Measurement 55 (2014) 101–109

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Measuring the quality of hybrid opinion mining model for e-commerce application

Vinodhini G *, Chandrasekaran RM

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar 608002, India

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form 5 February 2014
Accepted 23 April 2014
Available online 6 May 2014

Keywords:
Opinion
Classification
Unigram
Bigram
Feature
Mining
Reviews

ABSTRACT

With the rapid expansion of e-commerce over the decades, the growth of the user generated content in the form of reviews is enormous on the Web. A need to organize the e-commerce reviews arises to help users and organizations in making an informed decision about the products. Opinion mining systems based on machine learning approaches are used online to categorize the customer opinion into positive or negative reviews. Different from previous approaches that employed single rule based or statistical techniques, we propose a hybrid machine learning approach built under the framework of combination (ensemble) of classifiers with principal component analysis (PCA) as a feature reduction technique. This paper introduces two hybrid models, i.e. PCA with bagging and PCA with Bayesian boosting models for feature based opinion classification of product reviews. The results are compared with two individual classifier models based on statistical learning i.e. logistic regression (LR) and support vector machine (SVM). We found that hybrid methods do better in terms of four quality measures like misclassification rate, correctness, completeness and effectiveness in classifying the opinion into positive and negative.

1. Introduction

E-commerce has attracted more and more people to buy and sell products online, customer reviews that describe experiences with product and services are becoming more important in decision making [25,26]. Potential customers are interested to know the opinions of existing online customers to gather information about the products they plan to purchase, and businesses want to analyze the opinions of the customer of their products to monitor their brand. Customer reviews generally contain the product opinions of customers expressed using features of the product [1,9,18,27,39].

Opinion mining is a branch of data mining that analyzes individual subjective opinions such the orientations of the opinions [14,20,29,33,36]. Within this broad field, much of the work has been focused on opinion polarity classification, where a text opinion is classified as positive or negative. Specifically, our focus is on feature-based opinion mining, in which the task applies to the sentence level to discover customers opinion about various aspects of a product [43]. Various machine learning classifiers have been used for opinion classification in the literature [23,28,32,42,47]. Also, many works in machine learning communities have shown that combining individual classifiers is an effective technique for improving classification accuracy [21,31,44,46]. The emerging interest and importance of text sentiment classification in the real world applications, motivates us to perform a comparative study of hybrid methods in opinion classification. This study will greatly benefit application developers as well as researchers in the areas related opinion mining.

In this work, we introduce two hybrid models for opinion classification i.e. PCA with bagging and PCA with

* Corresponding author. Tel.: +91 9626885482.
E-mail addresses: g.t.vino@gmail.com (G Vinodhini), aurmc@sify.com (RM Chandrasekaran).

http://dx.doi.org/10.1016/j.measurement.2014.04.033
0263-2241/

Bayesian boosting models, using product attributes as features for classification model. They are empirically validated using a product review data set containing reviews collected from Amazon reviews. To analyze the relationship clearly two data models are developed. Model I using only unigram product attribute as features for classification. Model II using a combination of unigram, bigram and trigram product attribute as features for classification. The results are compared with two individual statistical model i.e. logistic regression and support vector machine.

This paper is outlined as follows. Section 2 narrates the background. Section 3 discusses the problem outline used. Dimension reduction technique used is discussed in Section 4. Section 5 presents the various evaluation measures used. The various methods used to model the prediction system are introduced in Section 6. Data source used is reported in Section 7. Section 8 summarizes the results and Section 9 concludes our work.

2. Background

Much research exists on opinion mining of user opinion data, which mainly judges the polarities of user reviews [1,5,10,16,26,37,38]. Machine learning approaches applicable to opinion mining, mostly belongs to supervised classification [17]. Machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and support vector machine (SVM) have achieved great success in opinion categorization. The other most well known machine learning methods such as *K*-Nearest neighborhood, ID3, C5, Centroid and Winnow classifier are also used for opinion mining [2,19,22,26,30,35]. Among the various machine learning methods, SVM achieved great success in opinion categorization. Naive Bayes ranks next highest in performance [4,6,14,38]. Though many comparative studies exist, previous work on efficient integration of different types of classifiers to improve the opinion classification performance is rare [13,15,31,40,45].

The literature also reveals that the result of opinion mining classifier varies according to the composition method of features. Opinion mining is carried out using only unigram as features [34] and combination of unigram, bigram and trigram as features [11]. One major difficulty of the text opinion classification problem is the high dimensionality of the features used to describe texts. The aim of feature selection methods is to obtain a reduction of the original feature set by removing some features that are considered irrelevant for text opinion classification to yield an improved classification accuracy of learning algorithms [3,24]. In machine learning approaches various feature selection and reduction approaches such as information gain, mutual information, chi square test, document frequency, hybrid methods for feature selection and fisher's discriminant ratio are employed [41]. But the literature does not contribute any work using popular feature reduction method, principal component analysis in opinion classification.

