King Saud University
Journal of King Saud University – Computer and Information Sciences
www.ksu.edu.sa
www.sciencedirect.com

A comparative performance evaluation of neural network based approach for sentiment classification of online reviews

G. Vinodhini *, R.M. Chandrasekaran

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

Received 27 July 2013; revised 5 January 2014; accepted 13 March 2014
Available online 31 October 2015

KEYWORDS

Artificial neural networks;
Sentiment analysis;
Text classification;
Opinion mining;
Support vector machine

Abstract The aim of sentiment classification is to efficiently identify the emotions expressed in the form of text messages. Machine learning methods for sentiment classification have been extensively studied, due to their predominant classification performance. Recent studies suggest that ensemble based machine learning methods provide better performance in classification. Artificial neural networks (ANNs) are rarely being investigated in the literature of sentiment classification. This paper compares neural network based sentiment classification methods (back propagation neural network (BPN), probabilistic neural network (PNN) & homogeneous ensemble of PNN (HEN)) using varying levels of word granularity as features for feature level sentiment classification. They are validated using a dataset of product reviews collected from the Amazon reviews website. An empirical analysis is done to compare results of ANN based methods with two statistical individual methods. The methods are evaluated using five different quality measures and results show that the homogeneous ensemble of the neural network method provides better performance. Among the two neural network approaches used, probabilistic neural networks (PNNs) outperform in classifying the sentiment of the product reviews. The integration of neural network based sentiment classification methods with principal component analysis (PCA) as a feature reduction technique provides superior performance in terms of training time also.

1. Introduction

Sentiment analysis is an interdisciplinary area which comprises of natural language processing, text analysis and computational linguistics to identify the text sentiment. Web has been a rapidly growing platform for online users to express their sentiment and emotion in the form of text messages. As the opinionated texts are often too many for people to wade through to make a decision, an automatic sentiment classification method is necessary

* Corresponding author. Tel.: +91 9626885482.
E-mail address: g.t.vino@gmail.com (G. Vinodhini).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.03.024

to classify text messages into different sentiment orientation categories (e.g. positive/negative). Researchers explored various interesting approaches on the development of sentiment classification models for the classification of review sentiments, but work is still in progress. Many researchers investigated using various machine learning methods on English text sentiment classification (Prabowo and Thelwall, 2009; Chaovalit and Zhou, 2005; Turney, 2002; Pang et al., 2002) and a few studies on Chinese text sentiment classification (Ye et al., 2005, 2009; Wang et al., 2007; Tan and Zhang, 2008).

Neural networks have seen a rapid growth over the years, and are being applied successfully in various application domains for the classification problems. But the state of the art technique for neural network based text sentiment classification are found to be rare from the literature (Zhu et al., 2010; Chen et al., 2011; Sharma and Dey, 2012; Moraes et al., 2013). In recent years, there has been a growing interest in using ensemble learning techniques, which combine the outputs of several base classification techniques to enhance classification accuracy. However, compared with other research domains, related work about neural network based ensemble methods contributing to sentiment classification are still limited and more extensive experimental work is needed in this area (Wilson et al., 2006; Tsutsumi et al., 2007; Abbasi et al., 2008; Lu and Tsou, 2010; Whitehead and Yaeger, 2010; Xia et al., 2011; Su et al., 2013; Li et al., 2012). To fill this research gap, this paper makes a comparative study of the effectiveness of neural network based ensemble learning for sentiment classification. Feature selection is a crucial problem in the text classification. The aim of feature selection methods is to reduce the original feature set by removing irrelevant features for sentiment classification in order to improve accuracy of classification models (Wang et al., 2007; Tan and Zhang, 2008; Ahmed et al., 2008). PCA is used in this study in order to extract mathematically the most common features in all the models.

In this study, three different neural network models such as PNN, BPN and a homogeneous ensemble of PNN (HEN) are compared. The investigation is carried out for predicting the sentiment of product reviews with product attributes as features. Results of the evaluation of three different ANN based models are compared with that of the two different statistical methods (support vector machine and linear discriminant analysis) by computing five different quality attributes. Further, to analyze the relationship more clearly three different feature vector models are developed in each method. Model I is created using only unigram features. Model II is created using unigram and bigram features. Model III is obtained using unigram, bigram and trigram features.

1.1. Motivation and contribution

Existing works on the effectiveness of neural network based classification methods have been mainly conducted on text based topic classification (Ghiassi et al., 2012). There is lack of a comparative study on the effectiveness of neural networks based methods in text sentiment classification. The emerging interest and importance of text sentiment classification in the real world applications, motivates us to perform a comparative study of neural network based methods in sentiment classification. This study will greatly benefit application developers as well as researchers in the areas related to sentiment analysis.

Specifically, in this paper, we study the effectiveness of the neural networks based methods in sentiment classification as the interest of this study for three reasons.

- First, neural network based models has been very successfully applied to text classification and many other supervised learning tasks (Ur-Rahman and Harding, 2012; Ghiassi et al., 2012).
- The deep architectures of neural networks with layers (hidden) represents intelligent behavior more efficiently than “shallow architectures” like support vector machines (SVMs).
- The major features of neural networks such as adaptive learning, parallelism, fault tolerance, and generalization provide superior performance.

In spite of the above mentioned features of neural network methods, a few of the present research work on sentiment classification addressed the importance of integrating classification results provided by multiple classifiers (Xia et al., 2011). In addition, not much investigation has been carried out in sentiment classification to evaluate the benefits of combining neural network algorithms in order to increase the accuracy. Moreover, most existing studies in sentiment classification used the traditional measures for performance evaluation. A recent study (Kanmani et al., 2007), however, showed that various quality measures can be proposed to evaluate the accuracy of classification models in another domain like software fault prediction. This further motivates this study to evaluate the various performance evaluation metrics.

