Assignment 2 Report

-Atish Telang Patil; 20793593

1. **Accuracy Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| Stopwords removed | text features | Accuracy (test set) | Accuracy (validation set) |
| Yes | Unigrams | 82.97% | 75.84% |
| Yes | Bigrams | 79.74% | 73.59% |
| Yes | Unigrams + Bigrams | 84.30% | 77.01% |
| No | Unigrams | 83.16% | 76.34% |
| No | Bigrams | 83.26% | 78.67% |
| No | Unigrams + Bigrams | 84.91% | 79.26% |

1. **According to the results with stopwords condition performed slightly better than without stopwords condition.**

The reason for this is after we remove the stopswords the meaning of the sentence (or semantic) changes for few reviews.

For Eg: **The dress I got was not beautiful.** In this case if we remove the stops words the sentence becomes ‘**dress’, ‘got’, ‘beautiful’.** The classifier thinks this as a positive review and misclassifies.

If the stopwords are not removed than the sentence is correctly classified.

1. **Unigrams + Bigrams condition performed better in both the cases (with and without stopwords).**

In our case without stopwords Unigrams performed better than bigrams as the bigrams generated are rarely used in the corpus hence making it not efficient when compared to the unigrams.

A combination of Unigrams and Bigrams gave a better accuracy as it creates a more features for the classifiers to classify improving the efficiency.

In case with stopwords Bigrams performed slightly better than Unigrams as the frequency of the bigrams were more in the corpus.

The combination of Unigrams and Bigrams performed better than both the cases as it had more features and helped the classifier to classify the labels correctly.