The performance of opinion classification algorithms is also domain dependent. Many studies exist in mining customer opinions on product domain [5,8,12]. However, much of these studies mainly focus on identifying customer's opinion polarities towards product using single classifier.

The result of an opinion mining classifier varies according to the composition method of domain used, features selected and the type of learning algorithm. But the research that performs opinion mining by variously combining a feature reduction method and an ensemble learning algorithm is rare. Thus an attempt has been made to study the possibility to enhance the performance of the classification through the use of hybrid combination such as bagged SVM with PCA and Bayesian boosting with PCA. In this paper, we aim to make an intensive study of the effectiveness of hybrid combination of ensemble techniques with a dimension reduction method for opinion classification tasks. Another contribution of this work is to study the effect of different type of features (unigram, bigram and trigram) that can be employed to build the opinion learning models. We design two models of feature sets that are particular to opinion mining: unigram product attributes based (model I) and the other is combination of unigram, bigram and trigram based (model II). In contrast to previous work, in this paper, we describe a system that computes reduced features using PCA. To show the accuracy of reduced feature set, the features thus selected is evaluated using SVM and NB classifier. Two models are reconstructed using reduced feature sets. For each model, we apply two types of hybrid methods (bagged SVM + PCA, Bayesian boosting + PCA). A wide range of comparative experiments is conducted and the results are compared with two individual classifier models based on statistical learning i.e. logistic regression and support vector machine.

This work distinguishes itself from others in the following ways: In this work, combining a feature reduction and an ensemble learning algorithm is used for opinion classification which is not considered so far in opinion mining literature. Reduced principle component obtained is further analyzed to eliminate the least influencing attributes based on the attribute weights. In order to evaluate the prediction models rather than using usual accuracy measures like precision, recall and f-measure, four different quality parameters are used which capture the various aspects of the model quality.

3. Problem outline

We used the following methodology to develop the prediction models with two categories of word features unigram only (model I), unigram, bigram and trigram (model II). The following is the summary of our methodology for developing and validating the prediction models.

i. Perform data preprocessing and segregate unigram, bigram and trigram features (product attributes).
ii. Develop word vector for model I (unigram) and model II (unigram, bigram and trigram) using pre-processed reviews and grouped features respectively.

iii. Perform principal component analysis on the model I and model II features to produce reduced feature set for both models. The effectiveness of the features thus selected is evaluated using SVM & NB classifier.
iv. Develop the word vector models (I and II) with dimension reduced feature set to be used as training data set for the learning models.
a. Develop logistic regression model.
b. Develop the support vector machine model.
c. Develop the ensemble model using Bayesian boosting.
d. Develop the ensemble model using bagging.
v. Predict the output class (positive or negative) of each review in the test data set.
vi. Compare the predicted class with actual class.
vii. Compute the four quality parameters – misclassification rates, correctness, completeness and effectiveness and compare the predicted results of the hybrid model with the baseline methods (SVM, logistic regression).

4. Principal component analysis

Principal components analysis (PCA) is the widely used statistical method to reduce the dimension of feature set. Assuming $X(n \times m)$ matrix as the standardized word vector data with n reviews and m product attributes, the principal components algorithm works as follows [15]:

i. Calculate the covariance matrix.
ii. Calculate eigen values and eigen vectors.
iii. Reduce the dimensionality of the data.
iv. Calculate a standardized transformation matrix T.
v. Calculate domain features (p) for reviews.

The final result F is an $n \times p$ matrix of domain features.

5. Evaluating the accuracy of the model

Ten fold cross validation is used for evaluating the quality of the prediction models. In this work the results obtained for the test data set are evaluated using the following model quality parameters [7].

5.1. Misclassification rate

Misclassification rate is defined as the ratio of number of wrongly classified reviews to the total number of reviews classified by the prediction system. The wrong classifications fall into two categories. If negative reviews are classified as positive ($C1$), it is named as a type I error. If positive are classified as negative ($C2$), it is named as type II error.

$$\text{Type I error} = C1/(\text{Total no. of positive reviews})$$

$$\text{Type II error} = C2/\text{Total no. of negative reviews})$$

$$\text{overall misclassification rate} = (C1 + C2)/(\text{Total no. of reviews})$$

5.2. Correctness

Correctness is defined as the ratio of the number of reviews correctly classified as positive to the total number of reviews classified as positive. Low correctness means that a high percentage of the classes being classified as positive, which is not actually positive.