This work distinguishes itself from existing works in the following ways: In this work, probabilistic neural network and an ensemble of PNN are used for sentiment prediction which is not considered so far in sentiment analysis literature. This paper also provides a comparative study of existing neural network methods for sentiment classification through extensive experiments on a review dataset. Though proposing new techniques for sentiment prediction is not the main focus, we developed a homogeneous ensemble of PNN for classification which is not done so far in sentiment classification literature (Xia et al., 2013). Most of the earlier studies used various feature reduction methods but we attempted to use a hybrid combination of PCA and neural network (Cambria et al., 2013). In order to evaluate the prediction models in addition to traditional measures, five different quality parameters are used to capture the various quality aspects of the classification model. Training time is measured to show the superiority of feature reduction with the neural network based approach.

This paper outline follows the paper model of Kanmani et al. (2007). Section 2 is for the methodology used to develop the models. The data source used is reported in Section 3. The methods used to model the classification are introduced in Section 4. Section 5 lists out the various evaluation measures used. The findings from the experiments are discussed in Section 6. Section 7 describes the related work and Section 8 concludes the work done and proposes future works.

2. Model methodology

The methodology of the work is summarized below for developing and validating the classification models.

i. Perform data pre-processing and segregate the features (product attributes).
ii. Develop word vector for model I using unigram features, model II using unigram and bigram features and model III using unigram, bigram and trigram features.
iii. Perform PCA on the model I, II and III to produce reduced feature set for all the models.
iv. Develop the classification methods using the respective training data set with the dimension reduced feature set.
   a. Develop support vector machine model.
   b. Develop linear discriminant analysis model.
   c. Develop the BPN based neural network model.
   d. Develop the PNN based neural network model.
   e. Develop the homogeneous ensemble model based on PNN.
v. Classify the class (positive or negative) of each review in the test data set.
vi. Compare the classification results with actual results.
vii. Compute the quality parameters such as the overall error rate (misclassification), completeness, correctness, efficiency and effectiveness and compare the classification accuracy of the methods and compute training time of learning models.

3. Data source

The polarity data set used is a set of product review sentences which were labeled as positive or negative or neutral. We collected the review sentences from the publicly available customer review website www.amazonreviews.com. The domain chosen for the study is digital camera reviews. A Java web crawler was developed to download 970 positive reviews and 710 negative reviews randomly. In the crawled reviews, it is found that, there are borderline and neutral reviews in between along with the clear positive and negative reviews. We discard a review if it is not clearly aligned toward positive or negative sentiment. Outliers analysis is performed (Briand and Wust, 2002). Twenty-five sentences are identified as outliers and are not considered for further processing. As a result, there are 950 positive and 705 negative reviews. For our binary classification problem, to avoid the imbalanced class distribution, we selected 600 positive and 600 negative reviews randomly to establish the data set.

3.1. Data pre-processing

Previous studies revealed that pre-processing of text messages can improve the performance of text classification (Salton et al., 1997). The steps involved in data pre-processing are tokenization and transformation to reduce ambiguity. Then stop words are filtered to remove common English words such as 'a' and 'the' etc. Porter stemmer is then used for stemming. After pre-processing, the reviews are represented as unordered collections of words (bag of words).

3.2. Feature identification

Each product has its own set of features. As product reviews are about product features (also defined as product attributes) the product features are good indicators in classifying the sentiment of product reviews for product review based sentiment classification. Hence, the right features can be selected based on product features. To construct a feature space for product feature based sentiment classification, product features can be included and treated as features in the feature space. For each of the positive and negative review sentences represented as bag of words, the product features in the review sentences are collected. From the machine learning perspective, it is useful for the features to include only relevant information and also to be independent of each other .The unique characteristic of a product feature is that they are mostly nouns and noun phrases by part of speech tagging. In order to identify the nouns and noun phrases, part of speech (Stanford POS) tagging is applied and then association mining is done on the review sentences of nouns and noun phrases to identify frequent features. Compactness pruning and redundancy pruning are applied on the frequent features to obtain more accurate features (Hu and Liu, 2004).

The product features extracted from review sentences are unigram, bigram and trigrams. Table 1 shows the description of the data models used.

3.3. Feature vector

Converting a piece of text into a feature vector model is an important part of machine learning methods for sentiment classification. A word vector representation of review sentences is created using the features identified. The feature vector model is constructed by using term presence method. Another focus of the work is to compare the influence of using different n-gram schemes. For this reason, the product features identified are grouped based on the word granularity as unigram, bigram and trigram (Table 2). In order to find the effect of the word size in the classification, three different models are developed with varying levels of word granularity. Model I is represented as feature vector with only unigram features, model II is represented as a feature vector with a combination of unigram and bigram features and model III is represented as feature vector with combination of unigram, bigram and

Table 1 Properties of data source.

	Model I	Model II	Model III
Product	Camera	Camera	Camera
No. of reviews	1200	1200	1200
Positive reviews	600	600	600
Negative reviews	600	600	600
Feature	Unigram	Unigram, bigram	Unigram, bigram & trigram
No. of attributes	155 attributes,1 class label (sentiment)	196 attributes, 1 class label (sentiment)	215 attributes, 1 class label (sentiment)
Attribute type	Integer	Integer	Integer
Class attribute	Binomial	Binomial	Binomial
Vector space	1200 × 156	1200 × 197	1200 × 216

Table 2 Sample unigram, bigram and trigram features identified.

Word granularity	Sample product features identified
Unigrams (155)	Camera, digital, price, battery, flash, quality, setting, lens, lcd, manual, etc
Bigrams (41)	Raw format, exposure control, zoom option, indoor picture, indoor image, manual function, etc
Trigrams (19)	Mb memory card, image raw format, indoor image quality, etc

trigram features. In order to reduce the dimensionality of the feature space, PCA is employed as a dimensionality reduction technique.

