ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 6, June 2013

# Effect of Feature Reduction in Sentiment analysis of online reviews

G.Vinodhini
Department of Computer Science and Engineering, Annamalai University,
Annamalai Nagar-608002, India.

RM.Chandrasekaran
Department of Computer Science and Engineering, Annamalai University,
Annamalai Nagar-608002, India

## ABSTRACT

Sentiment analysis is the task of identifying whether the opinion expressed in a document is positive or negative about a given topic. In sentiment analysis, feature reduction is a strategy that aims at making classifiers more efficient and accurate. Unfortunately, the huge number of features found in online reviews makes many of the potential applications of sentiment analysis infeasible. In this paper we evaluate the effect of a feature reduction method with both Support Vector Machine and Naive Bayes classifiers. The feature reduction method used is principle component analysis. Our results show that it is possible to maintain a state-of-the art classification accuracy while using less number of the features through Receiver operating characteristic curves and accuracy measures.

***Keywords:*** *sentiment, opinion, feature, learning, support vector.*

## 1. INTRODUCTION

Sentiment analysis is to build a system that can classify documents as positive or negative, according to the overall sentiment expressed in the documents. Early approaches to sentiment analysis did not require domain-specific training data, their accuracy was quite poor. Subsequent research focused on supervised learning techniques that are common in text categorisation tasks, such as Support Vector Machine (SVM) and Naive Bayes (NB) classifiers. Though these techniques are far more accurate than the earlier text-based approaches, they are a lot more computationally expensive to run due to the large number of features. Very few of these features actually provide useful information to the classifier, so feature reduction can be used to reduce the number of features.

One major difficulty of the sentiment classification problem is the high dimensionality of the features used to describe texts, which raises problems in applying many sophisticated learning algorithms to text sentiment classification. The aim of feature reduction methods is to obtain a reduction of the original feature set by removing some features that are considered irrelevant for sentiment classification to yield improved classification accuracy and decrease the running time of learning algorithms (Ahmed, Chen, & Salem, 2008; Tan & Zhang, 2008; Wang et al., 2007). In this paper, we aim to make an intensive study of the effectiveness of reduced features using Principal Component Analysis (PCA) for sentiment classification tasks. The effectiveness of the features thus selected is evaluated using SVM and Naive bayes classifier. We design two models of feature sets that are particular to sentiment analysis. The unigram product attributes based and the other is dimension reduced unigram product attributes based. For each model, we apply Support vector machine and naive bayes approach.

This paper is outlined as follows. Section 2 narrates the related work. Section 3 discusses the methodology used. The Data source used is reported in Section 4. Dimension reduction technique used is discussed in Section 5. The various methods used to model the prediction system are introduced in Section 6. Section 7 summarizes the results and Section 8 concludes our work.

## 2. RELATED WORK

Most researchers employ basic feature reduction in their work in order to improve computational performance, with a few using more complicated approaches [1,2,4]. Pang & Lee (2004) used a SVM trained on subjective and objective text to remove objective sentences from the corpus. The other work that used sophisticated feature reduction was by Abbasi et al. (2008). They found that using either information gain or genetic algorithms (GA) resulted in an improvement in accuracy. They also combined the two in a new algorithm called the Entropy Weighted Genetic Algorithm, which achieved the highest level of accuracy in sentiment analysis. Positional information of the words in text can also be taken into account. In this perspective, use multinomial Naïve Bayes, together with the position information in the feature set. They conducted their experiments on the movie dataset achieving an 89% accuracy using unigrams and bigrams, which is a slight improvement over the performance reported by Pang & Lee (2004). Another variation of the SVM method was adopted by Mullen and Collier (2004 ), who used WordNet syntactic relations together with topic relevance to calculate the subjectivity scores for words. They reported an accuracy of 86% on the Pang &Lee's movie review dataset. Zaidan *et al.* (2007) used SVM and so-called "rationales" corresponding to words and phrases explaining a particular classification decision. Whitelaw *et al.* (2005) also employed SVM and a lexicon created with a semi-automated technique, which was then improved manually. Paltoglou & Thelwall (2010) suggested using tf- idf weighting schemes coupled with the SVM classifier. This solution achieved a significant improvement over the previous SVM-based approaches. Ahmed, Chen, & Salem (2008); Tan & Zhan ( 2008) and Wang et al. (2007) used feature reduction techniques to run through the corpus before the classifier has been trained and remove any features that seem unnecessary. This allows the classifier to fit a model to the problem set more quickly since there is less information to consider, and thus allows it to classify items faster.