5.3. Completeness

Completeness is defined as the ratio of the number of positive reviews, classified as positive to the total number of actual positive reviews. It is a measure of the percentage of positive that would have been found if we used the prediction model in the stated manner.

5.4. Effectiveness

Effectiveness is defined as the proportion of positive reviews considered high risk out of all reviews. Let, type II misclassification is $\Pr(n\text{fp}/\text{fp})$

$$\text{Effectiveness} = \Pr(\text{fp}/\text{fp}) = 1 - \Pr(n\text{fp}/\text{fp})$$

6. Methods

This section discusses the methods used in this work to develop the prediction system. The proposed hybrid approaches (bagging and Bayesian boosting with PCA) are employed using Weka 3.7.4 and rapid miner respectively.

6.1. Baseline methods

A support vector machine (SVM) is powerful classifier arising from statistical learning theory that has proven to be efficient for various classification tasks in text categorization. SVM belongs to a family of generalized linear classifiers. It is a supervised machine learning approach used for classification and regression to find the hyper plane, maximizing the minimum distance between the plane and the training points. The SVM model is employed using Weka tool. The kernel type chosen is a polynomial kernel with default values for kernel parameters like cache size and an exponent. Other parameters like tolerance, numFolds, epsilon and filter-type in SVM classifier also uses the default values available in Weka tool [32].

Logistic regression is a standard technique based on maximum likelihood estimation. The first step in logistic methods is identifying which combination of independent variables best estimates the dependent variable. This is known as model selection. Logistic regression model is also employed using Weka tool. The model is used with default values for classification parameters [38].

6.2. Hybrid methods

6.2.1. Bagging

The main idea is to construct each member of the ensemble from a different training dataset, and to predict the combination by uniform averaging over class labels

(Whitehead and Yaeger [21]). A bootstrap sample of S items is selected uniformly at random with replacement. This means each classifier is trained on a sample of examples taken with a replacement from the training set, and each sample size is equal to the size of the original training set [13,21]. Then, they are aggregated into to make a collective decision using majority voting. Therefore, bagging produces a combined model that often performs better than the single model built from the original single training set. The bagging model proposed is employed using Weka tool. SVM is used as base classifier and number of iterations used is 5. Other parameters for base learner use the default values available in the tool. SVM has been proved to provide a good classification result in opinion mining. But the practically implemented SVM is often far from the theoretically expected level because their implementations are based on the approximated algorithms due to the high complexity of time and space. To improve the limited classification performance of the real SVM, we propose to use the hybrid model of bagged SVM and PCA.

6.2.2. Bayesian boosting

Boosting is an iterative process, which adaptively changes the distribution of training examples so that the base classifiers will focus on examples that are hard to classify [21,40]. Boosting has become one of the alternative frameworks for classifier design, together with the more established classifier like Bayesian classifier.

Algorithm:
Input: D, Training data D of labeled examples d_i.
Output: A classification model.
Procedure

i. Initialize the weight $W_i = 1/n$ of each example d_i in D, where n is the total number of training examples.
ii. Generate a new dataset D_i with equal number of examples from D using selection with replacement technique.
iii. Calculate the prior probability $P(C_i)$ for each class C_j in dataset D_i.
iv. Calculate the class conditional probabilities $P(A_{ij}|C_j)$ for each attribute values in dataset D_i.
v. Classify each training example t_i in training data D with maximum posterior probabilities.
vi. Updates the weights of each training examples di in D, according to how they were classified. If an example was misclassified then its weight is increased, or if an example was correctly classified then its weight is decreased. To updates the weights of training examples the misclassification rate is calculated, the sum of the weights of each of the training example di in D that were misclassified.

$$\text{error}(M_i) = \sum_{i}^{d} W_i * \text{err}(d_i)$$

vii. Where $\text{err}(d_i)$ is the misclassification error of example d_i. If the example d_i was misclassified, then is $\text{err}(d_i)$ 1. Otherwise, it is 0. If a training example was correctly classified, its weight is multiplied by error $(M_i)/(1 - \text{err}(M_i))$. Once the weights of all of the correctly classified examples are updated, the weights for all examples including the misclassified examples are normalized so that their sum remains the same as it was before.
viii. To normalize a weight, the algorithm multiplies the weight by the sum of the old weights, divided by the sum of the new weights. As a result, the weights of misclassified examples are increased and the weights of correctly classified examples are decreased.
ix. Repeat steps 2–6 until all the training examples d_i in D are correctly classified.
x. To classify a new/unseen example use all the probability set in each round and considers the class of new example with highest classifier's vote.

