MSCI 641: ASSIGNMENT 2

QUESTION 2:

a) Which condition performed better: with or without stop words:

a. While stopword removal is expected to increase the accuracy in prediction, only a small variation is observed in the accuracy before and after the stopwords. In fact, removal has decreased the accuracy by 2.16, 0.0.06469 and 0.036 for unigram, bigram and uni+bigrams respectively. Hence the conclusion is that the data set with stopwords performs slightly better than the data without the stopwords.

b. Reason: (based on intuition + how I coded the removal of stopwords)

i. The first reason could be that, since we are dealing with the sentimental analysis of the data, there is a good chance that common adverb describing the sentiment of the review could have been removed.

ii. In addition, the list of stopwords may also contain some of the very commonly used words that pertain to the sentiment of the sentences which, upon elimination may reduce the accuracy.

iii. An input of “I did not like the car” resulting in [‘like’,’car’] will give a bad accuracy.

b) Which condition performed better: unigrams, bigrams or unigrams+bigrams?

a. The accuracy dropped by 0.043125 from unigram to bigram, 0.008125 from uni+bigram to unigram and 0.05125from uni+bigram to bigram (3.75E-05, 0.008125, 0.0224375 with stopwords). Precisely, the order of accuracy (with or without stopwords) is Uni+Bigrams > Unigram > Bigrams.

b. Reason:

i. One reason bigram has a poor accuracy could be because of data sparsity. The output of CountVectorizer is a BoW which does not have an order and hence is very much possible that combination of words may not give the right context about the data. Hence, we see slightly better accuracy in Unigram where the words are random, disordered and considered individually.

ii. Now, with unigram that gives better accuracy than bigrams, we combine both which helps the model understand the context of the bigrammed words better.

Stop Words Removed	Text Features	Accuracy (Test Set)	Accuracy %
Yes	Unigram	0.786875	78.6875
Yes	Bigram	0.74375	74.375
Yes	Unigram+Bigram	0.795	79.5
No	Unigram	0.808475	80.8475
No	Bigram	0.8084375	80.84375
No	Unigram+Bigram	0.830875	83.0875