Introduction

This report describes the implementation of a Naïve Bayes classifier in the context of classifying movie reviews into positive and negative categories. After stemming the terms within the corpus of reviews, the NB classifier adopts the Bernoulli model. The terms' weights are then smoothed to account for the zero counts problem, resulting in significant improvements in $F1$ performances. All reported performance measurements are acquired by averaging on a 10-fold-cross-validation.

The first section discusses the adoption of the Bernoulli model, as well as the design considerations and the choice of parameters. The second section addresses some observations and possible performance boosts, i.e. the removal of stop-words as well as feature selection. In both cases performance is compared to that of the basic classifier.

Our Model

Statistical model

The Bernoulli model is implemented to assign weights for terms within the corpus of reviews. The following Maximum Likelihood Estimation formulas is used to train the classifier:

$$P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}, \tag{2.1}$$

$$P(c) = \frac{N_c}{N} \tag{2.2}$$

Where T_{ct} denotes the count of the term t in class c, N_c is the number of data-points in class c, and N is the total number of data-points. In this case, c stands for either $positive$ or $negative$ movie reviews.

The Porter Stemmer is then applied to the terms in the training data, and subsequently to the new documents to be classified.

Smoothing

Prior to smoothing, the basic Naïve Bayes classifier results in performance measures of $F1_{Positive} = 0.16$ and $F1_{Negative} = 0.66$, which is far from ideal. Zero counts is a possible explanation for the lackluster performance, which can be further eliminated by applying smoothing to the weights of the terms in (2.1). The term scores incorporating the smoothing parameter α is as follows:

$$P(t \mid c) = \frac{T_{ct} + \alpha}{\sum_{t' \in V} (T_{ct'} + \alpha)} \tag{2.3}$$

Figure 1 depicts the improvement in performance of the classifier for α values ranging from $\alpha = 0$ to $\alpha = 1$ with increments of 0.05. In Figure 1, a smoothing factor of $\alpha = 0.95$ improves the performance by $\sim 23\%$ for the *negative* class and $\sim 394\%$ for the *positive* class from $\alpha = 0$ to $\alpha = 0.95$. Performance measures improve significantly, with $F1_{Positive} = 0.814$ and $F1_{Negative} = 0.81$.

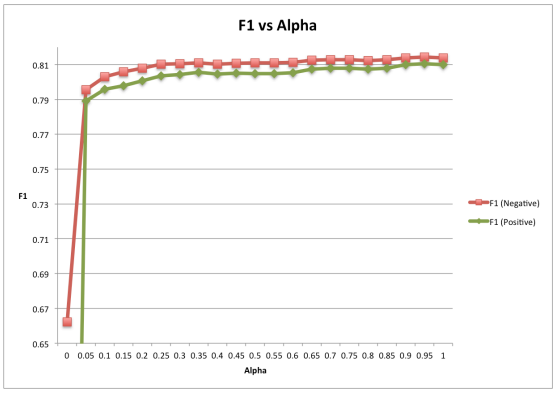

Figure 1: F1 vs Alpha for both classes (Negative and Positive)

Observations

Terms with highest weights

Table 1 shows the terms with the highest word count, or terms with the highest scores. Porter Stemmer results in the incomplete format of some words, e.g. *thi* is actually the stemmed version of *this*.

the	a	and	of	to	is	in	it
that	on	as	film	with	hi	for	he
thi	but	be	be	ar	by	i	movi
who	an	not	from	ha	her	have	at

Table 1: Highest scoring terms

The conjecture is that the highest scoring terms have equal probability of appearing in both *positive* and *negative* reviews. This hypothesis is taken into consideration in the following subsections, where two possible improvements are evaluated for their effect on performance.

Stop-words

Table 1 contains the mostly stop-words, hence stop-words are removed from the corpus of reviews in an attempt to improve performance. The results yield the F1 vs α curve in Figure 2. Contrary to hypothesis, removal of stop-words actually degrades, not improve performance. There is no need to remove stop-words.

One possible explanation for the performance degradation correlated with stop-words removal is that reviews from one class may consist of more stop-words than reviews from the other. This condition contradicts the previous conjecture that

stop-words appear equally in reviews from both classes. In other words, stop-words actually **improve** the classification process. This can be verified by comparing the lengths of reviews in both classes, which is beyond the scope of this report.

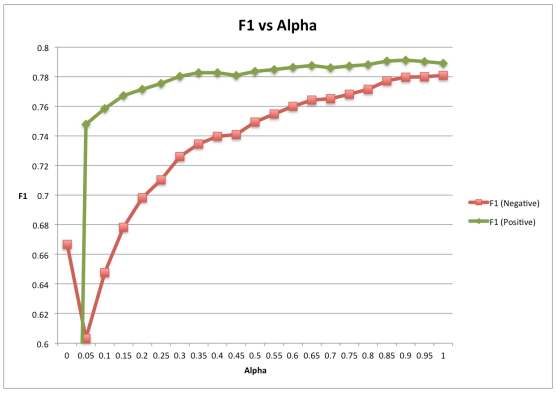

Figure 2: F1 vs Alpha for both classes after removing stop-words

Feature Selection

Feature selection is proposed as a second improvement that may refine the corpus of reviews. In so doing, terms are ranked by their *mutual information* scores. The smoothing $\alpha = 0.95$, chosen as the optimal performance α from Figure 1. Figure 3 depicts classifier performance as a function of the number of features selected. It can be inferred that the optimal performance is achieved when the entire corpus of reviews is used. This is the same as saying the number of features is greater than or equal to the corpus size. In order to achieve optimal performance, all words in the corpus of reviews are to be used in the classification process.

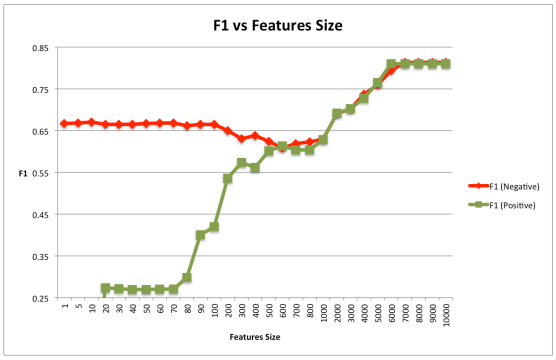

Figure 3: F1 vs Feature Size

Summary

In summary, the Naïve Bayes classifier implementation on a corpus of stemmed words resulted in a performance of $F1 = 0.814$, with a smoothing factor $\alpha = 0.95$. Smoothing improved the performance by $\sim 400\%$ as compared to the basic classifier without smoothing. Stop-words and feature selection were evaluated for possible performance improvements on the basic classifier, but are not to be adopted as they actually degrade, not improve performance.