CSIT (November 2014) 2(3):169–179
DOI 10.1007/s40012-014-0055-3

ORIGINAL RESEARCH

Opinion mining using principal component analysis based ensemble model for e-commerce application

G. Vinodhini · R M Chandrasekaran

Received: 13 December 2013 / Accepted: 17 August 2014 / Published online: 9 September 2014

Abstract With the rapid expansion of e-commerce over the decades, more and more product reviews emerge on e-commerce sites. In order to effectively utilize the information available in the form of reviews, an automatic opinion mining system is needed to organize the reviews and to help the users and organizations in making an informed decision about the products. Opinion mining systems based on machine learning approaches are used to categorize the reviews containing the customer opinion into positive or negative reviews. In this paper we explore this new research area of applying a hybrid combination of machine learning approaches tied with principal component analysis as a feature reduction technique. We introduce two hybrid ensemble based models (i.e. bagging and bayesian boosting based) for opinion classification. The results are compared with two individual classifier models based on statistical learning (i.e. logistic regression and support vector machine) using a dataset of product reviews. The other objective is to compare the influence of using different n-gram schemes (unigrams, bigrams and trigrams). We found that ensemble based hybrid methods perform better in terms of various quality measures in classifying the opinion into positive and negative reviews. We also applied a pairwise statistical test to compare the significance of the classifiers.

Keywords Opinion · Classification · Unigram · N-grams · Feature · Mining · Reviews

G. Vinodhini (✉) · R M. Chandrasekaran
Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar 608002, India
e-mail: g.t.vino@gmail.com

1 Introduction

With the expanse of the e-commerce and the social networking sites, there exists vast amount of information in the social media. Product reviews expressed in social network sites play an influential role in the market business analysis. As the number of reviews has been increasing at a rapid pace, it becomes difficult for the end user to analyze the opinions expressed in social media. Thus, there is a need for an opinion mining system which enable the retrieval of opinions. Such an opinion mining system can be used by the enterprises to determine how users perceive their products and how they stand with respect to competition. It is human nature to always depend on other people's opinion and experiences while buying products. So customers can also benefit through automated opinion mining systems. The features of the product distinguish one product from other similar products and also from different brands. Such product features play a crucial role in the decision making process of the potential customer [1, 2, 3–6]. Thus, our focus in this work is on feature level opinion mining.

As the online reviews are too many for people to go through, how to automatically classify them into different opinion orientation categories (e.g. positive/negative) has become an important research problem. Various machine learning classifiers dominate the opinion classification in the literature [1, 7–10]. Many of the previous studies, however, used single classifier for the classification task. Few works in opinion classification literature have shown that combining individual classifiers is an effective technique for improving classification accuracy [11, 12, 28]. One major difficulty of the opinion classification system is the dimensionality of the features used to describe texts. The higher dimension of the features, makes it difficult in

applying machine learning algorithms to opinion classification. Thus a reduction of the feature set by removing irrelevant features is essential in opinion classification [3, 13, 14].

In this work, we have applied supervised machine learning methods in order to classify reviews. Specifically, we have used ensemble based methods on the dataset. The classification models are empirically validated on a data set obtained by crawling opinions about digital cameras from the Amazon website. We chase several goals. First, we apply principal component analysis (PCA) and perform component level analysis to obtain the reduced feature set. Secondly, we compare the effectiveness of the ensemble based methods used with two individual statistical learning based models i.e. logistic regression (LR) and Support vector machine (SVM). Finally, we check our models applied over several n-grams combinations. The experimental and statistical results indicate that the hybrid method based on ensemble is effective for review text opinion classification.

The remainder of the paper is organized as follows. Section 2 narrates the related work about opinion mining. Section 3 discusses the problem outline used in this work. Section 4 reports various steps involved in data analysis. The various methods used to model the prediction system are introduced in Sect. 5. Section 6 presents the results and Sect. 7 concludes our work.

2 Review of literature

Many interesting works exist that focus on extracting the opinions from the customer reviews [5, 15–18]. Though many researchers have investigated opinion classification from different perspectives, use of machine learning for opinion classification counts more [1, 10, 11, 19–24]. Among the Machine learning techniques, it is observed that SVM, naive bayes (NB) and decision tree approaches have achieved great success in opinion categorization [1, 10, 11, 13, 14, 21, 25]. Besides theses, other machine learning methods such as K-nearest neighborhood, ID3, C5, centroid classifier and winnow classifier are also used for opinion mining, [10, 11, 19, 21, 24]. NB also achieved great success in opinion categorization [3, 17, 26]. Although some foundational studies have investigated potential ensemble approaches in the area of opinion classification, research has been limited and more in depth empirical comparative work is needed [11, 12, 21, 27, 28].

The literature also reveals that the result of an opinion mining varies according to the composition, method of feature [3, 13, 14]. Different levels of word granularity are used as features for opinion classification. Unigrams are used as feature for opinion classification [29, 30]. The combination of unigram, bigram and trigram are used as features [31] in classifying opinion. The high dimensionality of the features obtained from the text reviews increases the complexity of the text opinion classification. Various feature selection and reduction approaches such as information gain, mutual information, Chi square test and fisher's discriminant ratio are employed in the opinion classification [13, 14, 29, 32, 33]. Except the work of [25], the opinion mining literature does not contribute any work using PCA as feature reduction technique.