3.4. Feature reduction

Principal component analysis is a linear technique for dimensionality reduction which performs a linear mapping of the data to a lower dimensional space. The mapping is done in such a way that the variance of the data in the low dimensional representation is maximized (Cambria et al., 2013), thus resulting in new principal component variables (PC's). The steps involved are given in Fig. 1.

Using weka, the principal components of the models I, II and III are identified. The stopping rule used is 'eigenvalue > 1'. Due to this stopping rule used the number of principal components are reduced to 4, 6 and 6 for the models I, II and III respectively. Four components with 50.7% cumulative variance are obtained for model I. Six components are obtained for model II with a cumulative variance of 52.9%. Six components with cumulative variance of 53.4% are obtained for model III. The percentage of variance is less due to the stopping rule selected. The feature vector models for models I, II and III are reconstructed using the reduced principal components as features. The description of principal components obtained for models is shown in Table 3.

4. Methods

The methods used to develop the classification system in this work are three neural networks based and two statistical based. Support vector machine and linear discriminant analysis are statistical methods used which are employed using rapid miner tool. The neural network approaches used are BPN and PNN which are implemented using Matlab.

4.1. Linear discriminant analysis

Linear discriminant analysis (LDA) is a popular data classification method used in various application domains. LDA calculates a rule to classify reviews as positive or negative by reducing misclassification probability. LDA is suitable for cases where the within class frequencies are not equal. LDA aims to maximize the ratio of between-class variance to the within class variance in a data set thus assuring maximal separability (Li et al., 2008).

i. Calculate the covariance matrix.
ii. Obtain the eigen values and eigenvectors.
iii. Reduce the dimensionality of the data
iv. Calculate PC's for each review .

Figure 1 Steps of PCA.

Table 3 Description of models (feature reduced).

Properties	Model I	Model II	Model III
No. of components	PC1–PC4	PC1–PC6	PC1–PC6
Variance (%)	<50.7	<52.9	<53.4
Standard deviation	0.67	0.67	0.67
No. of features (original)	155	196	215
No. of principal components (reduced)	4	6	6
No. of reviews	1200	1200	1200
Positive reviews	600	600	600
Negative reviews	600	600	600

4.2. SVM

Support vector machine is a popular classifier arising from statistical learning theory that has proven to be efficient for various classification tasks in text categorization. SVMs are a supervised machine learning classification technique which uses a kernel function to map an input feature space into a new space where the classes are linearly separable (Joachims, 1998). The SVM model is employed using the rapid miner 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 use the default values available.

4.3. Back propagation neural network (BPN)

Neural networks have many influencing properties like adaptive learning, fault tolerance, and generalization. In this work, BPN is employed because of its superior classification ability. The BPN training pseudo code is summarized as follows (Sharma and Dey, 2012).

```
While the error is too large.

Step 1. For each training pattern presented in random order
        do the following.
  a. The inputs are applied to the network.
  b. Calculate the output for every neuron from the input
     layer, through the hidden layer(s), to the output layer.
  c. Calculate the error at the outputs.
  d. Use the output error to compute error signals for pre-
     output layers.
  e. Use the error signals to compute weight adjustments.
  f. Apply the weight adjustments.

Step 2. Periodically evaluate the network performance.
```

In order to obtain an optimal neural network architecture, different architectures are tested. The architectures are varied

Table 4 BPN parameters.

Model	Neurons in three layers	Learning rate	Momentum	Gain	Epochs
I	4, 7, 2	0.1	0.4	1	358
II	6, 19, 2	0.1	0.4	1	379
III	6, 19, 2	0.1	0.4	1	388

by changing the number of hidden layer neurons, learning rate, momentum rate and epochs. Table 4 summarizes the details of suitable architecture for the models I, II and III. The neural network architecture is designed using Matlab neural network tool box.

The logistic function is used as activation function. After training, the network is simulated for the validation data set and the classification outputs are obtained.

4.4. Probabilistic neural network (PNN)

A PNN is based on statistical Bayesian classification algorithm. The functions are organized into multilayered feed forward network with four layers such as input layer, pattern layer, summation layer and output layer. The input layer consists of input nodes which are the set of measurements. The pattern layer is fully connected to the input layer, with one neuron for each pattern in the training set. The pattern layer outputs are selectively connected to the summation units depending on the class of patterns. The following are the steps involved in the PNN model (Savchenko, 2013).

Step 1. The input layer neurons distribute input measurements to all the neurons in the pattern layer.

Step 2. The second layer has the Gaussian kernel function formed using the given set of data points.

Step 3. The third layer performs an average operation of the outputs for each review class.

Step 4. The fourth layer performs a vote, selecting the largest value and class label is then determined.

PNN was implemented using the Matlab tool box. The size of the training data set is the number of (1200) neurons in the hidden layer. The smoothing factor value is 1 for the models I, II and III.

4.5. Homogeneous ensemble (HEN)

Ensemble methods combine the predictions of multiple base models. Base models are created by resampling, of the training data. A homogeneous ensemble method integrates similar types of base classifiers. Homogeneous ensembles train base learners (PNN) each from a different bootstrap sample by calling a base learning algorithm. A bootstrap sample is obtained by sub sampling the training data set with replacement, where the size of a sample is the same as that of the training data set. After obtaining the base learners, ensemble model combines them by majority voting and the majority voted class is predicted (Su et al., 2013). The pseudo-code and design of HEN approach are shown in Figs. 2 and 3 respectively.

5. Evaluation measures

Tenfold cross validation is used for validating the classification methods. The classification methods are evaluated using the five different quality measures used by Kanmani et al. (2007) in their research work on software fault prediction. As these measures are not investigated so far in the sentiment analysis domain, we have used these measures.

