Hindawi Publishing Corporation
Scientific Programming
Volume 2015, Article ID 961454, 15 pages
http://dx.doi.org/10.1155/2015/961454

Research Article

Feature Reduction Based on Genetic Algorithm and Hybrid Model for Opinion Mining

P. Kalaivani[1] and K. L. Shunmuganathan[2]

[1]*Department of Computer Science and Engineering, Sathyabama University, St. Joseph's College of Engineering, Chennai 600119, India*
[2]*Department of Computer Science and Engineering, RMK Engineering College, Chennai, India*

Correspondence should be addressed to P. Kalaivani; vaniraja2001@yahoo.com

Received 26 December 2014; Revised 1 April 2015; Accepted 20 April 2015

Academic Editor: Bormin Huang

With the rapid growth of websites and web form the number of product reviews is available on the sites. An opinion mining system is needed to help the people to evaluate emotions, opinions, attitude, and behavior of others, which is used to make decisions based on the user preference. In this paper, we proposed an optimized feature reduction that incorporates an ensemble method of machine learning approaches that uses information gain and genetic algorithm as feature reduction techniques. We conducted comparative study experiments on multidomain review dataset and movie review dataset in opinion mining. The effectiveness of single classifiers Naïve Bayes, logistic regression, support vector machine, and ensemble technique for opinion mining are compared on five datasets. The proposed hybrid method is evaluated and experimental results using information gain and genetic algorithm with ensemble technique perform better in terms of various measures for multidomain review and movie reviews. Classification algorithms are evaluated using McNemar's test to compare the level of significance of the classifiers.

1. Introduction

A basic task in sentiment classification is classifying the polarity of a given text in the document, sentence, or feature level, whether the expressed opinion in a review document, a sentence, or an entity feature is positive, negative, or neutral. The WWW is frequently used medium for exchanging the opinions of user reviews about the product, movie, and music. It provides a review text containing consumer opinions, emotions, and service opinions stored in websites, blogs, and web forms. Nowadays, a number of review websites, web forms, and blogs are growing rapidly. The blogs are used to store important text and the user expresses their emotions, feelings, and opinions through blogs [1].

Sentiment analysis is one of the applications of natural language processing and text analytics to identify and extract subjective information in the source materials. It aims to determine the attitude of a writer with respect to some topic or the overall polarity of a document [1–3]. The attitude may be his/her judgment, affective state, or intended emotional communication. Web sites are used to express end user opinions, emotions and sentiment about the multi products reviews, movie reviews, music and story reviews. Sentiment analysis or opinion mining plays an important role and is difficult to analyze a lot of information individually.

Sentiment analysis not only helps people but also helps in business and an organization to evaluate sentiments or opinions. Based on the behavior of the customer an opinion about a product helps organizations during the decision making process.

To automate sentiment classification, there are several approaches that have been applied to review the documents. The approaches are natural language processing, machine learning algorithms such as maximum entropy, support vector machine, Naïve Bayes, K-nearest neighbor, decision tree algorithms combined with feature selection methods to predict the polarity of the user reviews, opinions, and emotions, such as positive, negative, and neutral [4–11].

In this paper, we have applied many supervised machine learning algorithms for opinion mining. A genetic algorithm

is a search and an optimized feature selection algorithm which integrates with ensemble methods to improve the performance and overcome the limitations of traditional method. An optimization is a process of finding the best or an optimal solution for a sentiment classification.

The proposed approach is based on the machine learning approach that uses information gain feature reduction technique and optimized feature selection, genetic algorithm that incorporate bagging with TF-IDF weighting scheme. The proposed method is evaluated and experimental results using information gain, genetic algorithm with ensemble technique, indicate higher performance result. A feature reduction method based on information gain is to decide the importance of a feature in the movie review and multidomain dataset. The disadvantage of this method is to select a large number of features and does not consider duplicates in the features. It can be reduced by using an optimized feature selection. Our main objective is to design and develop a new classification algorithm which will enable improving the performance of the sentiment classification.

The rest of the paper is organized as follows. Section 2 presents state-of-the-art, related to this study. Section 3 gives problem outline. Section 4 presents our proposed feature reduction methods. Section 5 gives methodology. Section 6 presents evaluation models used in this study. In Section 7 we discuss the empirical results and Section 8 gives the conclusion of the study and future research direction of this study.