The proposed Bayesian boosting model is employed using rapid miner tool. Naive Bayes classifier is used as inner classifier and the number of iteration to combine the classifier is 5. Other parameters are used with default values.

7. Data source

The polarity data set used is a set of product review sentences which were labeled as positive, negative or neutral. We collected the review sentences from the publicly available customer review dataset (http://www.cs.uic.edu/~liub/FBS/FBS.html). This data set contains annotated customer reviews of 5 products. Of those five products we have selected reviews of 2 different digital cameras (Canon G3 & Nikon Coolpix 4300) [37,38]. There are 988 annotated reviews and the data is presented in plain text format. Out of 988 reviews of camera, 139 are negative, 371 are positive and 478 are neutral reviews. Outliers are performed as suggested in Briand et al. [7]. Ten sentences are identified as outliers and are not considered for further processing. Thus the review dataset is of size 500 (365 positive and 135 negative reviews). For our binary classification problem, we have considered only 365 positive and 135 negative (500 reviews) reviews. For each of the positive and negative review sentences the product attributes discussed in the review sentences are collected. Unique product features are grouped, which results in a final list of product attributes (features) of size 115. Among 115 product attributes 96 are unigram attributes and 19 are bigram and trigram combinations. In terms of these, the descriptions of review dataset models to be used in the experiment are given in Table 1.

7.1. Data preprocessing

A word vector representation of review sentences is created for model I and II using the respective features. The word vector set can then be reused and applied to different classifiers. To create the word vector list, the review sentences are pre-processed. The following are the steps done in data pre-processing. Tokenize to split the texts of a review sentence into a series of words or tokens to remove the 'non-letters' words which will generate single

Table 1
Properties of data source.

Camera review	No. of reviews	Feature type	No. of features	Positive reviews	Negative reviews
Model I	500	Unigram only	96	365	135
Model II	500	Unigram + bigram + trigram	96 + 19 = 115	365	135

word tokens [38]. Transform the upper case letters to lower case to reduce ambiguity. Then stop words are filtered to remove common English words. Word like these are very noisy unless removed. Porter stemmer is then used for steaming. After pre-processing, the reviews are represented as unordered collections of words and the features (unigram, ngram) are modeled as a bag of words. A word vector is created for both model I and model II using the respective features based on the term occurrences. The binary occurrences of the each feature word in the processed review sentences results in a word vector for both the models.

7.2. Independent variable

In order to study the influence of the word size in the prediction, two models are developed in each of the methods – model I is represented as word vector with only unigram attributes, model II is represented as word vector with combination of unigram, bigram and trigram attributes. Using Weka, the principal components for each of the models (I and II) with their respective features (96 unigram and 115 (unigram + bigram + trigram)) are identified. The stopping rule used is 'eigen value > 1'. Due to this stopping rule the number of principal components for the models I and II are cut down to 1 (PC1). One component with 50.7% variance is obtained for model I. One component with a cumulative variance of 52.9% is obtained for model II. Due to the stopping rule chosen, the percentage of variance is less. Most of the literature showed that SVM and NB are perfect methods in single domain opinion classification [6,16,18,23]. Also SVM and NB classifiers are used as a base classifier in our ensemble approaches. PC1 represents the reduced dimension which is obtained by stopping rule. To justify the choice of PC1 alone as reduced component, the accuracy is measured for the models I and II using the classifiers SVM and NB in conjunction with (& without) the use of PCA and PC1. Table 2 shows the results of evaluation using 10 fold cross validation.

Since the Accuracy is better with PC1 alone as a component model (Table 2), an empirical analysis is done to find the influence of the PC1 attributes in the performance of classifiers. The attributes in PC1 are sorted in decreasing order of weights ranging from 1 to 0. An empirical analysis is done to measure the accuracy of the SVM and NB classifiers by varying the number of attributes of PC1. The number of attributes chosen is based on the attribute weight as shown in Table 3 for models I and II.

The classification performance is measured using tenfold cross-validation. It can be observed from Figs. 1 and 2, that the accuracy increases with increase in the number of attributes, but when value reached some boundaries performance of classifier is the same or worse. Thus the accuracy of the classifiers is influenced by the choice of a number of attributes used based on attribute weights of principal components (PC1). When number of attributes = 25 and number of attributes = 40 for models I and II, respectively, both classifiers significantly improved the classification accuracy. After that classification accuracy is reduced with little variations between classifiers. It suggests that this number of attributes is sufficiently optimal for the all classifiers to perform better input/output mapping. Thus the irrelevant attributes of PC1 can still be reduced to improve classifier performance. As a result of analysis, the reduced feature list for models I and II are shown in Tables 4 and 5 respectively.