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 model. The wrong classifications fall into two categories. If negative reviews are classified as positive (C1), it is named as type I error. If positive reviews are classified as negative (C2), it is named as Type II error (Kanmani et al., 2007) (Eqs. (1)–(3)).

$$\text{Type I error} = \text{C1}/(\text{Total no. of positive reviews}) \tag{1}$$

$$\text{Type II error} = \text{C2}/(\text{Total no. of negative reviews}) \tag{2}$$

$$\text{Overall misclassification rate} = (\text{C1} + \text{C2})/(\text{Total no. of reviews}) \tag{3}$$

5.2. Correctness

Correctness is defined as 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 are being classified as positive, which is not actually positive (Kanmani et al., 2007).

5.3. Completeness

Completeness is defined by Briand and Wust (2002) as the ratio of number of positive reviews classified as positive to the total number of 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 (Kanmani et al., 2007).

```
Input:
D, a set of d training tuples;
K, the number of models in the ensemble; (K=5)
A, learning scheme ( PNN)
Output: A composite model, M*.
Process:
for i = 1 to k do // create k models:
    create bootstrap sample, Di, by sampling D with replacement.
    use Di to derive a model, Mi;
end for
//To use the composite model on a tuple, X:
if classification then
    let each of the k models classify X and return the majority vote
```

Figure 2 Pseudo code of ensemble approach.

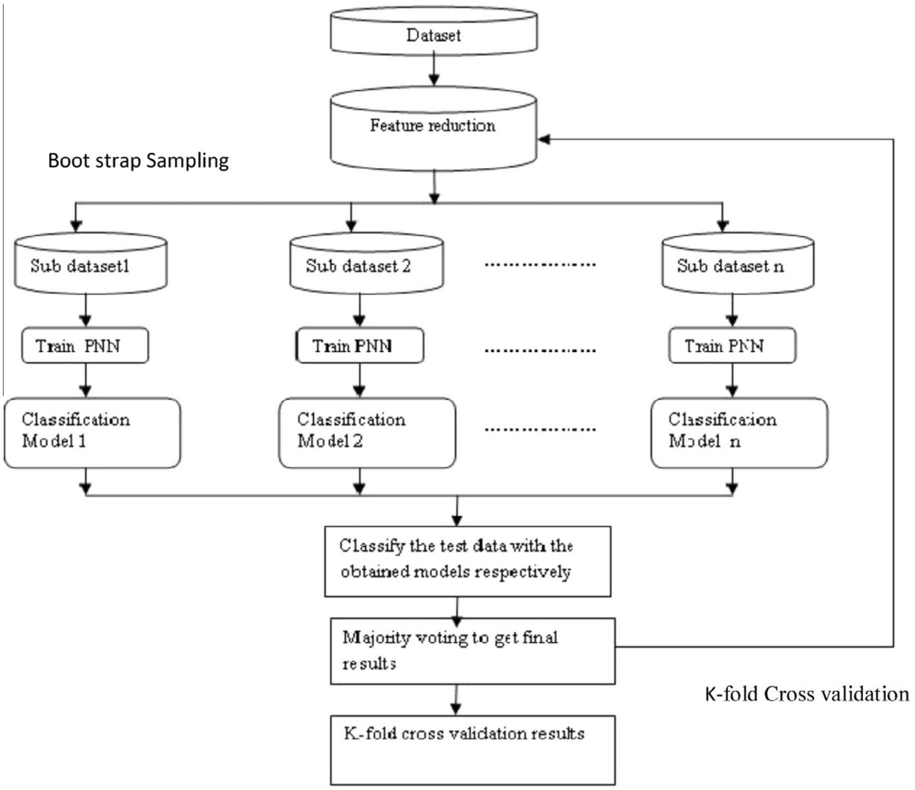

Figure 3 Design of HEN approach.

5.4. *Effectiveness*

Effectiveness is defined as the proportion of positive reviews considered high risk out of all reviews (Eq. (4)). Let, Type II misclassification is $\Pr(nfp/fp)$ and $\Pr(fp/fp)$ is number of reviews predicted and in actual are positive(Kanmani et al., 2007).

$$\text{Effectiveness} = \Pr(fp/fp) = 1 - \Pr(nfp/fp) \tag{4}$$

5.5. *Efficiency*

Efficiency is defined as the proportion of predicted positive reviews that are inspected out of all reviews (Eq. (5)). Type I misclassification is $\Pr(fp/nfp)$. $\prod_{nfp}$ be the expected proportion of negative review sentences. $\prod_{fp}$ be expected proportion of positive review sentences (Kanmani et al., 2007).

$$\text{Efficiency} = \frac{\Pr(fp/fp)\prod_{fp}}{\Pr(fp/nfp)\prod_{nfp} + \Pr(fp/fp)\prod[fp]} \tag{5}$$

6. Results and discussion

The classification systems are developed using the methods described in Section 4. Models I, II and III are used as feature vector models for classification. The obtained sentiment results are compared to the actual sentiment. Results of precision and recall measured for all the classification methods used in this study are shown in Figs. 4 and 5. It is observed from the Figs. 4 and 5. That the precision and recall values are higher for the

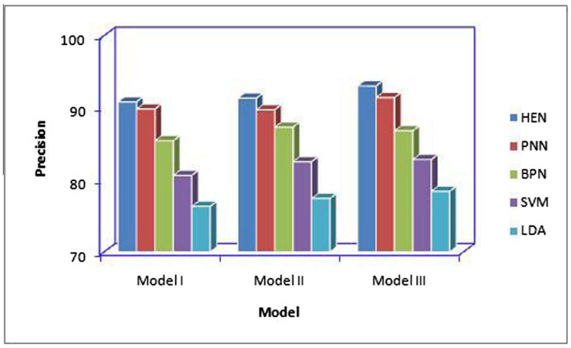

Figure 4 Precision of classifiers.

