# Assignment 2
# MSCI641 Report

| Stop words removed | Text Features | Accuracy(test set) |
|---|---|---|
| Yes | unigrams | 72.31% |
| Yes | bigrams | 72.32% |
| Yes | unigrams + bigrams | 75.45 % |
| No | unigrams | 72.46% |
| No | bigrams | 75.12% |
| No | unigrams + bigrams | 76.08% |

## Q1. Which condition performed better: with or without stop words, why is there a difference?

The highest accuracy is achieved with the presence of stop words and unigrams + bigrams as the text features.
The accuracy is overall better with the presence of stop words. Stop words do not contain meaning in but their presence is adding weight to sentiment analysis. Consider an example of sentence:
**" This movie is not good."**
If I remove (not) in Pre-processing step the sentence (movie good) indicates that it is positive which is false. Therefore, removing stop words would lower the accuracy of sentiment analysis. topic classification, stop words can be removed because they have no impact on the outcome of a topic of text.

## Q2. Which condition performed better: unigrams, bigrams or unigrams + bigrams? Discuss why you think there is a difference?

The accuracy results are better with the combination of unigrams and bigrams as tokens. This is because when unigrams are used in a silo they will not capture the contextual relationships between words or context of a particular word. For example, in the case of “not good”, the unigram model will capture “not” and “good” individually like term frequency approach but bigram model will capture “not good” as (‘not’, ‘good) combined hence keeping the contextual relationship between the words. In this way, the classifier recognizes that good is being preceded by not so it means that the movie was bad which wouldn’t be possible in unigram model.

With the combinations of unigram + bigram we are adding more redundancy to the dataset, which increases the term frequency of the word giving it more predictive power. Individual unigrams would be learnt along with bigrams making a more robust classifier.