Assignment2 Report

Student Name: Guohao Ma

Student Number: 20676560

The classification accuracy is listed in the following table:

Table 1. Multinomial Naïve Bayers classifier accuracy with or without stopwords

Stopwords removed	Test features	Accuracy
Yes	unigrams	80.65%
Yes	bigrams	82.43%
Yes	unigrams + bigrams	83.14%
No	unigrams	80.74%
No	bigrams	82.42%
No	unigrams + bigrams	83.18%

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

A1: Nearly for all the three cases (unigrams, bigrams or both), the Multinomial Naïve Bayers gives a better result on the dataset without stopwords. The reason behind this is that stopwords do not have too much meanings and indications. The words after stopwords may have a random distribution and will decrease the accuracy if included. However, we can see the difference is not too much with or without stop words. That's because the stopwords will have the similar effect, though small, on positive and negative examples.

Q2: Which condition performed better: unigrams, bigrams or unigrams+bigrams? Briefly discuss why you think there is a difference?

A2: The combination of unigrams and bigrams performs better than each one alone, with or without stopwords. And bigrams seem to be better than unigrams alone. Models built on unigrams takes the randomness of each word into consideration but ignore the relation between each word within context. Models built on bigrams include features like collocations and also consider some semantic structures. That's why bigrams perform better than unigrams by 1.5%. However, the combination of unigrams and bigrams beat each of them and improve accuracy by 0.7%, compared to bigrams. That accounts for the advantage of having the strengths of both.