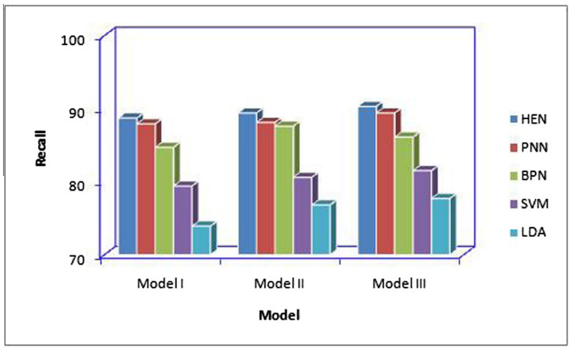

Figure 5 Recall of classifiers.

HEN approach. Moreover, precision and recall values are higher for PNN compared to the individual classifiers used. For all classification methods used, model III performs better than other models.

In addition to precision and recall, all the five quality parameters mentioned above are calculated. Tables 5–9 summarize the classification results of all five classification methods. In the table representation of the classification results (Tables 5–9), type II error is represented as G1G2 (actual positive and predicted negative) and type I error is represented as G2G1 (actual negative and predicted positive – type I error). The overall misclassification is also computed.

Tables 5–9 summarize the classification results. The classification results obtained for the LDA model are tabulated in Table 5. Results in Table 5 depict that the model III has better performance in terms of type I error and type II error compared to models I and II. Table 6 gives the classification results for the SVM method. The classification results show that the type I and type II errors are considerably lesser when compared to LDA for models I, II and III. This shows the superiority of SVM compared to LDA in sentiment classification. As type I and type II errors are less, the overall misclassification is also less for SVM compared to LDA. Back propagation based neural network prediction results are presented in Table 7. The type I and II error rates of the BPN method is much lower than the two individual classification models (SVM, LDA). But, a variation in performance of the feature vector model is noted. The overall misclassification rate is less for model II

Table 5 Results of LDA method.

	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)	458	142	600	464	136	600	471	129	600
	76.3%	23.7%	100%	77.4%	22.6%	100%	78.4%	21.6%	100%
	Type II error			Type II error			Type II error		
Actual negative (G2)	161	439	600	140	460	600	135	465	600
	26.8%	73.2%	100%	23.3%	76.7%	100%	22.5%	77.5%	100%
	Type I error			Type I error			Type I error		
Total %	6191	581	1200	604	596	1200	606	594	1200
	50.6%	49.4%	100%	50.3%	49.7%	100%	50.5%	49.5%	100%
	Overall misclassification: 25.3%			Overall misclassification: 23%			Overall misclassification: 22.1%		

Table 6 Results of SVM method.

	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)	483	117	600	495	105	600	497	103	600
	80.6%	19.4%	100%	82.5%	17.5%	100%	82.8%	17.2%	100%
	Type II error			Type II error			Type II error		
Actual negative (G2)	125	475	600	119	481	600	113	487	600
	20.8%	79.2%	100%	19.9%	80.1%	100%	18.8%	81.2%	100%
	Type I error			Type I error			Type I error		
Total %	608	592	1200	614	586	1200	610	590	1200
	50.6%	49.4%	100%	51.2%	48.8%	100%	50.8%	49.2%	100%
	Overall misclassification:20.2%			Overall misclassification: 18.7%			Overall misclassification: 18%		

Table 7 Results of BPN method.

	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)	513	87	600	530	70	600	521	79	600
	85.4%	14.6%	100%	87.3%	11.7%	100%	86.8%	13.2%	100%
	Type II error			Type II error			Type II error		
Actual negative (G2)	93	507	600	73	527	600	84	516	600
	15.5%	84.5%	100%	12.9%	87.%	100%	14%	86%	100%
	Type I error			Type I error			Type I error		
Total %	606	594	1200	603	597	1200	605	595	1200
	50.5%	49.5%	100%	50.3%	49.7%	100%	50.4%	49.6%	100%
	Overall misclassification: 15%			Overall misclassification: 12.3%			Overall misclassification: 13.6%		

Table 8 Results of PNN method.

	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)	539	61	600	541	59	600	549	52	600
	89.8%	10.17%	100%	90.2%	9.8%	100%	91.4%	8.6%	100%
	Type II error			Type II error			Type II error		
Actual negative (G2)	74	526	600	73	527	600	65	534	600
	12.3%	87.6%	100%	12.1%	87.9%	100%	10.8%	89.2%	100%
	Type I error			Type I error			Type I error		
Total %	613	587	1200	614	586	1200	614	586	1200
	51.1%	48.9%	100%	51.2%	48.8%	100%	51.2%	48.8%	100%
	Overall misclassification:11.2%			Overall misclassification:11%			Overall misclassification:9.7%		

Table 9 Results of HEN method.

	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)	545	55	600	548	52	600	558	42	600
	90.8%	9.17%	100%	91.3%	8.6%	100%	93%	7%	100%
	Type II error			Type II error			Type II error		
Actual negative (G2)	69	531	600	65	535	600	60	540	600
	11.3%	88.6%	100%	10.8%	89.2%	100%	10%	90%	100%
	Type I error			Type I error			Type I error		
Total %	614	586	1200	613	587	1200	618	582	1200
	51.2%	48.8%	100%	51.1%	48.9%	100%	51.5%	48.5%	100%
	Overall misclassification: 10.2%			Overall misclassification: 9.7%			Overall misclassification: 8.5%		

than the models I & III. Similarly a lesser type I and type II error are observed for of model II than the models I & III. BPN results show that the BPN method classifies better with unigram and bigrams combination rather than with unigram alone or with a combination of unigram, bigram and trigram features. It is also observed that BPN performs better compared to SVM and LDA in terms of error rate. This shows the predominant nature of neural networks in sentiment classification. Table 8 shows results of the PNN classification. Results in Table 8 depict that the model III has better performance in terms of type I error and type II error compared to models I and II of PNN. The classification results also show that the type I and type II errors are considerably lesser when compared to BPN, SVM and LDA for models I, II and III. This shows the superiority of PNN compared to other individual classification methods used. As type I and type II errors are less, the overall misclassification is also less for PNN.

