MSCI-641 Assignment 2 Report

Name: Nauman Ahmed

Student ID: 20743448

Stopwords removed	text features	Accuracy (test set)
yes	unigrams	80.47%
yes	bigrams	78.98%
yes	unigrams+bigram	82.10%
no	unigrams	80.83%
no	bigrams	82.58%
no	unigrams+bigram	83.31%

a) **With stopwords performed better**. Sentiment analysis tasks are sensitive to stopwords. This is because removing stopwords like ‘no’, ‘not’, ‘against’ from a corpus change the meaning of the text. If a customer wrote ‘I am not happy’, this indicates a negative review. Removing stopwords would make the text ‘happy’ and this indicates a positive review. Therefore, this will lead to the text being mis-classified.

b) **‘Unigrams + Bigrams’ performed the best**. Including bigrams in the prediction adds some context to the data as it captures the word relations. However, using bigrams alone makes the dataset matrix sparse as there is lower probability of bigrams repeating in different documents. This increases the probability of overfitting the model. Therefore, by including unigrams as well the dataset becomes less sparse and this prevents overfitting. The results are, therefore, better when both unigrams and bigrams are included.