To perform classification, the word vector for model I and model II are reconstructed using the reduced set of features represented in Tables 4 and 5 for all review sentences. The vector models are used to compare the classification performance using two ensembles based classification i.e. bagged SVM and Bayesian boosting models and two individual classifier models based on statistical learning i.e. logistic regression and support vector machine.

8. Results and discussion

The prediction systems are developed using each of the methods discussed in Section 6 for the two models I and II. The results are compared to the actual opinion and the four quality parameters are computed. Tables 6–9 summarize the misclassification results. P1 refers to the positive group and N1 refers to the negative group. The four possible output results are presented in the inner matrix of the tables, which are P1P1 (actual positive and predicted positive), P1N1 (actual positive and predicted negative – type II error), N1P1 (actual negative and predicted positive – type

Table 2
PCA performance (accuracy) of SVM and NB.

	Model I – accuracy (%)		Model II – accuracy (%)	
	SVM	NB	SVM	NB
Without PCA	75	68.8	75	71.5
With PCA	76.8	71.2	75.4	72.8
With PC1	77	73.8	75.8	73.2

Table 3
Number of attributes for attribute weights of PC1.

Model I		Model II	
Attribute weight	No. of attributes	Attribute weight	No. of attributes
⩾0.04	6	⩾0.04	9
⩾.02	13	⩾.02	30
⩾0.01	25	⩾0.007	42
⩾0.007	37	⩾0.005	61
⩾0.005	58	⩾0.002	77
⩾.002	81	⩾.001	93
⩽1	96	⩽1	115

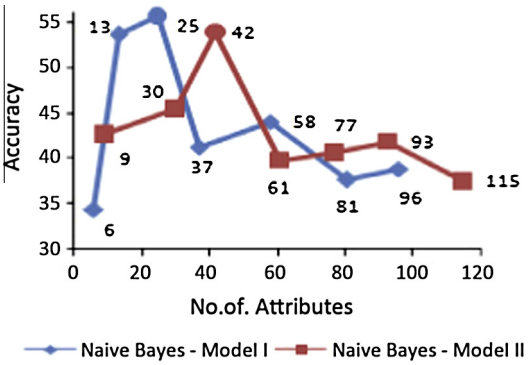

Fig. 1. Accuracy of NB with varying no. of attributes in PC1.

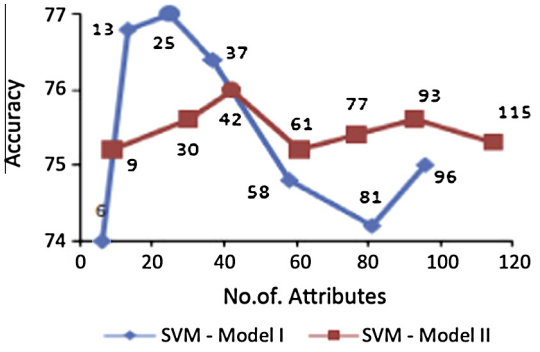

Fig. 2. Accuracy of SVM with varying no. of attributes in PC1.

Table 4
Attribute list for model I.

Component	Unigram features (25 features)	Attribute weight
PC1	Camera, digital, g, price, battery, flash, quality, setting, lens, lcd, manual, viewfinder, light, mode, zoom, use, software, optical, picture, canon, lag, mp, compact flash, download, speed	1–.01

I error) and N1N1 (actual negative and predicted negative). The overall misclassification is given at the bottom of the matrix. Table 6 presents the results obtained by logistic

Table 5
Attribute list for model II.

Component	Unigram + N-gram features (40 features)	Attribute weight
PC1	Camera, digital camera, canon g, price, quality, lens, lcd, viewfinder, light auto correction, manual, picture, strap, optical zoom, size, megapixel, lag, metering option, movie mode, battery life, mp, download, image download, compact flash, use, lag time, zoom, auto mode, software, speed, hot shoe flash, made, macro, mb card, raw format, casing, exposure control, option, indoor picture, indoor image, manual function	1–0.007

regression. Type I error is high compared to type II error. Model II has a lower misclassification rate than the model I. Table 7 shows the results of SVM. Even though the type I errors are comparatively lesser than the logistic regression method, (which is advantageous) the overall misclassification is again high due to the high values of type I error. Table 8 presents the results of hybrid bagged SVM prediction. The overall misclassification rate is reduced considerably for model I and model II compared to other three methods used. This represents the high accuracy in prediction. But, surprisingly bagged SVM with PCA reduction performs better for model I, a deviation from the previous two approaches. Table 9 gives the results of Bayesian boosting ensemble prediction. The overall misclassification rate is reduced considerably for models I and II compared to other three methods used. The high accuracy in prediction is for model II than model I.