2.1 Motivation and contribution

Opinion mining systems are highly domain dependant. The results can vary significantly from a domain to another which make the opinion mining a very interesting and challenging task. Prior studies have shown that many works in opinion mining exist on the product domain using single classifier [1, 32, 34, 16, 35–37]. This motivates us to conduct this analysis on product domain. PCA is a popular and effective feature reduction technique applied in various other applications [25]. Research on opinion mining by combining a feature reduction and an ensemble learning algorithm is not done so far in the literature. We therefore intuitively seek to integrate the feature reduction method and ensemble classification algorithms in an efficient way to enhance the performance of the classification. PCA is applied as a feature reduction technique to extract the reduced principal components. Reduced principle components, thus obtained from PCA is further analyzed to eliminate the least influencing attributes based on the attribute weights. In order to evaluate the prediction models different quality parameters are used to capture the various aspects of the model quality.

Another contribution of this work is to study the effect of different levels of features (unigrams, bigrams and trigrams) employed to build the opinion mining models. To analyze the relationship clearly three data models are developed. Model I using only unigram product attribute as features for classification. Model II uses a combination of unigrams and bigrams. Model III is developed using unigrams, bigrams and trigrams product attributes. For each data model (models I, II and III), wide range of comparative experiments are conducted by comparing ensemble based hybrid methods with individual classification methods. Given the importance of text sentiment classification in in the real-world applications, we believe a comparative study of ensemble based hybrid models in text sentiment classification will greatly benefit application development as well as researchers in related areas.

i. Perform data pre-processing and segreggate unigram, bigrams and trigrams product attributes (features) from reviews.
ii. Develop word vector for three models using pre-processed reviews and grouped features
    a. Model I using unigram features
    b. Model II using unigram and bigram features
    c. Model III using unigram, bigram and trigram features
iii. For each model perform principle component analysis to produce reduced feature set.
iv. Perform component level analysis to extract the dominating attributes from reduced feature set to decrease the complexity further .
v. Evaluate the effectiveness of the features thus selected using SVM and NB classifier.
vi. Develop the word vector models (Models I ,II and III) with the dominating attributes 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.
vii. Predict the output class (positive or negative) of each review in the test data set.
viii. Compute the four quality parameters – misclassification rates, correctness, completeness and effectiveness and compare the prediction results of the hybrid model with the baseline methods.

Fig. 1 Problem outline

3 Problem outline

This section describes the problem outline used to develop the prediction models. The following Fig. 1 shows the outline of our work.

4 Data source

We collected the review sentences from the publicly available customer review site (http://www.amazonreviews.com) using web crawler. We totally collected 937 customer reviews of digital camera. Out of these, 272 are negative, 355 are positive and 310 are neutral reviews. Outliers are performed as suggested in Briand et al. [38] and are not considered for further processing. In order to obtain a balanced data distribution for our binary classification problem, we have considered only 250 positive and 250 negative (500 reviews) reviews. For each of the positive and negative review sentences, the product attributes discussed in the review sentences are collected manually (bag of words). From the bag of words, 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, 12 are bigrams and 7 are trigram attributes. In terms of these, the descriptions of review dataset models to be used in the experiment are given in Table 1.

In order to study the influence of the word size in the classification, three word vector models (models I, II and III) are developed using the respective features mentioned in Table 1. To create the word vector models, the review sentences are preprocessed by tokenization, stop words removal and stemming. After pre-processing, the reviews are represented as bag of words. Model I is represented as word vector with only unigram attributes. Model II is represented as a word vector with a combination of unigram and bigram. Model III is represented as a word vector with a combination of unigram, bigram and trigram attributes. The word vector models are created based on the term occurrences. Each preprocessed product review sentences available in the polarity data set, thus obtained were labelled as positive or negative.

4.1 Feature reduction (Independent variable)

Principal component analysis is the widely used statistical method to reduce the dimension of feature set. Assuming X (n · m) matrix as the standardized word vector data with n reviews and m product attributes, the principal components algorithm works as follows.

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

The final result is a n × p matrix of domain features. Using rapidminer tool, the principal components for each of the models I, II and III are identified. The stopping rule used is 'eigen value >1′. Due to this stopping rule the number of principal components for the models (I, II and III) are cut down to 1 (PC1). PC1 represents the reduced dimension which is obtained by stopping rule. One component (PC1) with 50.7 % variance is obtained from model I. One component with a cumulative variance of 52.9 % are obtained for model II and 53.7 % for model III. Due to the stopping rule chosen, the percentage of variance is less. In order to justify the choice of PC1 alone as reduced

Table 1 Properties of data source

Camera review	No. of reviews	Feature type	No. of features	Positive reviews	Negative reviews
Model I	500	Unigrams only	96	250	250
Model II	500	Unigrams + bigrams	96 + 12 = 108	250	250
Model III	500	Unigrams + bigrams + trigrams	96 + 12 + 7 = 115	250	250

Table 2 PCA performance (accuracy) of SVM and NB

	Model I: accuracy (%)		Model II: accuracy (%)		Model III: accuracy (%)	
	SVM	NB	SVM	NB	SVM	NB
Without PCA	75	71.5	75	68.8	75.4	71.1
With PCA	75.4	72.8	76.8	71.2	77.1	73.4
With PC1	75.8	73.2	77	73.8	77.6	74.2

component, a component level analysis is done. Most of the literatures showed that SVM and NB are perfect methods in opinion classification [1, 6, 7, 39–43]. Also SVM and NB classifiers are used as base classifiers in our ensemble based approaches. So, in this component level empirical analysis, we will use SVM and NB classifiers. The accuracy is measured using the classifiers SVM and NB in conjunction with (and without) the use of PCA and PC1. Table 2 shows the results of evaluations using ten fold cross validation.

