REPORT

Name: Dishant Mittal

Student Id: 20710581

Stopwords removed	Text features	Accuracy (test set)
yes	unigrams	0.7316625
yes	bigrams	0.673075
yes	unigrams + bigrams	0.755625
no	unigrams	0.731325
no	bigrams	0.7536875
no	unigrams + bigrams	0.7666375

Q1) Which condition performed better: with or without stopwords? Write a brief paragraph (5-6 sentences) discussing why you think there is a difference in performance.

Ans) With stopwords performed slightly better than without stopwords.

In nltk library words like ‘not’ and ‘can’ are treated as stopwords. In a text sentiment classification related task these words are important for embedding the sentiment in the word vectors which are given as an input to the model.
For example, consider the sentence: ‘I do not like this car’. This has negative sentiment attached to it. But if we exclude the token ’not’ in the analysis, the overall meaning of the resulting sentence is positive, which is not at all the reality. So, removing stop words can be problematic if context is affected.

Q2) Which condition performed better: unigrams, bigrams or unigrams+bigrams? Briefly (in 5-6 sentences) discuss why you think there is a difference?

Ans) unigrams + bigrams performed the best.

In unigram-only approach the weights corresponding to each word are considered separately which is one of the way to build the machine learning model. However, we can potentially add more predictive power to our model by adding two word sequences (i.e. bigrams). For example, if a review had the three word sequence “didn’t love movie” we would only consider these words individually with a unigram-only model and probably not capture that this is actually a negative sentiment because the word ‘love’ by itself is going to be highly correlated with a positive review. But, if we also add the bigram features to the model then we can take into account ‘didn’t love’ as a unit which can clearly inform the model that this is most likely a negative review. Thus, adding bigram features along with unigrams would have helped the model to improve the accuracy.