The summary of misclassification in all prediction systems is depicted in Fig. 3. It shows the better and consistent performance of bagged ensemble method for model I and Bayesian boosting method for model II. Among the four classification methods, the ensemble methods give high prediction results for both models I and II independent of the level of granularity of the text features. Correctness is computed as the ratio of the actual positive reviews to the total positive reviews identified by the model. From the results in Table 10, it is found that the model I of bagged ensemble method and model II of Bayesian boosting model has comparatively high correctness. The completeness and effectiveness models are presented in Tables 11 and 12, respectively.

Though support vector machine dominates the logistic regression method, in general the ensemble methods have high performance in terms of completeness, correctness and effectiveness. Among the two models for four methods, model II prediction results in terms of overall misclassification rate are good in most cases. So the model II (combination of unigram, bigram and trigram) can be considered better in terms of feature set used. Among the classification methods, ensemble with PCA dominates individual classifier. Among the hybrid methods, Bayesian boosting performs better for combination of unigram, bigram and trigram. Also bagged SVM dominates for unigram model. Thus PCA is a suitable dimension reduction

Table 6
Results of SVM.

Group	Model I (predicted)			Model II (predicted)		
	Positive (P1)	Negative (N1)	Total	Positive (P1)	Negative (N1)	Total
Actual positive (P1)	350	15 4% *Type II error*	365	345	20 5% *Type II error*	365
Actual negative (N1)	100 74% *Type I error*	35	135	100 74% *Type I error*	35	135
Total	450	50	500	445	55	500
Percent	90.00%	10.00%	100%	89.00%	11.00%	100%
	Overall misclassification rate: 23%			*Overall misclassification rate: 24%*		

Table 7
Results of logistic regression.

Group	Model I (predicted)			Model II (predicted)		
	Positive (P1)	Negative (N1)	Total	Positive (P1)	Negative (N1)	Total
Actual positive (P1)	323	42 15% *Type II error*	365	315	50 14% *Type II error*	365
Actual negative (N1)	81 60.0% *Type I error*	54	135	73 54% *Type I error*	62	135
Total	404	96	500	388	112	500
Percent	80.80%	19.20%	100%	77.60%	22.40%	100%
	Overall misclassification rate: 24.6%			*Overall misclassification rate: 24.6%*		

Table 8
Results of bagged SVM.

Group	Model I (predicted)			Model II (predicted)		
	Positive (P1)	Negative (N1)	Total	Positive (P1)	Negative (N1)	Total
Actual positive (P1)	353	12 3% *Type II error*	365	347	18 5% *Type II error*	365
Actual negative (N1)	97 72% *Type I error*	38	135	101 75% *Type I error*	34	135
Total	450	50	500	448	52	500
Percent	90.00%	10.00%	100%	89.60%	10.40%	100%
	Overall misclassification rate: 21.8%			*Overall misclassification rate: 23.8%*		

Table 9
Results of Bayesian boosting.

Group	Model I (predicted)			Model II (predicted)		
	Positive (P1)	Negative (N1)	Total	Positive (P1)	Negative (N1)	Total
Actual positive (P1)	347	18 5% *Type II error*	365	343	22 6% *Type II error*	365
Actual negative (N1)	96 71.1% *Type I error*	39	135	89 66% *Type I error*	46	135
Total	443	57	500	432	68	500
Percent	88.60%	11.40%	100%	86.40%	13.60%	100%
	Overall misclassification rate: 22.6%			*Overall misclassification rate: 22.2%*		

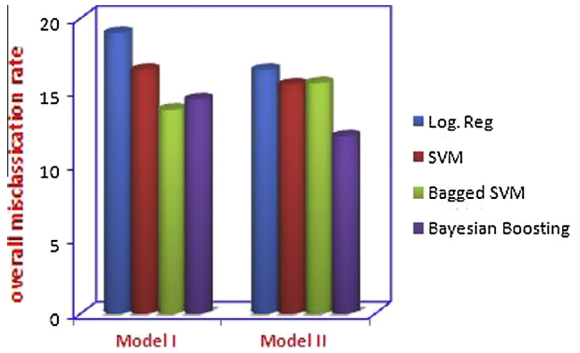

Fig. 3. Summarized overall misclassification rate.

Table 10
Results of correctness of classifiers.

	Mode I (%)	Model II (%)
SVM	90.6	90.5
LR	89.7	89.6
Bagged SVM	91.2	90.2
Bayesian boosting	90.9	92.5

Table 11
Results of completeness of classifiers.