The accuracy is better with PC1 alone as a component model (Table 2). To reduce the attributes of PC1, an empirical analysis is done to find the influence of attributes in PC1. Figure 2 shows the rapidminer work flow representation of PCA component level analysis.

4.2 Component level analysis

In order to find the dominating attributes of the reduced principle component PC1, a component level analysis is done. In this analysis the accuracy of the SVM and NB classifiers in conjunction with the use of different attribute weights of PC1 is measured. The attributes in PC1 are sorted in decreasing order of weights ranging from 1 to 0. The number of attributes chosen for models I, II and III are based on the attribute weights as shown in Table 3.

The classification performance is measured using ten-fold cross validation. It can be observed from the Figs. 3 and 4. That the accuracy increases with increase in the number of attributes, but when the accuracy value reaches some boundaries, the performance of classifier are the same or worse. Thus, it is evident that the accuracy of the classifiers is influenced by the choice of a number of attributes. The choice of the number of attributes is based on attribute weights of principal components (PC1). When number of attributes of PC1 are 25, 36 and 42 for models I, II and III respectively, both classifiers significantly improved the classification accuracy. After which the classification accuracy is reduced with little variations between classifiers for all models. This suggests that model I with 25 number of attributes, model II with 36 number of attributes

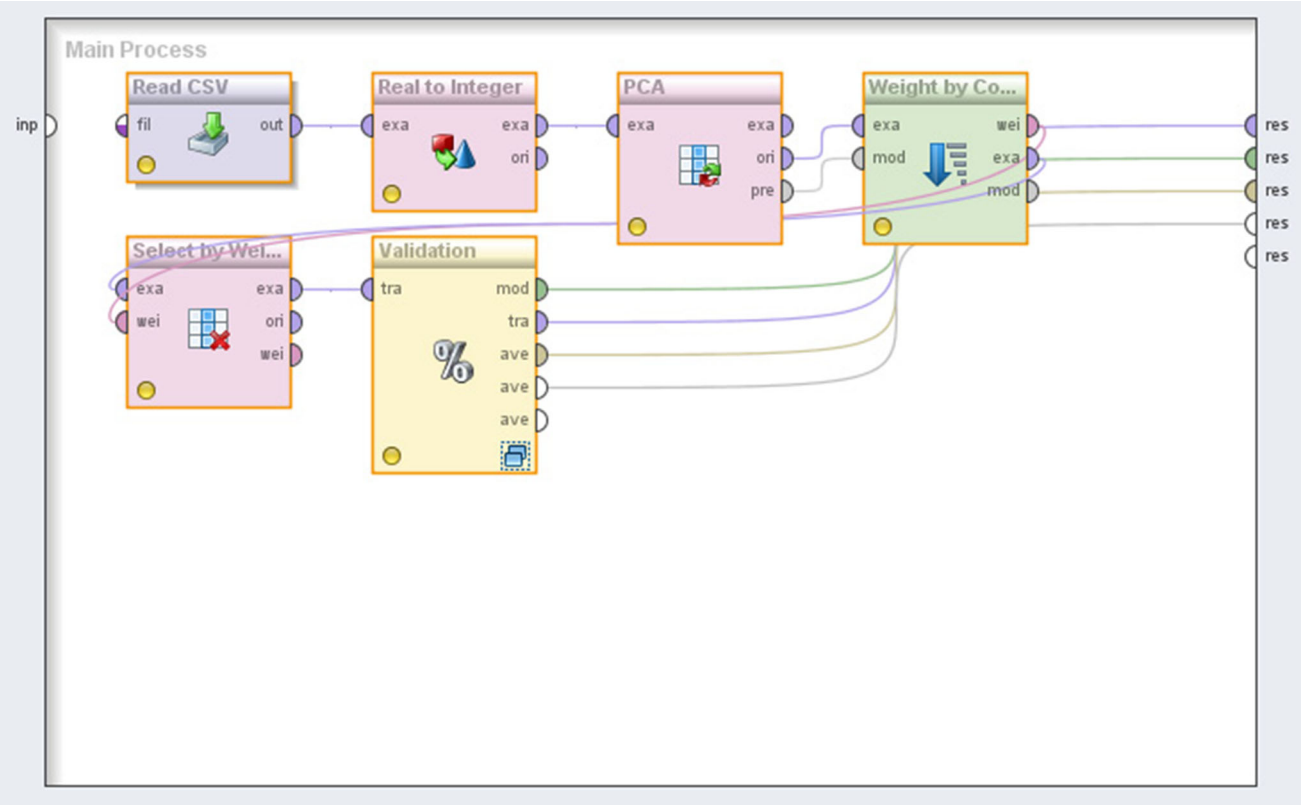

Fig. 2 Rapidminer work flow for PCA component level analysis

Table 3 Number of attributes for attribute Weights of PC1

Model I		Model II		Model III	
Attribute weight	No. of attributes	Attribute weight	No. of attributes	Attribute weight	No. of attributes
≥0.04	6	≥0.04	7	≥0.04	9
≥0.02	13	≥0.02	26	≥0.02	30
≥0.01	25	≥0.008	36	≥0.007	40
≥0.007	37	≥0.006	58	≥0.005	61
≥0.005	58	≥0.003	76	≥0.002	77
≥0.002	81	≥0.002	89	≥0.001	93
≤1	96	≤1	108	≤1	115

