**MSCI 641 Assignment -2**

**Name: Sagar Kulkarni**

**Student ID: 20767929**

Table:

|  |  |  |
| --- | --- | --- |
| Stopwords removed | Text features | Accuracy (test set) |
| Yes | Unigrams | **0.7852** |
| Yes | Bigrams | **0.7291** |
| Yes | Unigrams + bigrams | **0.7893** |
| No | Unigrams | **0.7883** |
| No | Bigrams | **0.7745** |
| No | Unigrams + bigrams | **0.7975** |

**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 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.

**Whichcondition 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.