So, our focus in this work is to make an intensive study of the effectiveness of reduced features using Principal Component Analysis (PCA) for sentiment classification of online product reviews. The effectiveness of the features thus selected is evaluated using SVM and Naive bayes classifier.

## 3. METHODOLOGY

The following is the summary of our methodology for developing and validating the prediction models on a sample of N reviews, using a set of M product attributes.

i. Perform pre-processing and segregate unigram features (product attributes) as bag of words.
ii. Develop word vector for Model using pre-processed reviews and unigram product features (Model I).
iii. Perform principle component analysis on the Model I to produce reduced feature set and develop word vector for Model using reduced feature set (Model II).

iv. Develop the classification models using the respective training data set (Model I & Model II).
   a. Develop the Support vector machine model.
   b. Develop the Naive bayes classifier model.
v. Predict the class (positive or negative) of each review in the test data set using k-fold cross validation for Model I & Model II.
vi. Compare the performance of the two methods described for Model I and Model II.

## 4. DATA SOURCE

The data set used contains product reviews sentences which were labelled as positive, negative or neutral. We collected the review sentences from the publicly available customer review dataset. This dataset contains annotated customer reviews of 5 different products. From those five products we have selected reviews of two different digital cameras only. Hu, and Liu, (2005) has employed this data set to analyze the performance of sentiment classification. In this binary classification problem, we have considered only 365 positive reviews and 135 negative reviews. The product attribute discussed in the review sentences are collected for each of the positive and negative review sentences .Pang et al. (2004) found that unigrams comprehensively out-performed bigrams and combinations. Unique unigram product features alone are grouped, which results in a final list of product attributes (features) of size 95. In terms of these, the descriptions of review dataset model (Model I) to be used in the experiment are given in Table 1.

**Table 1. Description of dataset (Model I)**

| Camera review | No.of reviews | Feature | No.of features | Positive Reviews | Negative reviews |
|---|---|---|---|---|---|
| Model I | 500 | Unigrams | 95 | 365 | 135 |

## 5. FEATURE REDUCTION

Feature reduction is selecting a subset of the features available for describing the data before applying a learning algorithm, is a common technique used. It has been widely observed that feature reduction can be a powerful tool for simplifying or speeding up computations,

Principal Components Analysis (PCA) is the widely used statistical method to reduce the dimension of feature set. Principal components analysis can transform the original data set of correlated variables into a smaller data set of uncorrelated variables that are linear combinations of the original ones. The new principal components variables are referred as domain metrics. Assuming (N X M) matrix as the word vector data with N reviews and M product attributes, the principal components algorithm works as follows.

i. Calculate the covariance matrix *D* of X.
ii. Calculate Eigen values $\lambda_j$ and eigenvectors $e_j$ of $\Sigma$, j = 1,. . .M.
iii. Reduce the dimensionality of the data (using a cutoff for the percentage of .95 variance for the eigen values).
iv. Calculate a standardized transformation matrix *Y* where each column is defined as

$$t_j = \frac{e_j}{\sqrt{\lambda_j}} \text{ for } j = 1,...., p$$

v. Calculate domain metrics for each review using

$$F_j = Xt_j \quad F = XT$$

The final result F is an n × p matrix of domain metrics with a mean of zero and variance of one, and each row is a set of domain metric values for a review.

A word vector representation of review sentences is created for Model I using the unigram features. To create the word vector list, the review sentences are pre-processed. Tokenize to split the texts of a review sentence. Transform the upper case letters to lower case to reduce ambiguity. Then stop words are filtered to removes common English words. Porter stemmer is the used for stemming to reduce words to their base or stem.

After pre-processing, the reviews are represented as unordered collections of words and the features (Unigram) are modeled as a bag of words. A word vector is created for Model I using the respective features based on the term occurrences. Using weka, the principal components for Model I with their unigram features (95 unigram) are identified. The principle component with variance less than 0.95 are obtained (Kennedy and Inkpen, 2006; Mullen and Collier, 2004; Pang and Lee, 2004; Whitelaw, Garg, et al., 2005). A word vector model for Model II is created using review sentences and the reduced principle components as features. The description of principle components obtained for Model II is shown in Table 2.

Table 2. Description of dataset (Model II)

| Properties | Value |
|---|---|
| | PC1 - |
| No.of Components | PC57 |
| Variance | <.95 |
| Standard Deviation | 0.67 |
| Proportion of variance | 0.003 |
| No.of.Features (original) | 95 |
| No.of.Features (Reduced) | 57 |
| No.of.Reviews | 500 |
| Positive Reviews | 365 |
| Negative Reviews | 135 |

## *6.* METHODS

This section discusses the methods used in this work to develop the prediction system. The statistical approach based SVM and proposed multi classifier approach are employed using weka tool.