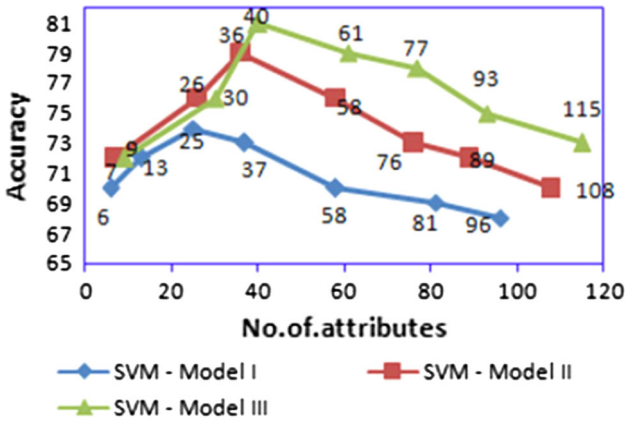

Fig. 3 Accuracy of SVM with varying no. of attributes in PC1

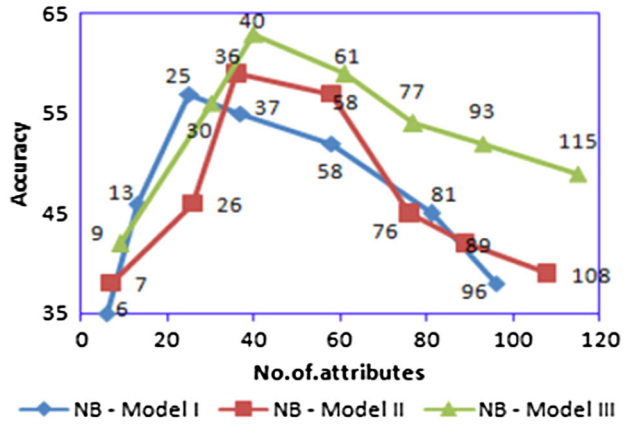

Fig. 4 Accuracy of NB with varying no. of attributes in PC1

and model III with 42 number of attributes are sufficiently optimal for the classifiers to perform better input/output mapping. Thus the irrelevant attributes of PC1 can be reduced to improve classifier performance. As a result of analysis, the reduced feature list for models I, II and III are shown in Tables 4, 5 and 6 respectively.

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

Table 5 Attribute list for Model II

Component	Unigram + bigram features (36 features)	Attribute weight
PC1	Camera, digital camera, canon g, price, quality, lens, lcd, viewfinder, manual, picture, strap, optical zoom, size, megapixel, lag, metering option, movie mode, battery life, mp, download, image download, compactflash, use, lag time, zoom, auto mode, software, speed, made, macro, raw format, casing, exposure control, option, indoor picture, indoor image, manual function	1–0.008

Table 6 Attribute list for Model III

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, compactflash, use, lag time, zoom, auto mode, software, speed, hot shoe flash, made, macro, mb memory card, raw format, casing, exposure control, option, indoor image quality, manual function	1–0.007

To perform classification, the word vector for models I, II and III are reconstructed using the reduced set of features represented in Tables 4, 5 and 6 for all review sentences. The vector models are used to compare the classification performance using two ensembles based classification i.e. bagging and bayesian boosting models and two individual classifier models.

5 Classification methods

This section discusses the classification methods used in this work to develop the prediction system. The classification methods are employed using weka tool with default values for all parameters.

5.1 Baseline methods

Support vector machines are powerful classifiers arising from statistical learning theory that have 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 n to find the hyper plane maximizing the minimum distance between the plane and the training points. LR 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. The model is used with default values for classification parameters [10].

5.2 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 [44]. The bagging algorithm creates an ensemble of models for a learning scheme where each model gives an equally weighted prediction [11, 21, 28]. 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. 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.

5.3 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. Boosting have become one of the alternative framework for classifier design, together with the more established classifier like bayesian classifier. NB classifier is used as inner classifier and the number of iterations to combine the classifier is 10. Other parameters are used with default values [11, 12, 21, 27, 28].

6 Results and discussion

The prediction systems are developed using each of the methods discussed in Sect. 5 for the models I, II and III. The results are shown in Tables 7, 8, 9, 10, 11, 12 and 13. For each 10-fold cross validation, the data set was first partitioned into ten equal sized sets and each set was in turn used as the test set while the classifier trains on the other nine sets. In this work the results obtained for the test data set are evaluated first using 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.