Among the individual models the overall misclassification rate of PNN shows better accuracy in the performance of the classification methods for models I, II &III. This is due to the nature of PNN model that as soon as one pattern representing each category has been observed, the network can begin to generalize to new patterns. As additional patterns are observed and stored into the network, the generalization will improve and the decision boundary can become more complex. Among the different models used, model III performs with low error rate for all classification methods (with BPN as an exception). As PNN is found to perform better from the tabulated results (Tables 5–8), a homogeneous ensemble of PNN is developed for classification in order to increase the accuracy further. Table 9 shows results of a homogeneous ensemble of PNN methods. Type I and type II errors are observed to reduce considerably for all the models (I, II and III). This result in a minimum overall misclassification rate compared to PNN. The lesser the overall misclassification rate the greater the accuracy of the classifier. The result shows that the ensemble approach performs better than individual PNN neural network model. Among the models, the performance of homogeneous ensemble model is appreciable for model III. The homogeneous ensemble method gave a relatively more reliable prediction than those using a single classifier (PNN) for models I, II and III.

From results in Table 10, it is found that the support vector machine and linear discriminant analysis methods have less correctness for models I, II and III. This represents that a larger number of non positive review values would have been examined. The correctness value is much higher for HEN method compared to other methods used for models I, II and III (88.7%, 89.4% and 90.3%). Among the individual classifiers, highest correctness of 89% is achieved by model

Table 10 Results of correctness (in%).

Classifier	Model-I	Model-II	Model-III
HEN	88.7	89.4	90.3
PNN	87.9	88.1	89.4
BPN	84.7	87.6	86.1
SVM	79.4	80.6	81.5
LDA	73.9	76.8	77.7

III of PNN in classifying positive review sentences. Among the models I, II and III, classification results are good for model III for most of the classifiers in terms of correctness. This proves that the combination of the unigram, bigram and trigram has a strong relationship to review sentiment classification. Thus the PNN based model classifies the reviews very accurately with high correctness.

Completeness of the classification models is shown in Table 11. Table 11 shows that a homogeneous ensemble of neural network predicts the maximum positive and negative reviews present compared to other methods used for all models I, II and III. Among three models, model III of HEN predicts the maximum positive and negative reviews. Among the individual classification methods used, probabilistic neural network predicts the maximum positive and negative reviews present compared to BPN, SVM and LDA for the models I, II and III. All individual classifiers perform better for model III, but with an exception that the BPN model performs better with unigram and bigram features rather than including trigram features. The effectiveness of the classification methods is represented in Table 12. The efficiency of the model is represented in the Table 13. Both effectiveness and efficiency capture the productive effort to be spent in inspecting the real positive and negative review sentences. Homogeneous ensemble of PNN proves to be more effective and efficient for all models I, II and III. The efficiency of model III of HEN is 90%, this is because of the 558 actual positive review sentences being classified as the positive review. Among the individual classification methods used, BPN and PNN models classify the positive review with better efficiency and effectiveness. The higher effectiveness indicates that the waste of effort during analysis is very minimum. The efficiency of model III of PNN is 87%, this is because of 539 actual positive review sentences being classified as the positive review. This is due to the fact that PNN has the single pass training and ability to approximate any PDF by the sum of multivariate Gaussian functions.

Table 11 Results of completeness (in%).

Classifier	Model-I	Model-II	Model-III
HEN	90.8	91.3	93
PNN	89.8	90.1	91.5
BPN	85.5	88.3	86.8
SVM	80.5	82.5	82.8
LDA	76.3	77.3	78.5

Table 12 Results of effectiveness (in%).

Method	Model-I	Model-II	Model-III
HEN	89.7	90.25	91.5
PNN	88.7	89.05	90.3
BPN	84.95	87.15	86.4
SVM	79.9	81.3	82
LDA	74.75	77.05	77.95

Table 13 Results of efficiency (in%).

Method	Model-I	Model-II	Model-III
HEN	85.7	87.9	90.1
PNN	84.3	85.9	87.1
BPN	80.2	81.8	81.6
SVM	74.4	75.3	76.6
LDA	70.4	70.9	72.3

In general, our experimental results show that among the classification methods used, a homogeneous ensemble of PNN performs better on all quality measures. Among the individual classification methods PNN achieves better performance in all quality measures. Model III suites better for almost all classification methods except for BPN with a minimum variation. Thus the inclusion of bigrams and trigrams provides better performance compared to the performance of the classifiers using unigrams alone.

Next, we wanted to know the effect of feature reduction (PCA) by measuring the training time. The average training time of the classification methods used is summarized in Table 14. There is a drastic reduction in training time with PCA being used as a feature reduction method. This shows that high volume of data dimension of textual data will degrade the performance of classifiers and lead to a long training time. Table 14 shows that the training time is reduced considerably for individual neural network method. Among the neural network methods used, the PNN method has more reduction in the percentage of training time. This is because of the practical advantage of PNN, unlike BPN, PNN operates totally in parallel without a need for feedback from the individual neurons to the inputs.

The training time of the ensemble method with PCA is very minimal compared to that of without using PCA. But the training time of the ensemble method is high compared to other individual classification methods with PCA because of the multiple classifier combination. Thus the proposed ensemble approach is applicable where more reliable prediction is needed rather than considering the training time. But the use of PCA shows a drastic reduction in training time for the HEN method also.

6.1. Threats for validity

A few numbers of threats are there to the validity of this study on a reasonable number of reviews. The proposed methods need to be investigated on other domains because of the domain specific nature of sentiment analysis. The POS tagging approach involved in segregating the nouns describing the product attributes in review sentences may not be guaranteed as 100% complete, because there are rare cases where part of speech of product attributes may not be a noun. The performance of neural network based models used need

Table 14 Summary of the training time.