### *6.1. SVM*

Support Vector Machines are powerful classifiers arising from statistical learning theory that have proven to be efficient for various classification tasks in text categorization. Support vector machine belong to a family of generalized linear classifiers. It is a supervised machine learning approach used for classification to find the hyper plane maximizing the minimum distance between the plane and the training points. An important property of SVMs is that

they simultaneously minimize the empirical classification error and maximize the geometric margin; hence known as maximum margin classifiers. Support vector machine is a new machine-learning paradigm. In the binary classification case, the basic idea behind the training procedure is to find a hyper plane, represented by vector $\vec{p}$ , that not only separates the document vectors in one class from those in the other, but for which the separation(margin), is as large as possible. This search corresponds to a constrained optimization problem; let $h_j \in \{1,-1\}$ (corresponding to positive and negative) be the correct class of review $d_j$ , the solution can be written as

$$\vec{p} := \sum_j \alpha_j h_j \vec{d}_j , \quad \alpha_j \geq 0$$

where the $\alpha_j$'s are obtained by solving a dual optimization problem. Those $\vec{d}_j$ such that $\alpha_j$ is greater than zero are called as support vectors, since they are the only document vectors contributing to $\vec{h}$. Classification of test instances consists simply of determining which side of p's hyperplane the data points fall on.

### *6.2. Naive Bayes*

One approach to text classification is to assign to a given document *d* the class *c. T*he *Naive Bayes* (NB) classifier is derived by the probability model. The probability model for a classifier is a conditional model

$$p(C|F_1, \ldots, F_n)$$

over a dependent class variable *C* with a small number of outcomes or *classes*, conditional on several feature variables $F_1$ through $F_n$. The problem is that if the number of features *n* is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. Therefore reformulate the model to make it more tractable. Using Bayes' theorem,

$$p(C|F_1, \ldots, F_n) = \frac{p(C)\ p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.$$

## 7. EVALUATION

In order to study the influence of the feature size in the prediction, two models are developed in each of the methods – Model I is represented as word vector with unigram product features, Model II is represented as word vector with reduced principle components. Most of the literatures showed that SVM and Naive Bayes are perfect methods in single domain opinion classification (Kennedy and Inkpen, 2006; Mullen and Collier, 2004; Pang and Lee, 2004; Whitelaw, Garg, et al., 2005). Accuracy is measured for the classifiers SVM and NB in conjunction with (& without )the use of PCA. Table 3. shows the results of evaluations using 10 fold cross validation.

Table 3. Results of evaluations

| Classifier | Accuracy -Model I (without PCA) (%) | Accuracy -Model II ( with PCA ) (%) |
|---|---|---|
| SVM | 75.2 | 77 |
| Naive Bayes | 38.8 | 45.2 |

Accuracy is better with reduced PCA alone as a component model (Table 3). An empirical analysis is done to find the influence of the Principle components attributes in the performance of classifiers. The analysis measures the accuracy of the SVM and NB classifiers for Model I and Model II . The classification performance is measured using ten-fold cross-validation. It can be observed from the Table 3 that the accuracy is more for both classifier with reduced feature set.

Thus the accuracy of the classifiers is influenced by the choice of number of attributes used .It suggests that this number of attributes is sufficiently optimal for the all classifiers to perform better input/output mapping.

ROC curves are very popular for performance evaluation and are used as an alternative metric for accuracy. The ROC curve plots the false positive rate (FPR) on the x-axis and true positive rate (TPR) on the y-axis. The FPR measures the fraction of negative examples that are misclassified as positive. The TPR measures the fraction of positive examples that are correctly labelled.

$$\text{True Positive Rate} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} * 100$$

$$\text{False Positive Rate} = \frac{\text{FalsePositive}}{\text{FalsePositive} + \text{TrueNegative}} * 100$$

The diagonal divides the ROC space. Points above the diagonal represent good classification result and points below the diagonal line represent poor results. The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. As a result of ROC analysis, the performance of SVM and Naive bayes classifier with and without PCA is shown in Figure 1 and Figure2.

