**Introduction:**

The second project is based off of Natural Language Processing. We used a dataset containing the yelp reviews of various services. The Sentiment field contains 2 values: 1 for negative sentiments(1 or 2 stars), while 2 represents positive sentiments(3 or mores stars).

**Preprocessing:**

There were no preprocessing conducted except to equalize the number of both outcome in the used data set. This was done in order to prevent the bias of the machine learning model towards any particular outcome, as it would leave it unusable even if it had a lower error rate.

Cleaning of the comments is taken as part of project and has been explained in the ‘Project Summary’ section.

**EDA:**

No EDA was performed since we are only using the comments to predict the gender, and the remaining data is ignored.

**Project Summary:**

The most important part of the project was the cleaning of the data.

The nltk and string packages for used for this process. A nested loop structure was used to clean the data, where the outer the loop was used to extract each comment and consequently tokenize the words in the comment. The inner loop is responsible of cleaning the comment by checking word by word if any were stop words, after finding their root-word using the PortStemmer() inbuilt function in nltk. The inner loop was also responsible to remove the punctuation marks in the comments.

The cleaned comments were then added to the data frame and were used to create a sparse matrix using sklearn feature CountVectorizer(). The sparse matrix and the cleaned comments were then used to train a multinomial naive bayes theorem model MultinomialNB() using the fit() function. Since the target feature was discrete, we used accuracy\_score() to measure the accuracy of the trained model.

The average accuracy was found to be around 85% for using 50000 records.