Introduction

In this report, we discuss the implementation details and considerations of a Naïve Bayes classifier in the context of classifying movie reviews into two categories: i.e. positive and negative reviews. After stemming the corpus, the 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 our reported performance measurements are acquired by averaging on a 10-fold-cross-validation.

The first section discusses the adoption of the assumed model, the design considerations and the choice of design parameters. The second section addresses some observations and possible performance boosts, i.e. 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

In implementation, the Bernoulli model is adopted to assign weights for terms within the corpus of reviews. The choice of model results in the following Maximum Likelihood Estimation formulas to train the classifier:

\[
P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}
\]

(2.1)

\[
P(c) = \frac{N_c}{N}
\]

(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 positive or negative movie reviews.

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

Smoothing

Prior to smoothing, the basic classifier results in the performance as measured by \(F_{1_{\text{positive}}} = 0.16\) and \(F_{1_{\text{negative}}} = 0.66\), which is far from ideal. Zero counts is a possible explanation for the lackluster performance, which can be eliminated by applying smoothing to the weights of terms in (2.1). The term scores is thereby calculated as follows:

\[
P(t \mid c) = \frac{T_{ct} + \alpha}{\sum_{t' \in V} (T_{ct'} + \alpha)}
\]

(2.3)
Figure 1 depicts the improvement in performance of the classifier for different values of $\alpha$ ranging from $\alpha = 0$ to $\alpha = 1$ with increments of 0.05. By looking at Figure 1, a smoothing factor of $\alpha = 0.95$ is applied, resulting in an improvement in performance of $\sim 23\%$ for the negative class and $\sim 394\%$ for the positive class, ranging from $\alpha = 0$ to $\alpha = 0.95$. In this case, $F_{1,\text{positive}} = 0.814$ and $F_{1,\text{negative}} = 0.81$.

![Figure 1: F1 vs Alpha for both classes (Negative and Positive)](image)

Observations

Terms with highest weights

Table 1 shows the terms that scored highest (i.e. with most word counts). Words that appear incomplete are a direct result of the stemming process, e.g. *thi* is the stemmed version of *this*.

<table>
<thead>
<tr>
<th>the</th>
<th>a</th>
<th>and</th>
<th>of</th>
<th>to</th>
<th>is</th>
<th>in</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>on</td>
<td>as</td>
<td>film</td>
<td>with</td>
<td>hi</td>
<td>for</td>
<td>he</td>
</tr>
<tr>
<td>thi</td>
<td>but</td>
<td>be</td>
<td>be</td>
<td>ar</td>
<td>by</td>
<td>i</td>
<td>movi</td>
</tr>
<tr>
<td>who</td>
<td>an</td>
<td>not</td>
<td>from</td>
<td>ha</td>
<td>her</td>
<td>have</td>
<td>at</td>
</tr>
</tbody>
</table>

Table 1: Highest scoring terms

The conjecture is that the highest scoring terms have equal probability of appearing in both *positive* and *negative* reviews. Therefore, in the following subsections we describe two possible improvements and discuss their performances.

Stop-words

Since most of the words in Table 1 are stop-words, stop-words are proposed to be removed from the corpus. This yields the F1 vs $\alpha$ curve in Figure 2. Contrary to assumption, removal of stop-words actually degrades the performance, and are kept going forward.

One possible explanation for degradation in performance as a result of removing stop-words is that reviews in one class may consist of more stop-words than reviews from another. If this hypothesis is true, then the previous conjecture
that stop-words appear equally in reviews from both classes is false. In other words, stop-words indeed do improve the classification process. This can be verified by comparing the lengths of reviews in both classes, which is a possible future undertaking that is beyond the scope of this report.

![Figure 2: F1 vs Alpha for both classes after removing stop-words](image)

**Feature Selection**

The second proposal is to refine the corpus using feature selection. Terms are ranked by their mutual information scores. Smoothing is continuing to be used at $\alpha = 0.95$, the most optimal $\alpha$ given performance from Figure 1. Figure 3 depicts classifier performance as a function of the number of features selected. It can be inferred that the best performance is obtained when the entire corpus is used (i.e. number of features is greater than or equal to the corpus size). All words in the corpus are to be kept for the classification process.

![Figure 3: F1 vs Feature Size](image)

**Summary**

In summary, the implemented Naïve Bayes classifier on stemmed words produced a performance of $F1 = 0.814$ with a smoothing factor $\alpha = 0.95$. Smoothing improved the performance by ~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 were not adopted as they actually degrade performance of the classifier.