MSCI 641 Assignment -2

Name: Sagar Kulkarni

Student ID: 20767929

Stopwords removed	Text features	Accuracy (test set)
Yes	Unigrams	**0.8074**
Yes	Bigrams	**0.8063**
Yes	Unigrams + bigrams	**0.8239**
No	Unigrams	**0.8076**
No	Bigrams	**0.8281**
No	Unigrams + bigrams	**0.8314**

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.

The model performed better with the stop words because we are performing a rudimentary form of sentiment analysis. This means that the context of the statement would matter a lot and the stopwards provide the meaning for the sentence. Stopwords like ‘not’, can change the meaning of the sentences. For eg: “I do not like the movie” without ‘not’ is “I do like the movie” which is positive, but the label given to this sentence in training set is negative and hence causes the context to be missed. On the other hand, removing the stopwards speeds up the process as this reduced the number of words the Countvector must fit. This is a case of speed vs accuracy trade-off.

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

Unigrams + bigrams performed better than either one separately. And bigrams performed better than unigrams for the case when stopwords were not removed. This may be due to the bigrams preserving the relationship between words and get a better sense of meaning of the sentence. But unigrams and bigrams perform similarly in the case when stopwords were removed. This would again largely be a result of the stopwords being important to preserve the context of the sentence. Eg: “not like” will not exist in the bigram set without stop words and hence the sentence loses its semantic meaning.