In this lab exercise, I developed two sentiment classifiers: Naïve Bayes and Logistic Regression classifiers. These classifiers were developed using Python libraries and their main function is to predict whether a sentiment or comment placed on a product as a review is ether positive or negative. I used the *pandas* library since it provides fast and flexible ways of managing data especially related or labelled data. In this case, our dataset was related since each review was linked to either 1 or 0 which represented a positive or negative review respectively. This library also allows easy manipulation of data when grouped together, for example getting the count of either positive reviews or negative reviews based on a label (1 or 0). I used the *nltk (Natural Language Toolkit)* librarybecause it has a set of natural language processing tools and datasets such as *stopwords* which allows easy and fast way of text normalization intended for building prediction classifiers. I also used the *sklearn* library because it has most of the standard machine-learning tools and algorithms which assist in producing tasks such as classification and regression for faster data processing.

I trained each of the two classifiers on two versions: *n* for normalized text and *u* for unnormalized text. With the training data being a combination of all the datasets provided (train\_data.txt), while using the Naïve Bayes on normalized text and unnormalized text, I got an accuracy of 0.79 (79%) and 0.77 (77%) respectively. While using the Logistic Regression classifier on normalized text and unnormalized text, I got an accuracy of 0.80 (80%) and 0.74 (74%) respectively. From these results, it can be inferred that:

1. The Logistic Regression classifier has a higher accuracy (80%) than the Naïve Bayes classifier (79%) when the text is normalized because the Logistic Regression classifier extracts features easily from normalized text than the Naïve Bayes classifier.

The Naïve Bayes classifier has a higher accuracy (77%) than the Logistic Regression (74%) when the text is not normalized. This can be attributed to the fact that the not-normalizing of the text makes it hard for the Logistic Regression to map out the features from the dataset for classification.

1. However, this does not mean that Naïve Bayes is better than Logistic Regression because we cannot use *accuracy* as a measure of comparing which model is better than the other. Logistic Regression needs a large data set to improve its accuracy which will later be better than the Naïve Bayes since it will be able to extract more features from more data therefore making better predictions