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

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

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

The obtained results are compared to the actual opinion and the four quality parameters are computed. Tables 7, 8, 9 and 10 summarize the misclassification results. G1 refers to the positive group and G2 refers to the negative group. The possible output results are presented in the inner matrix of the tables, which are G1G2 (actual positive and predicted negative—type II error) and G2G1 (actual negative and predicted positive—type I error). The overall misclassification is given at the bottom of the matrix.

6.1 Performance of individual classifiers

The classification results obtained for LR and SVM methods are given in Tables 7 and 8 respectively. In Table 7, the classification results of LR show that type II error is comparatively lesser than type I error for all three models (models I, II and III). This indicates that the LR method predicts positive reviews more accurately than negative reviews for models I, II and III. Among the models used, model I has better performance in terms of type I error and type II error compared to other two data models (models II and III). Due to less type I and II error, the overall misclassification of LR is also less for model I

Table 7 Results of LR

	Model I (Predicted)			Model II (Predicted)			Model III (Predicted)		
	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total
Actual positive (G1)	157	93 37.3 % Type II error	250	158	92 36.7 % Type II error	250	142	108 43 % Type II error	250
Actual negative (G2)	88 35 % Type I error	162	250	97 38.8 % Type I error	153	250	118 47 % Type I error	132	250
Total (%)	245 (49)	255 (51)	500 (100)	255 (51)	245 (49)	500 (100)	260 (52)	240 (48)	500 (100)
	Overall misclassification: 36.1 %			Overall misclassifiction: 37.8 %			Overall misclassification: 45 %		

Table 8 Results of SVM

	Model I (Predicted)			Model II (Predicted)			Model III (Predicted)		
	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total
Actual positive (G1)	183	67 26.8 % Type II error	250	182	68 27.2 % Type II error	250	176	74 29.6 % Type II error	250
Actual negative (G2)	59 23.6 % Type I error	191	250	70 28 % Type I error	180	250	85 34 % Type I error	175	250
Total (%)	242 (48)	258 (51.6)	500 (100)	252 (50.4)	248 (49.6)	500 (100)	261 (52.2)	249 (49.8)	500 (100)
	Overall misclassification: 25.2 %			Overall misclassifiction: 27.6 %			Overall misclassification: 31.3 %		

Table 9 Results of bagged SVM

	Model I (Predicted)			Model II (Predicted)			Model III (Predicted)		
	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total
Actual positive (G1)	204 Type II error	46 18.6 %	250	202 TypeII error	48 19.1 %	250	201 Type II error	49 19.6 %	250
Actual negative (G2)	48 19.2 % TypeI error	202	250	50 19.8 % TypeI error	200	250	54 21.7 % Type I error	196	250
Total (%)	252 (50.4)	248 (49.6)	500 (100)	252 (50.4)	248 (49.6)	500 (100)	255 (51)	245 (49)	500 (100)
	Overall misclassification: 18.97 %			Overall misclassifiction: 19.5 %			Overall misclassification: 20.7 %		

compared to models II and III. Table 8 gives the classification results in terms of error measures for SVM method. The type I and type II errors are considerably lesser when compared to LR method, which shows the superiority of SVM. Due to less type I and II error, the overall misclassification is also less compared to LR for all three models (models I, II and III). Though the overall misclassification is less compared to LR, SVM method also predicts positive reviews more accurately (type II error is lesser than type I error) than negative reviews for models II and III. Among the models used, SVM again performs better for model I than models II and III.

6.2 Performance of ensemble based hybrid classifiers

Table 9 and 10 presents the results of hybrid bagged SVM and hybrid boosting prediction respectively. Table 9 shows that the overall misclassification rate of bagged SVM is

Table 10 Results of bayesian boosting

	Model I (Predicted)			Model II (Predicted)			Model III (Predicted)		
	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total	Positive (G1)	Negative (G2)	Total
Actual Positive (G1)	197	53 21.4 % Type II error	250	206	44 17.7 % Type II error	250	208	42 16.7 % Type II error	250
Actual negative (G2)	58 22.9 % Type I error	192	250	46 18.5 % Type I error	204	250	47 19 % Type I error	203	250
Total (%)	255 (51)	245 (49)	500 (100)	252 (50.4)	248 (49.6)	500 (100)	255 (51)	245 (49)	500 (100)
	Overall misclassifiction: 22.28 %			Overallmisclassifiction: 18.1 %			Overallmisclassification: 17.8 %		

Table 11 Results of correctness of classifiers

	Model I (%)	Model II (%)	Model III (%)
SVM	73.4	72.8	70.5
LR	64.7	63.3	56.9
Bayesian boosting	78.6	82.3	83.3
Bagged SVM	81.4	80.9	80.4

Table 12 Results of completeness of classifiers

	Mode I (%)	Model II (%)	Model III (%)
SVM	75.6	72.2	67.4
LR	64.1	61.9	54.6
Bayesian boosting	78.8	80.2	80.9
Bagged SVM	81.6	81.7	77.3

Table 13 Results of effectiveness of classifiers

	Model I (%)	Model II (%)	Model III (%)
SVM	79	79.5	78.5
LR	65.1	65.3	69.7
Bayesian boosting	80.8	82.1	83.3
Bagged SVM	81	80.3	80.5

reduced considerably for models I, II and III compared to SVM (the best individual classification method found in Sect. 6.2). This represents the high accuracy in prediction of the bagged ensemble based method compared to the individual classifiers. The classification results also show that type II error is comparatively lesser than type I error for all three models (models I, II and III). This indicates that the bagged SVM based hybrid method also predicts positive reviews more accurately than negative reviews for models I, II and III. In general, bagged SVM with PCA reduction performs better for model I compared to models II and III.

Table 10 gives the results of bayesian boosting based hybrid prediction. The overall misclassification rate is reduced considerably for models I, II and III compared to the best individual classification method identified in Sect. 6.2 (SVM). Bayesian boosting ensemble based hybrid method performs better than SVM. But, bagged SVM based hybrid method dominates bayesian boosting ensemble based hybrid method in terms of type I error and type II error for model I. Thus, in terms of overall misclassification rate for model I. Comparing with bagged SVM based model bayesian boosting method dominates for models II and III with lesser type I and type II error.

In general, the results in Tables 7, 8, 9 and 10 shows that the ensemble based hybrid approach performs better than other individual classification methods. The performance of bagged SVM ensemble model is appreciable for model I and bayesian boosting is better for all other models (II & III). Among the models, model I performs with high accuracy for all classification methods used except bayesian boosting (Figs. 5, 6).

6.3 Quality metrics

In addition to the misclassification rate, the following quality metrics are evaluated.

6.3.1 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.

6.3.2 Completeness

Completeness is defined as the ratio of number of positive reviews classified as positive to the total number of actual positive reviews.

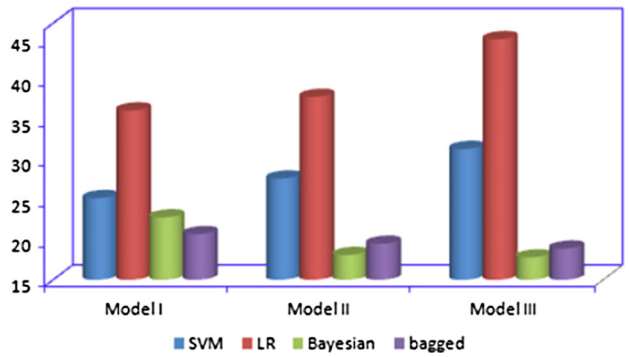

Fig. 5 overall misclassification rate of data models

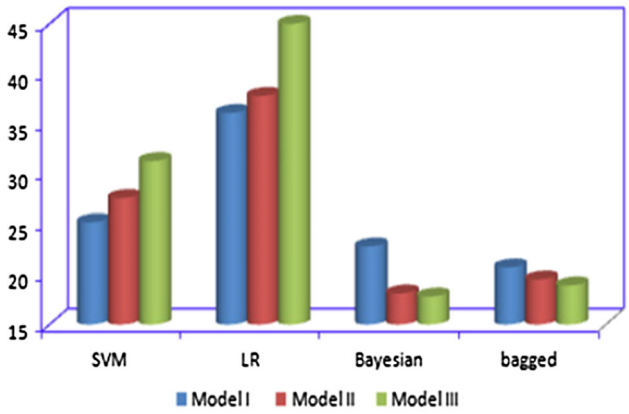

Fig. 6 overall misclassification rate of classifiers

6.3.3 Effectiveness

Effectiveness is defined as the proportion of positive reviews considered high risk out of all reviews. Let, Type II misclassification is Pr(nfp/fp)

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

Tables 11, 12 and 13 summarize the various quality measures of all the classification methods used in the analysis. These measures used for evaluation are discussed in Sect. 6.1. From the results in Table 11, it is found that the SVM and LR based models (I, II and III) lead to low correctness, which implies that a large number of sentences that are not positive/negative would have been inspected. The correctness value is much higher for hybrid ensemble methods compared to other methods used for models I, II and III. Among the hybrid classifiers, highest correctness of 83.3 % is achieved by model III of bayesian boosting in classifying review sentences. Among the models I, II and III in each classification method, classification results are good for model I of SVM, LR and bagged SVM. This proves that rather than the combination of the unigram, bigram and trigram, unigrams alone have a strong relationship to review classification. In general hybrid ensemble based models classifies the reviews very accurately with high correctness.

Completeness of the classification models is shown in Table 12. Table 12 shows that hybrid ensemble based methods predicts the maximum positive and negative reviews present compared to other methods used for all models. Among three models, model I of bagged SVM predicts the maximum positive and negative reviews with high completeness of 81.6 %. The effectiveness of the models are represented in the Table 13. Effectiveness captures the productive effort to be spent in inspecting the real positive and negative review sentences. Bagged SVM proves to be more effective and for models I and bayesian boosting is effective for models II and III. Among the classification model used, hybrid methods classify the review with better effectiveness. The higher effectiveness of model III for bayesian boosting (81.3 %) indicates that the waste of effort during analysis is very minimum.

In general, our experimental results show that among the classification methods used, hybrid ensemble methods perform better on all quality measures. Among the hybrid classification methods bagged SVM achieves better performance in all quality measures for model I. Model II and III suites better for bayesian boosting classification method. Thus, for bayesian boosting, the inclusion of bigrams and trigrams provides better performance compared to the performance of the classifier using unigrams alone. Moreover the results also shows that PCA is a suitable dimension reduction method for ensemble based methods.

