Text Classification for Sentiment Analysis

The task is to classify the given hotel reviews as positive or negative using the Naïve Bayes classification.

Approach: Naïve Bayes with add-1 smoothing

I extracted the log prior and log likelihood information and used the Naïve Bayes approach with add-1 smoothing to classify the reviews into one of the classes by assigning them to the highest probable class.

Modified Approach: Naïve Bayes with add-1 smoothing and stop words list

In order to improve the accuracy, I used a stop word list and disregarded their occurrences in the test set when computing the probability using Naïve Bayes with add-1. Though this approach was used to increase the accuracy, the accuracy value did not change from the base approach. But, if the training and testing data is considerably huge, then using stop words might lead to an increase in accuracy compared to the Naïve Bayes approach with just add-1 smoothing.

Improved Approach: Binary multinomial Naïve Bayes

Here, the accuracy is improved by taking into account the fact that, in text classification tasks, occurrence of a word in a document is important more than its frequency. Implementing this, the algorithm's accuracy improved to some extent.

The Naïve Bayes approach gave an accuracy greater than 90%. To train the model, a k-fold validation technique was used. The accuracy varies greatly over the validation technique but is expected to give an approximate accuracy of over 90%.