Natural Language Processing Assignment 1

Mini-report

This assignment saw us reproduce the famous paper by Pang et al (2002) that expands on simple Naïve Bayes BoW-based sentiment classification of movie reviews. The changes explored were the use of Support Vector Machines, which outperforms Naïve Bayes in many NLP tasks, as well as some slight implementation modifications. I will detail any points in my implementation where I had to deviate or extrapolate from the paper, regarding only those areas that we were tasked with reproducing.

- removed punctuation and one letter words from the dictionary

**Naïve Bayes**

* (think about whether you should use a frequency threshold and how you would set it
* Make sure the NB version you use uses at least Laplace smoothing
* Please describe clearly which implementation you use

**Support Vector Machines**

We recommend SVM Light as implementation for the practical.

**Stemming**

What effect could stemming potentially have?

**Cross Validation**

in their use of crossvalidation, and use Round-Robin splitting (on the first part of the file names) for comparability. It is assumed that you will use stratified crossvalidation.

**Results**

What would it mean for a difference not to be statistically significant?

Thus, it is possible that observed differences in the reported performance are really just noise

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Features | # of Features | Frequency or Presence? | Stemmed? | NB mean accuracy | SVM mean accuracy | P-Value |
| (1) | Unigrams |  | Frequency |  |  |  |  |
| (2) |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |