Assignment 2

Fuat Can Beylunioğlu
20803408

June 7, 2019

1.

Stopwords removed	Text features	Accuracy (test set)
No	Unigrams	0.807175
No	Bigrams	0.830288
No	Unigrams+Bigrams	0.832438
Yes	Unigrams	0.806100
Yes	Bigrams	0.808150
Yes	Unigrams+Bigrams	0.824338

2.a) The model performs better on non-cleaned data. The reason is probably due to some stopwords that are very indicative when classifying into sentiment categories. Namely negation words such as ”no”, ”not” and ”but” can easily change the direction of the comment sentiment. Even though the difference reduces when unigrams and bigrams are used, it is still lower. Since Naive Bayes has conditional independence and BOW assumptions, information that any token contains can increase accuracy separately, therefore absence of such tokens, which are directly linked with sentiments, reduces the accuracy significantly.

2.b) Bigrams perform better than Unigrams and Unigrams+Bigrams model performs the best. Since Multinomial NB assumes conditional independence and BOW, it cannot capture patterns such as ”no good”, ”don't like” etc. Bigrams cover this deficiency to a degree and allow for such negations. When there is no stopword that can negate the meaning, however, bigrams perform similar to unigrams since it doesn't capture any pattern; but it can only increase in performance when model is augmented with unigrams. Whereas when there are stopwords, bigrams perform as well as Unigrams+Bigrams model because bigrams include enough information about negations to decently classify documents and making the information from unigrams be redundant.