s the different parts. This makes it the ideal choice because it has the required features and the learning curve is relatively shallow due to the amount of useful content on it.

The NLTK library was used in conjunction with scikit-learn for text regularization, specifically lemmatization and stemming. NLTK also has features that can be used to implement Naïve Bayes models, however, it doesn’t provide an intuitive interface for that which makes it difficult to use. Using it for lemmatization and stemming was straight forward.

**Performance Evaluation**

As mentioned earlier, the performance of the models was evaluated by analyzing the accuracy of the model as well as precision and recall.

<Discussion of results>