MACHINE LEARNING (BITS F464)
FIRST SEMESTER 2017

Document Classifier with Naive Bayes and extensions

	Positive Precision	Positive Recall	Positive F1	Negative Precision	Negative Recall	Negative F1
Basic Naive Bayes	**85.89%**	**75.09**	**.80131**	**77.87%**	**87.66**	**.8248**
After Removing Stopwords	86.5	77.36	0.8170	79.54	87.99	0.8355
Binary Naive Bayes	**87.23%**	**77.32%**	**0.8198**	**79.63**	**88.68%**	**0.8391**
After Removing Stopwords	87.18%	79.32%	0.8306	81.036	88.336	0.8452

Removing stop words from the Naive Bayes implementation causes a small but perceptible boost in the observed Precision/ Recall values. One possible reason for this could be because of the basic assumption of the naive bayes classifier that each attribute is conditionally independent of one another.

However, the presence of stop words (often prepositions) are almost entirely dependent on the context of the usage of the word, and hence dependent on the words that appear in close proximity to it.

The application of Binary Naive Bayes produces increased results because in the case of classifications like sentiment analysis of movie reviews, the number of times a word like “Bad” occurs is irrelevant. Even if there is one occurrence of Bad in the review, it is likely that it is a negative result. Hence making it a 0-1 not present-present sort of a situation removes redundancies which could lead to false positives or negatives.

Further removing stop words further removes intra-attribute dependencies and slightly improves the results.