Assignmet2-MSCI641 Report

Results

Stop words removed	Text Features	Accuracy(test set)
Yes	unigrams	80.88%
Yes	bigrams	81.45%
Yes	unigrams + bigrams	82.90%
No	unigrams	80.70%
No	bigrams	82.84%
No	unigrams + bigrams	83.12%

- The highest accuracy is achieved with the presence of stop words and unigrams + bigrams as the text features.

(a) The results(accuracy) are better with the presence of stop words. Even though, stop words do not contain meaning in themselves, increase the size of the vocabulary, increase complexity and training time of the algorithm, their presence is essential in tasks like sentiment analysis. Consider an example of sentence : “She was not happy”. Now, suppose we have two classes “happy”(0) and “sad”(1). Without the removal of stop words , the sentiment would be “sad”(1) however removing the stop word would miss-classify the sentiment as “happy”(0) because now the token would only be [‘happy’]. Therefore, removing stop words would immensely dent the accuracy of sentiment analysis. However, in some tasks like topic classification, stop words can be removed because they have no impact on the outcome of a theme/topic of text. My results are consistent with the above explained analysis as the accuracy with stop words is higher as compared to accuracy without stop words.

(b) The results(accuracy) are better with the combination of unigrams and bigrams as tokens. The problem with unigrams only is that they do not capture the relationships between words or context of a particular word. For example in the case of “not happy”, the unigram model will capture “not” and “happy” individually like a bag of words approach but bigram model will capture “not happy” as (‘not’, ‘happy’) combined hence preserving the relationship between the two words. In this way, the classifier recognizes that happy is being preceded by not so it means unhappy which was difficult to capture in unigram model because the context of “not” would have been unclear(it could have occurred with any other token in the sentence other than happy). Bigrams alone would also not produce optimal results because to find specific bigrams in a text is highly unlikely therefore they would result in a sparse matrix representation. Bigrams alone would also result in poor generalization of the results because the model will train on specific examples of bigrams which are difficult to find in the test set-*overfitting*. While only unigrams would result in *underfitting* and bigrams alone would result in *overfitting* ,both of them combined should give good results as combined the relationship between words would also be preserved plus there would be far lesser sparsity in the matrix representation/lesser number of specific occurrences of tuples. My results are consistent with the above explained analysis as the accuracy with unigrams and bigrams combined is the maximum as compared to unigrams and bigrams only.