	Model I (%)	Model II (%)
SVM	95.8	94.5
LR	95.6	94.7
Bagged SVM	97.2	96.7
Bayesian boosting	95.8	98

Table 12
Results of effectiveness of classifiers.

	Model I (%)	Model II (%)
SVM	89	89.5
LR	85	85.3
Bagged SVM	**93.8**	92.1
Bayesian boosting	91	**93.3**

method for bagged SVM and Bayesian boosting methods. The better performance of hybrid classifier is because SVM is invariant under PCA transform.

8.1. Threats to validity

When the models need to be developed in real time, the negative reviews are required to be considered equally important. The data set used is slightly positively skewed (365:135). For an imbalanced data, receiver operating characteristic (ROC) is the suitable measure to evaluate the performance of the classifiers with respect to both the classes of reviews. This work does not consider neutral reviews for classification, i.e. Multi class classification. This work is carried out with an imbalanced dataset and results of models may vary if class distribution is balanced. Further, we restricted our analysis with product features of maximum word size to 3 (trigram). Rare possibilities may exist with words describing the product of word size of higher *n*-grams. Moreover, the performance of the classifiers is evaluated for product reviews. Opinion mining is domain specific, so the hybrid methods need to be evaluated on other application domains.

9. Conclusion

In the development of prediction models to classify the reviews, more reliable approaches are expected to reduce the misclassifications. In this paper, two ensembles based hybrid approaches, which perform better than the statistical baseline approaches are introduced. Among the methods used, two hybrid methods are highly robust in nature for models I and II, which is studied through the four quality parameters. Our results also proved that the PCA is a suitable dimension reduction method for bagged SVM and Bayesian boosting methods. However, in terms of type I error measured for negative reviews, each model has a poor prediction of negative reviews compared to positive reviews. The possible explanation for the poor classification result might be nature of the product review domain used. It could happen that a large number of negative reviews were classified into the positive category because the data set was slightly positively skewed. The accuracy of hybrid methods can be increased by increasing the number of iterations (greater than 5) used for combining the classifiers. Further work needs to be done to improve the classification accuracy of negative opinion.

References

[1] A. Reyes, P. Rosso, Making objective decisions from subjective data: detecting irony in customers reviews, Decis. Support Syst. 53 (2012) 754–760.
[2] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inf. Syst. 26 (2008).
[3] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, IEEE Trans. Knowledge Data Eng. 23 (2011) 447–462.
[4] A. Ahmed, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inf. Syst. 26 (3) (2008).
[5] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, Mining Text Data (2012) 415–463.
[6] P. Beineke, T. Hastie, S. Vaithyanathan, The opinion factor: improving review classification via human-provided information, in: Proceedings of the 42nd ACL Conference, 2004.
[7] Briand, J. Wust, J.W. Daly, D. Victor Poter, Exploring the relationships between design measures and software quality in object-oriented systems, J. Syst. Software 51 (2000) 245–273.
[8] C. Bosco, V. Patti, A. Bolioli, Developing corpora for sentiment analysis and opinion mining: the case of irony and Senti-TUT, IEEE Intell. Syst. 28 (2) (2013) 55–63.
[9] C. Lin, Y. He, R. Everson, S. Ruger, Weakly supervised joint sentiment-topic detection from text, IEEE Trans. Knowl. Data Eng. 24 (6) (2012) 1134–1145.
[10] L.-S. Chen, C.-H. Liu, H. Chiu, A neural network based approach for sentiment classifica tion in the blogosphere, J. Inf. 5 (2011) 313–322.
[11] Y.H. Cho, K.J. Lee, Automatic affect recognition using natural language processing techniques and manually built affect lexicon, IEICE Trans. Inf. Syst. E89 (12) (2006) 2964–2971.
[12] K. Dave, S. Lawrence, D. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in: Proceeding of 12th Int. Conference on the WWW, 2003, pp. 519–528.
[13] G. Wang et al., Sentiment classification: the contribution of ensemble learning, Decision Support Syst. (2013).