2. Related Works

Several techniques were used for opinion mining tasks in history. The following few works relate to this study. The field of machine learning has provided many models that are used to solve various sentiment classification problems.

Among them are support vector machine, Naïve Bayes, decision trees, maximum entropy, and hidden Markov models. So far, the most popular machine learning approaches used as baselines are support vector machine (SVM) and Naïve Bayes (NB) [2].

In Pang et al. [2] study several machine learning algorithms were analyzed on a movie review dataset, together with different feature selection techniques. They used a binary unigram representation of patterns and directly apply the machine learning techniques. Training patterns are represented by the presence or absence of words instead of that counting the number of occurrences of words in the documents. When the machine learning approach was applied to the document they report the best performance using SVM method with unigram text representation using a movie review dataset. They achieved the best result using SVM based in unigram. They utilized Naïve Bayes (NB), maximum entropy (ME), and support vector machines (SVM). As per the results, on the movie review dataset 82.9% accuracy was achieved, while the NB method gave lower accuracy.

In the later study Pang and Lee [3] proposes first separate subjective sentence from the rest of the text. They assume that two consecutive sentences would have a similar subjective label, as the author is inclined not to change sentence subjectivity too often. Thus, labeling all sentences as objective and subjective, they reformulate the task of finding the minimum s-t cut in a graph. They carried out experiments on the movie reviews and movie plot summaries mined from the Internet Movie Data Base (IMDB), achieving an accuracy of around 85%.

To use the prior knowledge besides a document, Mullen and Collier [5] attempted to use the semantic orientation of words defined by Pang et al. [2] and several kinds of information from the Internet and thesaurus. They evaluated the same dataset used in Pang et al. [2] study and achieved 75% accuracy with the lemmatized word unigram and the semantic orientation of words.

Wiebe et al. [6] used review data for automobiles, banks, movies, and travel destinations. She classified words into two classes (positive or negative) and counts the overall positive or negative score for the text. If the documents contain more positive than negative terms, it is assumed as a positive document; otherwise, it is negative. These classifications are based on document and sentence level classification. These classifications are useful and improve the effectiveness of sentiment classification but cannot find what the opinion holder liked or disliked about each feature.

Zhang et al. [7] use customer feedback review and product review. They use decision learning method for sentiment classification. Decision tree learning is a method for approximating discrete valued target functions, in which the learned function is represented by a decision tree. Learned trees are also re-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants.

Chen and Chiu [12] proposed a Neural Network (NN) based index, which combines the advantages of machine learning techniques and semantic orientation indices to effectively classify sentiment. Tao and Tan [13] used emotional function words instead of emotional keywords to evaluate emotional states. Hu and Liu [14] used adjective synonym sets and antonym sets in WordNet to judge the semantic orientations of adjectives.

Ye et al. [10] report an evaluation of three supervised machine algorithms of Naïve Bayes, SVM, and character based N-gram model for sentiment classification of the reviews and, in this study, they reported that all three approaches reached accuracies of at least 80% and also that SVM and N-gram approaches outperformed the Navïe Bayes approach.

Zhang et al. [15] proposed a lexicon enhanced method for sentiment classification by combining machine learning and semantic orientation approaches into one framework. Specifically, they used the words with semantic orientations as an additional dimension of features for the machine learning classifiers. In general, sentiment analysis is concerned with analysis of direction based text, that is, text containing opinions and emotions. Sentiment classification studies attempt to determine whether a text is objective

or subjective or whether a subjective text contains positive or negative sentiments. The common two-class problem involves classifying sentiments as positive or negative [3, 16]. Additional variations include classifying sentiments as opinionated/subjective or factual/objective [6]. Some studies have attempted to classify emotions, including happiness, sadness, anger, and horror, instead of the sentiments.

Xia et al. [17] ensemble framework is applied to sentiment classification tasks with the aim of integrating different feature sets and different classification algorithms to produce a more accurate classification procedure. The author has applied two types of feature sets for opinion mining and three well-known text classification algorithms, namely, Naïve Bayes, maximum entropy, and support vector machines, which are employed as a base classifiers for each of the feature sets and proposed three types of ensemble methods, namely, the fixed combination and weighted combination and the meta-classifier combination is evaluated for three ensemble strategies.