6.4 Statistical significance test

We applied the nonparametric McNemar's statistical test to compare the performance of the best trained classifier. Comparison of the classifiers based on McNemar's nonparametric statistical test showed that the ensemble based hybrid method performs better. The null hypothesis (H0) for this experimental design suggests that different classifiers perform similarly whereas the alternative hypothesis (H1) claims otherwise suggesting that at least one of the classifiers performs differently. The z scores will indicate whether we should accept H0 and reject H1 or vice versa. In order to calculate the z scores, the classification results of the three classifiers must be identified for each individual instance.

In Tables 14, 15 and 16, the arrowheads ↑ denote the classifier mentioned in the table row header performed better in the given dataset and ← denote the classifier mentioned in the table column performed better in the given dataset. Z scores are given next to the arrowheads as a measure of how statistically significant the results are. By looking at the Mc nemar's test results for the model I

Table 14 McNemar's test results: model I

	Model I			
	SVM	LR	Bagged SVM	Bayesian boosting
SVM	0	← 3.26	↑5.2061	↑2.43
LR		0	↑1.9431	↑0.82
Bagged SVM			0	←2.76
Bayesian boosting				0

Table 15 McNemar's test results: model II

	Model II			
	SVM	LR	Bagged SVM	Bayesian boosting
SVM	0	← 2.6	↑1.2061	↑2.0
LR		0	↑1.96	↑1.86
Bagged SVM			0	↑2.42
Bayesian boosting				0

Table 16 McNemar's test results: model III

	Model III			
	SVM	LR	Bagged SVM	Bayesian boosting
SVM	0	← 1.94	↑3.99	↑3.66
LR		0	↑2.05	↑1.75
Bagged SVM			0	↑3.0
Bayesian boosting				0

(Table 14), it is deduced that bagged SVM has produced significantly better results than SVM, LR and bayesian boosting classifiers (H1 is accepted with a confidence level of more than 99.5 %). SVM classifier performed better than LR for model I. In Table 15, the Mc nemar's test results for the model II data set shows that bagged SVM performs better than SVM and LR. It is also observed that bayesian boosting performs better than bagged SVM for model II. The performance differences between ensemble classifiers and other individual classifiers were found to be statistically significant for the model I (H1 is accepted with a confidence level of more than 99 %). For all three models SVM performs statistically better than LR. For model III (Table 16), among the ensemble methods used bayesian boosting performs better than bagged SVM. So, the hypothesis H1 is accepted with a confidence level of more than 99.5 %. We calculated the p values for the one-tailed Mc nemar's tests comparing our ensemble based approach with the baselines. The resulting p values shows that bagged SVM based hybrid approach is significantly better than the other approaches for model I. This improvement is statistically significant at $p < 0.005$. For model II, bayesian boosting is statistically significantly better than the other approaches ($p < 0.001$). For model III, bayesian boosting is statistically significantly better than the other approaches ($p < 0.005$).

6.5 Threats to validity

This work does not consider neutral reviews for classification i.e. multi class classification. Moreover, the performance of the classifiers is evaluated for product reviews, but opinion analysis is domain specific. So the hybrid methods need to be evaluated on other application domains. Product attributes are selected from review sentences manually, which cannot be assured as 100 % accurate. So a suitable part of speech tagging approach may be employed.

7 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, the combination of ensembles and PCA methods were highly robust in nature for models I, II and III which was studied through the various quality parameters. Bagged SVM dominates for unigrams model with a reduction in overall misclassification rate of 3.3 % compared to bayesian boosting based hybrid model. Bayesian boosting performs better for combination of of unigrams, bigrams and trigrams with a reduction in overall misclassification rate of 2.9 % compared to bagged SVM based method. In future, the performance of hybrid classifier is to be evaluated on various other domains. Different hybrid combinations of soft computing techniques can also investigated. The use of PCA as feature reduction technique must be analyzed on various other feature selection methods like information gain & mutual information in the future. The effect of feature reduction methods (PCA, fisher's linear discrimination ratio, latent semantic indexing) combined with other ensemble methods such as stacking and voting can be done as an extension of this work.

References

1. Pang B, Lee L, Vaithyanathan S (2002), Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in natural language processing (pp 79–86)