[14] G. Vinodhini, R.M. Chandrasekaran, Sentiment analysis and opinion mining: a survey, Int. J. Adv. Res. Comput. Sci. Software Eng. 2 (6) (2012).
[15] G. Vinodhini, R.M. Chandrasekaran, Sentiment Mining Using SVM-Based Hybrid Classification Model, Computational Intelligence, Cyber Security and Computational Models, Advances in Intelligent Systems and Computing, vol. 246, Springer, 2014.
[16] M. Gamon, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in: Proceeding of the 20th Intl Conference on Computational Linguistics, 2004, p. 84.
[17] H. Chen, Intelligence and security informatics: information systems perspective, Decis. Support Syst. 41 (3) (2006).
[18] A. Kennedy, D. Inkpen, Opinion classification of movie reviews using contextual valence shifters, Comput. Intell. 22 (2) (2006) 110–125.
[19] S. Li, R. Xia, C. Zong, C.-R. Huang, A framework of feature selection methods for text categorization, in: Proceedings of the 47th Annual Meeting of the ACL, 2009, pp. 692–700.
[20] M. Chau, J. Xu, Mining communities and their relationships in blogs: a study of online hate groups, Int. J. Hum Comput Stud. 65 (1) (2007).
[21] M. Whitehead, L. Yaeger, Opinion mining using ensemble classification models, in: International Conference on Systems, Computing Sciences and Software Engineering (SCSS 08), Springer, 2008.
[22] Melville Prem, Gryc Wojciech, D. Richard, Sentiment analysis of blogs by combining lexical knowledge with text classification, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009.
[23] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources. in: Proceedings of EMNLP-2004, pp. 412–418.
[24] T. O'Keefe, I. Koprinska, Feature selection and weighting methods in sentiment analysis, in: Proceedings of the Australasian Document Computing Symposium, 2009, pp. 67–74.
[25] A. Ortigosa, R. Carro, J. Quiroga, Predicting user personality by mining social interactions in facebook, J. Comput. Syst. Sci. Special Issue Intell. Data Anal. (2013).
[26] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retrieval 2 (1–2) (2008) 1–135.
[27] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002, pp. 79–86.
[28] Bo Pang, L. Lee, A optional education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings 42nd ACL, 2004.
[29] A. Reyes, P. Rosso, D. Buscaldi, From humor recognition to irony detection: the figurative language of social media, Data Knowl. Eng. 74 (2012) 1–12.
[30] Rudy Prabowo, Mike Thelwall, Sentiment analysis: a combined approach, J. Inf. 3 (2009) 143–157 (Barcelona, Spain (pp. 271–278)).
[31] Rui Xia, Chengqing Zong, Shoushan Li, Ensemble of feature sets and classification algorithms for opinion classification, Inf. Sci. 181 (2011) 1138–1152.
[32] M. Rushdi Saleh, M.T. Martın-Valdivia, A. Montejo-Raez, L.A. Urena-Lopez, Experiments with SVM to classify opinions in different domains, Expert Syst. Appl. 38 (12) (2011) 14799–14804.
[33] F. Salvetti, S. Lewis, C. Reichenbach, Automatic opinion polarity classification of movie reviews, Colorado Res. Linguist. 17 (1) (2004).
[34] S. Sista, S. Srinivasan, Polarized lexicon for review classification, in: Proceedings of the International Conference on Machine Learning; Models, Technologies & Applications, 2004.
[35] Songho Tan, Jin Zhang, An empirical study of sentiment analysis for Chinese documents, Expert Syst. Appl. 34 (2008) 2622–2629.
[36] T.S. Raghu, H. Chen, Cyberinfrastructure for homeland security: advances ininformation sharing, data mining, and collaboration systems, Decis. Support Syst. 43 (4) (2007).
[37] H. Tang, S. Tan, X. Cheng, A survey on sentiment detection of reviews, Expert Syst. Appl. 36 (7) (2009) 10760–10773.
[38] M. Tsytsarau, T. Palpanas, Survey on mining subjective data on the web, Data Mining Knowledge Discovery (2011) 1–37.
[39] P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics, 2002.
[40] W. Li, W. Wang, Y. Chen, Heterogeneous ensemble learning for Chinese sentiment classification, J. Inf. Comput. Sci. 9 (15) (2012) 4551–4558.
[41] S.G. Wang, Y.J. Wei, W. Zhang, D.Y. Li, W. Li, A hybrid method of feature selection for Chinese text opinion classification, in: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Computer Society, 2007, pp. 435–439.
[42] C. Whitelaw, N. Garg, S. Argamon, Using appraisal groups for opinion analysis, in: Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, Bremen, DE, 2005, pp. 625–631.
[43] X. Fu, G. Liu, Y. Guo, Z. Wang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowl.-Based Syst. 37 (2013) 186–195.
[44] Y. Su, Y. Zhang, D. Ji, Y. Wang, H. Wu, Ensemble learning for sentiment classification, Chinese Lexical Semantics (2013) 84–93.
[45] Y.P. Chen, A hybrid framework using SOM and fuzzy theory for textual classification in data mining, Modeling with Words LNAI2873, 2003, pp. 153–167.
[46] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall, 2012.
[47] Ziqiong Zhang, Q. Ye, Z. Zhang, Y. Li, Sentiment classification of internet restaurant reviews written in Cantonese, Expert Syst. Appl. 38 (6) (2011) 7674–7682.