Method	Without PCA (in sec)			With PCA (in sec)		
	Model-I	Model-II	Model-III	Model-I	Model-II	Model-III
HEN	243.3	262.7	277.5	82.4	89.9	92.3
PNN	102.8	143.3	150.4	33.3	36.4	41.0
BPN	127.6	174.2	198.7	42.1	45.5	53.4
SVM	140.5	186.9	220.2	48.7	62.5	75.7
LDA	148.1	197.1	237.5	68.6	95.0	128.1

to be investigated for multi class classifications i.e. consider neutral reviews for classification. The dominating nature of the neural network based model is shown with a balanced dataset and results of models may vary if class distribution is unbalanced. Further, we restricted our analysis with product features of maximum word size to 3 (trigrams). Rare possibilities may exist where product attributes can be of higher n-grams. Though most of the sentiment mining work on product reviews has been carried out using reviews collected from Amazon reviews, very few benchmark dataset are also available for product reviews. So the investigation is to be carried out using benchmark datasets available. Though a reasonable number of reviews (1200) are used in this analysis, the performance needs to be proved by increasing the number of reviews.

7. Related work

In this section, a brief review of related work on sentiment classification methods that have been so far proposed is discussed. The focus is on sentiment classification studies to classify the text into positive or negative sentiments. Among the many studies conducted on sentiment classification using machine learning algorithms, SVM and naive bayes have been used widely for classification of online reviews (Pang et al., 2002; Wilson et al., 2005; Wang et al., 2007; Tan and Zhang, 2008; Prabowo and Thelwall, 2009), The comparative studies in the literature also showed that SVM outperformed other classifiers such as naive bayes, centroid classifier, K-nearest neighbor, winnow classifier (Tan and Zhang, 2008). Among the application domains, researchers have focused much on sentiment classification of product reviews for business intelligence (Wu et al., 2006; Tang et al., 2009; Prabowo and Thelwall, 2009). Moreover, in recent years we witnessed the advance in neural network methodology, like fast training algorithm for deep multilayer neural networks. Few researchers attempted back propagation neural network based sentiment prediction (Zhu et al., 2010; Chen et al., 2011; Sharma and Dey, 2012; Moraes et al., 2013). Experiments indicated that ANN produces superior results. Ghiassi et al. (2013) very recently used dynamic neural network for sentiment analysis on twitter sentiments. This motivated us to evaluate the use of another popular neural network based approach the PNN. From the literature work done, the PNN model is not applied so far in sentiment mining of product reviews to our knowledge. But many researchers have proved that the PNN model is more effective than other models for data classification in various other domains (Savchenko, 2013; Ciarelli and Oliveira, 2009). Also, in recent years there has been a growing interest in using ensemble learning techniques, which combine the outputs of several base classification techniques to form an integrated output, to enhance classification accuracy. Related work about ensemble methods contributing to sentiment classification are still limited compared with other research domains and more extensive experimental work is needed in this area (Wilson et al., 2006; Tsutsumi et al., 2007; Abbasi et al., 2008; Lu and Tsou, 2010; Whitehead and Yaeger, 2010; Xia et al., 2011; Su et al., 2013; Li et al., 2012). However, to the best of our knowledge, no work has investigated neural networks ensemble impact on improving the accuracy of feature level sentiment mining.

Though several machine learning approaches have been developed, selection of feature size results in improved performance. Previous work used various methods for selecting the features such as gradable adjectives, parts of speech as features, log likelihood ratio, information gain, mutual information, chi square test and document frequency (Hatzivassiloglou and Wiebe, 2000; Yu and Hatzivassiloglou, 2003; Dave et al., 2003; Gamon, 2004; Wang et al., 2007, 2011; Tan and Zhang, 2008). Except the work of Cambria et al. (2013) the literature does not contribute much work using the popular feature reduction method PCA in sentiment classification. Thus in the proposed method, the probabilistic neural network, BPN and a homogeneous ensemble of neural networks has been selected as the classification model. The effect of PCA as a feature reduction technique with neural network based sentiment classification is also investigated, as SVM is the most commonly used sentiment classification method and linear discriminant analysis is another popular statistical method used in classification problems. The performances of three neural network models are shown by comparing with two statistical models such as SVM and LDA.

8. Conclusion

Performances of neural network based approaches are compared with two statistical approaches. The homogeneous ensemble method performs better than other classification methods used. Among the individual neural network approaches used, PNN was highly robust. The performance was analyzed through the five quality parameters along with traditional techniques. The proposed approach of combining the neural network with PCA shows its superiority not only in quality measures, but also in training time. This indicates that feature reduction is an essential issue for learning methods in sentiment classification. Our experimental analysis shows that a hybrid combination of PNN and PCA could be a better solution for reducing the training time and increasing the classification performance. Our analysis also shows that the compound combination of unigram, bigram and trigram performs better for almost all the prediction models. The possible reason for the better performance of PNNs is because of the combined effect of the computational capability and flexibility, by retaining its simplicity. The prediction accuracy of the ensemble method can still be increased by increasing the number of classifier combinations. To test the limitations of the proposed method, future works could use different data domains and classification approaches probably with a data set of much a larger number of reviews.

References

Abbasi, A., Chen, H., Thoms, S., Fu, T., 2008. Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans. Knowl. Data Eng. 20 (9), 1168–1180.

Ahmed, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans. Inf. Syst. 26 (3).

Briand, L., Wust, J., 2002. Empirical studies of quality models in object-oriented systems. In: Zelkowitz, Marvin (Ed.), . In: Advances in Comput-ers, 56. Academic Press, pp. 1–44.