2. Turney PD (2002) 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
3. Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst 26, 12:1–12:34
4. Boiy E, Moens M-F (2009) A machine learning approach to sentiment analysis in multilingual web texts. Inf Retr 12(5): 526–558
5. Vinodhini G, Chandrasekaran RM (2012) Sentiment analysis and opinion mining: a survey. Int J Adv Res Comput Sci Softw Eng 2(6)
6. Pang Bo, Lee L (2004). A opinional education: Opinion analysis using subjectivity summarization based on minimum cuts. In: Proceedings 42nd ACL
7. Mullen T, Collier N (2004) Opinion analysis using support vector machines with diverse information sources. In: Proceedings of EMNLP-2004, Barcelona, Spain (pp 412–418)
8. Rushdi Saleh M, Martın-Valdivia MT, Montejo-Raez A, Urena-Lopez LA (2011) Experiments with SVM to classify opinions in different domains. Exper Syst Appl 38(12):14799–14804
9. Kim S-M, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of the 20th international conference on computational Linguistics, (Association for Computational Linguistics), p 1367
10. Ziqiong Zhang Z, Ye Q, Zhang Z, Li Y (2011) Sentiment classification of Internet restaurant reviews written in Cantonese. Expert Syst Appl 38(6):7674–7682
11. Xia Rui, Zong Chengqing, Li Shoushan (2011) Ensemble of feature sets and classification algorithms for opinion classification. Inf Sci 181:1138–1152
12. Li W, Wang W, Chen Y (2012) Heterogeneous ensemble learning for Chinese sentiment classification. J Inf Comput Sci 9(15):4551–4558
13. Tan Songho, Zhang Jin (2008) An empirical study of opinion analysis for chinese documents. Expert Syst Appl 34:2622–2629
14. Wang SG, Wei YJ, Zhang W, Li DY, 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
15. Liu B (2010) Sentiment analysis and subjectivity. Handb Nat Lang Process, pp 627–666
16. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135
17. Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. Expert Syst Appl 36(7):10760–10773
18. Tsytsarau M, Palpanas T (2011) Survey on mining subjective data on the web. Data Min Knowl Discov 24:1–37
19. Li S, Xia R, Zong C, Huang C-R (2009) A framework of feature selection methods for text categorization. In: Proceedings of the 47th annual meeting of the ACL (pp 692–700)
20. Melville, Wojciech Gryc, "Opinion Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD'09, June 28–July 1, 2009, Paris, France. Copyright 2009 ACM 978-1-60558-495-9/09/06
21. Prabowo R, Mike T Barcelona, Spain (2009) Opinion analysis: a combined approach. J Inf 3:143–157
22. Wang S, Deyu L, Yingjie W, Hongxia L (2009) "A feature selection method based on fisher's discriminant ratio for text sentiment classification." In: Web information systems and mining. Springer, Berlin, Heidelberg, pp 88–97
23. Abbasi A, France S, Zhang Z, Chen H (2011) Selecting attributes for sentiment classification using feature relation networks. IEEE Trans Knowl Data Eng 23:447–462
24. Chen L-S, Liu C-H, Chiu H (2011) A neural network based approach for sentiment classification in the blogosphere. J Informetr 5:313–322
25. Cambria E, Schuller B, Xia Y, Havasi C (2013) "New avenues in opinion mining and sentiment analysis." IEEE Intell Syst 28(2):15–21
26. Tsytsarau M, Palpanas, T (2012) Survey on mining subjective data on the web. Data Min Knowl Discov 24(3):478–514
27. Chen JS (2003) Market segmentation by tourists' sentiments. Ann Tour Res 30(1):178–193
28. Whitehead M, Yaeger L (2010) "Sentiment mining using ensemble classification models." In: Innovations and advances in computer sciences and engineering, Springer, Netherlands, 509–514
29. Vinodhini G, Chandrasekaran RM (2013) Effect of feature reduction in sentiment analysis of online reviews. Int J Adv Res Comput Eng Technol (IJARCET) 2(6):2165–2172
30. Sista S, Srinivasan S, (2004) Polarized lexicon for review classification. In: Proceedings of the international conference on machine learning; models, technologies and applications 2004
31. Cho YH, Lee KJ (2006) Automatic affect recognition using natural language processing techniques and manually built affect lexicon. IEICE Tran Inf Syst E89(12):2964–2971
32. Gamon M (2004) 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 (p 84)
33. O'Keefe T, Koprinska I (2009) Feature selection and weighting methods in sentiment analysis, In: Proceedings of the Australasian document computing symposium (pp 67–74)
34. Dave K, Lawrence S, Pennock D (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In Proceeding of 12th intl. conference on the WWW, (pp 519–528)
35. Chen H (2006) Intelligence and security informatics: information systems perspective. Decis Support Syst 41(3):555–559
36. Chau M, Xu J (2007) Mining communities and their relationships in blogs: a study of online hate groups. Int J Hum-Comput Stud 65(1):57–70
37. Raghu TS, Chen H (2007) Cyberinfrastructure for homeland security: advances in information sharing, data mining, and collaboration systems. Decis Support Syst 43(4):1321–1323
38. Briand L, Daly J, Wust J (2000) A unified framework for coupling measurement in object-oriented systems. IEEE Trans Softw Eng 25(1):91–121
39. Kennedy A, Inkpen D (2006) Opinion classification of movie reviews using contextual valence shifters. Comput Intell 22(2):110–125
40. Gamon M, Aue A (2005) Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In: Proceedings of the ACL workshop on feature engineering for machine learning in natural language processing. Association for Computational Linguistics, pp 57–64
41. Salvetti F, Lewis S, Reichenbach C (2004) Automatic opinion polarity classification of movie reviews. Colorado research in linguistics. University of Colorado, Boulder (vol. 17, no. 1)
42. Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for opinion analysis. In: Proceedings of CIKM-05, 14th ACM international conference on information and knowledge management, Bremen (pp 625–631)
43. Beineke P, Trevor H, Shivakumar V (2004) "The sentimental factor: improving review classification via human-provided information." In: Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics
44. Whitehead M, Yaeger L (2008) Opinion mining using ensemble classification models. In: International conference on systems, computing sciences and software engineering (SCSS 08), Springer, Berlin