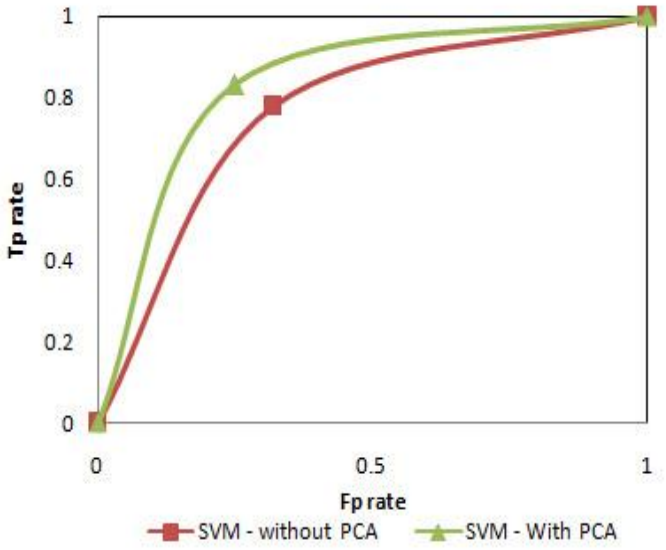


Figure 1. ROC performance of SVM classifier

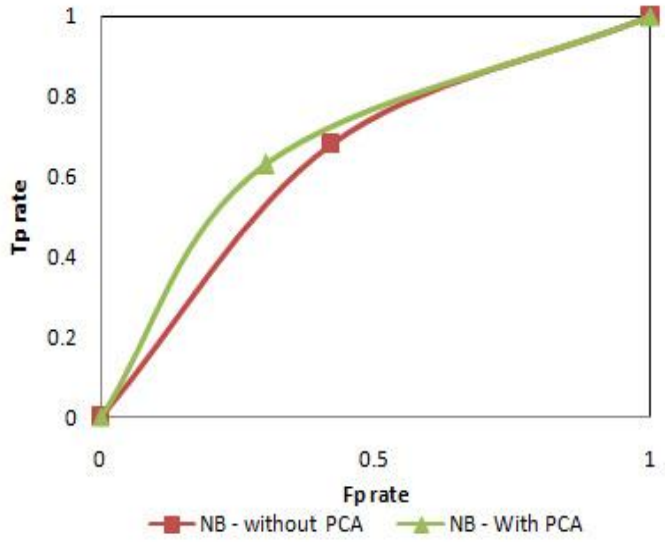


Figure 2. ROC performance of Naive bayes classifier

## 8. CONCLUSION

In this paper we empirically evaluated the performance of a feature reduction method for sentiment analysis. In particular, we introduced principle component analysis. The experiments are conducted using two classifiers, SVM and NB, on the product review data set. We achieved an increase in accuracy using PCA as a feature reduction method for naive Bayes and SVM as a classifier. This is a promising result as it is comparable with previous state-of-the-art results. Future work will include evaluating more feature reduction methods, particularly some of the common ones from text categorisation, such as information gain and chi- square test. It would also be valuable to combine some of the feature selectors to see if better feature sets can be produced. Lastly, there would be significant value in repeating these tests on another data set.

## REFERENCES

[1]. A Abbasi, HC Chen and A Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions On Information Systems, Volume 26, Number 3, 2008.

[2]. Ahmed, A., Chen, H., & Salem, A. (2008). Opinion analysis in multiple languages: feature selection for opinion classification in web forums. ACM Transactions on Information Systems, 26(3).

[3]. A.Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, Volume 22, Number 2, pages 110–125, May 2006.

[4]. B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proc. of the ACL, pages 271–278. ACL, 2004.

[5]. Mullen, T., & Collier, N. (2004). "Sentiment analysis using support vector machines with diverse information sources". In Proceedings EMNLP'04, pp. 412-418.

[6]. Paltoglou, G., & Thelwall, M. (2010). "A study of information retrieval weighting schemes for sentiment analysis", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1386-395.

[7]. Tan, S. B., & Zhang, J. (2008). An Empirical study of opinion analysis for Chinese documents. Expert Systems with Application, 34(4), 2622–2629.

[8]. Wang, S. G., Wei, Y. J., Zhang, W., Li, D. Y., & Li, W. (2007). A hybrid method of feature selection for chinese text opinion classification [C]. In Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (pp. 435–439). IEEE Computer Society.

[9]. Whitelaw C., Garg N., & Argamon, S. (2005). "Using appraisal groups for sentiment analysis". In Proceedings of the 14th ACM CIKM, pp. 625-631.

[10]. Yang, Y., & Pedersen, J.O. (1997). "A comparative study of feature selection in text categorization". In Proceedings ICML, pp. 412-420.

[11]. Zaidan. O.F., Eisner J., & Piatko, C.D. (2007). "Using annotator rationales to improve machine learning for text categorization". In Proceedings of NAACLHLT, pp. 260-267.