MSCI 641: Assignment #2

Dan Wang 20796378
d347wang@uwaterloo.ca

University of Waterloo — June 5, 2019

Question1: Accuracy of different models
See the table 1.

Table 1:

Stopword removed	text features	Accuracy(test set)
yes	unigrams	0.8060
yes	bigrams	0.7881
yes	unigrams+bigrams	0.8238
no	unigrams	0.8045
no	bigrams	0.7857
no	unigrams+bigrams	0.8213

Question 2:

(a) Datasets without stopwords performed better than other datasets. We can see from the above table that the classification performed on dataset without stopwords consistently have better performance, achieving a slightly higher accuracy score than those with stopwords, which is not a coincidence. However, the difference is not prominent. When using count vectorizer, removing stopwords is great as it lowers the dimensional space and also a few stop words won't drive the analysis. On the other hand, when we are exploring the semantics of the given text, removing stopwords will omit the context and we will end up with ambiguous results.

(b) As is shown in the table, unigrams bigrams inputs produce better results. And sole bigrams input produce the worst. It's not common for bigrams to perform worse than unigram. Here, the reason might be that longer n-grams will be rarer, leading to higher idf values. And in particular, adding extra features may lead to overfitting. So bigrams perform the worst. On the other hand, some meaningful phrases may consist of more than one word and have an opposite meaning to the content. Thus by considering some bigrams, we are allowed to make use of the context to help with semantics analysis. That might be the reason why the mix of unigrams and bigram perform better than only unigrams or bigrams.