Liu et al. [18] proposed designs and developed a movie rating and review summarization system in a mobile environment. They used a sentiment classification approach based on Latent Semantic Analysis (LSA) to identify product features.

Hai et al. [19] proposed a method to identify opinion and features from online reviews and used one domain specific corpus as well as one domain-independent corpus. They used a measure called domain relevance, which characterizes the relevance of a term for a text collection. They used syntactic dependency rules to extract a list of candidate opinion and features of the domain review corpus and then estimated its intrinsic domain relevance and extrinsic domain relevance scores on the domain dependent corpus and domain specific corpus. Candidate features that are less generic are opinion features.

Kalaivani and Shunmuganathan [20, 21] examine how a classifier works with various sizes of feature set. In this study, information gain feature reduction method is applied to reduce the original feature set by removing irrelevant feature for sentiment classification of movie reviews and to select top p% ranked attributes for training the classifier.

The method also evaluated the accuracy of movie domain data sets and used different feature weight schemes along with information gain feature selection method. They compared three supervised machine learning approaches such as SVM, Naïve Bayes, and KNN for sentiment classification of movie reviews.

2.1. Motivation. Automatic classification of sentiment is important for numerous applications such as opinion mining, opinion summarization, contextual advertising, and market analysis. The sentiment classification has been modeled as the problem of training a binary classifier using reviews for positive or negative. It is a growing field of research, driven by both commercial applications and academic interest. The sentiment analysis is used for identifying the rate of accuracy in positive, negative, and neutral reviews.

Various studies show sentiment classification on product review using machine learning algorithms [3, 17, 22–24]. This helps us to conduct opinion mining on multidomain product reviews and movie reviews. Information gain is a popular feature reduction technique and is used in opinion mining [12]. The sentiment classification literature does not contribute any work using collective optimized feature reduction technique, information gain, and an ensemble method. In this study, we used information gain and optimized feature reduction technique, genetic algorithm and an ensemble method, bagged SVM, and Bayesian Boosting NB to perform the opinion mining task.

The main contribution of this study is to find the effect of unigram feature and joint feature. To build our opinion mining model, we used unigram and bigram features. In Test I only unigram is used as a feature and in Test II unigram and bigram are used as a feature for classification.

For each test various machine learning algorithms NB, LR, SVM, and ensemble methods, bagged SVM and Bayesian NB are used to conduct the experiment. The accuracy result and overall error rate are compared. The comparative results show that the hybrid model gives better result than single classifier.

3. Problem Outline

In this work, various machine learning algorithms are applied to classify the documents and to find set of opinion as positive or negative. To overcome the drawbacks such as unstable outcomes in the unigram feature selection, the IG integrated with genetic optimized feature selection for the supervised classification algorithm is formulated. Information gain feature reduction is applied to dataset to extract the relevant features for the domain. Reduced attributes are further analyzed to eliminate irrelevant attributes using the optimized feature selection based on the attribute weights. The attribute weight relation is set to top p% and the p value is set to 0.7. This section describes the opinion mining problem. The prediction model is as follows.

Input is as follows:

The review dataset D, a set of d training review dataset classifiers, is used as a learning scheme. In this work, we used three machine learning classifiers NB, LR, and SVM and hybrid model.

Output is as follows:

A predicted model.

Method is as follows:

(a) To prepare, review documents, we performed tokenization and transformed all characters to lower case, stemming, and filter stop words. Tokenization operation splits the review documents into a sequence of words. Filter stopword operation removes every word which equals stop words from review documents using the predefined stop word list.

(b) Feature measure scheme TF-IDF to convert text representation vector is as follows:

(i) Test I uses unigram with TF-IDF.

(ii) Test II uses unigram and bigram with TF-IDF.

Table 1: Description of multidomain dataset and movie review dataset (unigram).