Cambria, E., Mazzocco, T., Hussain, A., 2013. Application of multidimensional scaling and artificial neural networks for biologically inspired opinion mining. Biol. Insp. Cogn. Archit. 4, 41–53.

Chaovalit, P., Zhou, L., 2005. Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of the 38th Annual HICSS.

Chen, Long-Sheng, Liu, Cheng-Hsiang, Chiu, Hui-Ju, 2011. A neural network based approach for sentiment classification in the blogosphere. J. Inf. 5, 313–322.

Ciarelli, Patrick Marques, Oliveira, Elias, 2009. An enhanced probabilistic neural network approach applied to text classification. In: LNCS, 5856. Springer Verlag, pp. 661–668.

Dave, K., Lawrence, S., Pennock, D.M., 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: The 12th WWW.

Gamon, M., 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 841.

Ghiassi, M., Olschimke, M., Moon, B., Arnaudo, P., 2012. Automated text classification using a dynamic artificial neural network model. Exp. Syst. Appl. 39 (12), 10967–10976.

Ghiassi, M., Skinner, J., Zimbra, D., 2013. Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Exp. Syst. Appl.

Hatzivassiloglou, V., Wiebe, J., 2000. Effects of adjective orientation and gradability on sentence subjectivity. In: International Conference on Computational Linguistics (COLING-2000).

Hu, M., Liu, B., 2004. Mining opinion features in customer reviews. In: Proceedings of 19th National Conference on Artificial Intelligence, pp. 755–760.

Joachims, 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceeding 10 European Conference on Machine Learning (ECML). Springer Verlag, pp. 137–142.

Kanmani, S., Uthariaraj, V.R., Sankaranarayanan, V., Thambidurai, P., 2007. Objected-oriented software fault prediction using neural networks. Inf. Software Technol. 49 (5), 483–492.

Li, Tao, Zhu, Shenghuo, Ogihara, Mitsunori, 2008. Text categorization via generalized discriminant analysis. Inf. Process. Manage. 44, 1684–1697.

Li, W., Wang, W., Chen, Y., 2012. Heterogeneous ensemble learning for Chinese sentiment classification. J. Inf. Comput. Sci. 9 (15), 4551–4558.

Lu, B., Tsou, B.K., 2010. Combining a large sentiment lexicon and machine learning for subjectivity classification, machine learning and cybernetics (ICMLC). In: 2010 International Conference on (IEEE), pp. 3311–3316.

Moraes, Rodrigo, Valiati, João Francisco, Gavião Neto, Wilson P., 2013. Document-level sentiment classification: an empirical comparison between SVM and ANN. Exp. Syst. Appl. 40 (2), 621–633.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques. In: EMNLP.

Prabowo, Rudy, Thelwall, Mike, 2009. Sentiment analysis: a combined approach. J. Inf. 3, 143–157.

Salton, G., Singhal, A., Mitra, M., Buckley, C., 1997. Automatic text structuring and summarization. Inf. Process. Manage. 33 (2), 193–207.

Savchenko, A.V., 2013. Probabilistic neural network with homogeneity testing in recognition of discrete patterns set. Neural Networks 46, 227–241.

Sharma, Anuj, Dey, Shubhamoy, 2012. A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. ACM SIGAPP Appl. Comput. Rev. 12 (4), 67–75.

Su, Y., Zhang, Y., Ji, D., Wang, Y., Wu, H., 2013. Ensemble Learning for Sentiment Classification, Chinese Lexical Semantics. Springer, pp. 84–93.

Tan, S.B., Zhang, J., 2008. An empirical study of sentiment analysis for Chinese documents. Expert Syst. Appl. 34 (4), 2622–2629.

Tang, H., Tan, S., Cheng, X., 2009. A survey on sentiment detection of reviews. Expert Syst. Appl. 36 (7), 10760–10773.

Tsutsumi, K., Shimada, K., Endo, T., 2007. Movie review classification based on a multiple classifier. In: The 21th Pacific Asia Conference on Language, Information and Computation (PACLIC).

Turney, P.D., 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, Philadelphia, PA, pp. 417–424.

Ur-Rahman, N., Harding, J.A., 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Syst. Appl. 39 (5), 4729–4739.

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

Wang, Suge., Li, Deyu, Song, Xiaolei, Wei, Yingjie, Li, Hongxia, 2011. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Syst. Appl. 38, 8696–8702.

Whitehead, M., Yaeger, L., 2010. Sentiment mining using ensemble classification models. In: Innovations and Advances in Computer Sciences and Engineering. Springer, pp. 509–514.

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 347–354.

Wilson, T., Wiebe, J., Hwa, R., 2006. Recognizing strong and weak opinion clauses. Comput. Intell. 22 (2), 73–99.

Wu, C.-H., Chuang, Z.-J., Lin, Y.-C., 2006. Emotion recognition from text using semantic labels and separable mixture models. ACM Trans. Asian Lang. Inf. Process. 5 (2), 165–182.

Xia, Rui, Zong, Chengqing, Li, Shoushan, 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. 181, 1138–1152.

Xia, Rui, Zong, Chengqing, Xuelei, Hu, Cambria, Erik, 2013. Feature ensemble plus sample selection: domain adaptation for sentiment classification. IEEE Intell. Syst. 28 (3), 10–18.

Ye, Q., Lin, B., Li, Y.J., 2005. Sentiment classification for Chinese reviews: a comparison between SVM and semantic approaches. In: The 4th International Inference on Machine Learning and Cybernetics ICMLC.

Ye, Q., Zhang, Z.Q., Rob, L., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. 36 (3), 6527–6535.

Yu, H., Hatzivassiloglou, V., 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the Conference on Empirical Methods in Natural.

Zhu, Jian, Xu, Chen, Wang, Han-shi, 2010. Sentiment classification using the theory of ANNs. J. China Univ. Posts Telecommun. 17 (Suppl.), 58–62.