NLP Practical 1 - Sentiment Classification Revisited

Charles J.Y. Yoon

11th November, 2018

Introduction

For the first practical, document sentiment classification was revisited. Pang et al. describes multiple methods of classifying movie reviews; this domain is experimentally convenient, since there is a large dataset with classification taken by the review ratings.[1] We have been given a similar dataset that has been tokenized for classification, in the context of the NLP course.

The original experiment used various supervised learning methods with multiple models of feature extraction, but without any validation of statistical significance test on different classification models. We aim to select sub-models that may yield a statistically different result, and evaluate their effectiveness. We, however, exclude Maximum Entropy (ME) model from the original experiment.

Methods

Classifiers

We will be using Naive Bayes and Select Vector Machine("SVM") classifiers on binary classes. Laplace smoothing was used for Naive Bayes, and SVM classification was handled by `scikit-learn.svm.SVC`.

Bag of Words

Each document is represented as a vector by a bag-of-words model, an accumulation of either frequency or occurrences of words seen in the training set. Any n-gram model can be used to create the feature vectors– here we have used unigram and bigrams as per the original.

Frequency vs Presence

By testing on both types of bag-of-words models, the difference in accuracy between frequency and presence was investigated.

n-gram feature extraction

The unigram and bigram models used in sentiment classification is different to that of language generation. While such models may calculate the bigram probability as $P(x_i|x_{i-1})$, here we *take each unigram and bigram as a feature* although the features may be correlated.

Stemming

Unlike the original experiment, we will use the Porter Stemmer to preprocess the document. Since the meanings of words do not change with morphological change, stemming the words may increase the accuracy of the classifier.

Feature Cutoff

The original paper removes certain number of features by removing tokens that have occurred less than a given amount. The SVM classifier implements feature

		Naive Bayes		SVM	
		Frequency	Presence	Frequency	Presence
Unigram	Stemmed	79.95	81.45	83.10	83.75
	Unstemmed	79.85	81.55	82.95	84.95
Unigram + Bigram	Stemmed	79.70	79.20	84.40	86.55
	Unstemmed	79.20	76.95	83.65	86.55
Bigram	Stemmed	77.15	75.85	80.60	82.30
	Unstemmed	74.10	72.00	79.90	81.10

Figure 1: Accuracies in percentage given configuration with no cutoff

selection by default, which results in no difference between feature selected models and their originals. For the Naive Bayes classifiers, however, we have set the boundary to 3 occurrences, and the following result is below.

Evaluaton and Results

We have performed accuracy comparison and sign test in order to determine the efficacy of each classification model. A stratified 3-fold cross-validation set by round-robin, and the predictions were concatenated to be evaluated at once. Figure 1 shows accuracies in percentages given different configurations of each model, albeit without any feature selection. In order to verify that the different components make a significant effect, sign tests have been performed.

Change in model	p-value
Stemming	0.982
Unigram+Bigram	0.789
Bigram	0.011
SVM	0.461
Occurances	0.173
Feature Cutoff	0.893

Figure 2: Sign test results between baseline and individual change

We chose a baseline classifier (Naive Bayes, no stemming, unigrams, and frequency based bag-of-words without feature cutoff) to be compared pairwise to a set of models that has only one element changed. Figure 2 shows the `p-values` of each comparison, where higher `p-value` notes less significance in change (in favor of the null hypothesis).

Note that the sign test results are not complete as we have not done for all pair. The results above assume that each treatment's efficacy is independent of each other; such can be investigated further with more detailed results.

Conclusion

From the results of figures 1 and 2, we cannot confidently assert that the null hypothesis was defeated for all models except for bigrams, since they are all below the confidence interval.

The significant difference in the various n-gram model was anticipated as the feature vector yielded differs; bigram models have performed worse than unigram models, however, which can be the lack of each token in data affecting the results.

Reference

[1] Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79–86.