Dataset	Number of stratified samples	Positive reviews	Negative reviews	Total attributes	Total attributes (weight by IG)	Total attributes (optimized selection)		
						NB, LR, SVM	BSVM	BNB
Book	191	100	91	377	341	264	231	231
DVD	199	99	100	364	318	255	228	219
Electronics	200	100	100	274	209	192	169	175
Kitchen	190	99	91	207	193	145	123	131
Movie	200	100	100	1569	1568	1098	1069	958

Table 2: Description of multidomain dataset and movie review dataset (joint feature).

Dataset	Number of stratified samples	Positive reviews	Negative reviews	Total attributes	Total attributes (weight by IG)	Total attributes (optimized selection)		
						NB, LR, SVM	BSVM	BNB
Book	191	100	91	407	367	285	297	250
DVD	199	99	100	392	342	274	279	250
Electronics	200	100	100	303	236	212	196	192
Kitchen	190	99	91	234	214	164	146	145
Movie	200	100	100	1719	1718	1203	1069	1049

(c) The stratified sampling creates a random review subset of the whole document.

(d) Evaluate the performance of the SVM classifier with and without IG and optimize selection.

(e) Each test uses an IG feature reduction technique, optimized feature selection, and genetic algorithm that incorporates bagging and Bayesian with TF-IDF weighting scheme.

(f) Calculate the relevance of attributes based on information gain and assign attribute weights to them accordingly.

(g) Select attributes from input words whose weight satisfies the specified condition (with highest weight top 7%) with respect to the input weight.

(h) Remove useless attributes.

(i) The proposed model is used as a training dataset for learning models.

  (1) Develop a model using Naïve Bayes.
  (2) Develop a model using logistic regression.
  (3) Develop a model using support vector machine.
  (4) Develop a model using IG, optimized feature selection (GA), and an ensemble bagging technique incorporated support vector machine.
  (5) Develop a model using IG, an optimized feature selection (GA) and Ensemble Bayesian is boosting technique incorporated Naïve Bayes.

(j) Effectiveness of each model is evaluated and prediction model is compared with the baseline method.

3.1. Corpora Description. The user's opinions are the valuable sources of data which helps to improve the quality of service rendered. Blogs, review sites, and microblogs are some of the platforms where user expresses his/her opinions.

To conduct the study, movie reviews and multidomain datasets are considered here. The Cornell movie-review corpora (http://www.cs.cornell.edu/people/pabo/movie-review-data) consists of movie reviews dataset which contains 1000 positive reviews and 1000 negative reviews. The multidomain dataset (http://www.cs.jhu.edu/~mdredze/datasets/sentiment) contains product reviews, book, DVD, electronics, and kitchen; each of these contains 1000 positive and 1000 negative reviews.

In order to obtain a reduced feature for our problem, we applied stratified sampling for each domain. The number of samples, number of positive reviews and negative reviews, and total number of attributes, attributes reduced after applying information gain, and reduced attributes after applying an optimized feature reduction for each classification algorithms are given in Tables 1 and 2. The properties of the data source for unigram and joint feature word vector models are developed and Test I uses only unigram; Test II is represented as a word vector that uses unigram and bigram attributes.

4. Proposed Sentiment Classification Using Genetic Algorithm

The main objective of the feature selection is to reduce the number of features and the computational cost and to improve the performance of classification. It has been proved that feature reduction method is to remove the irrelevant and redundant feature and also increase the learning task, so it improves the efficiency of sentiment classification.

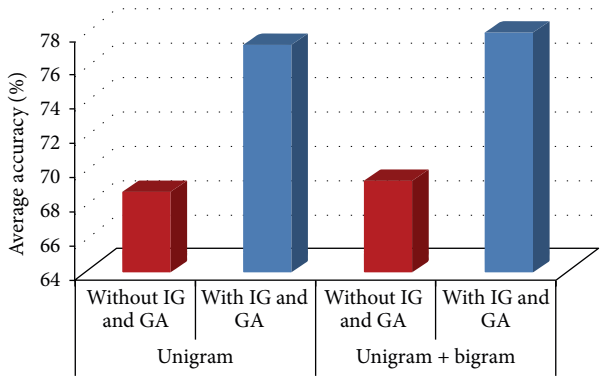

Figure 1: An accuracy of unigram and joint feature.

In this study, we use movie review dataset and multidomain dataset for evaluation, which involves splitting the available dataset into a training set and a testing set. We used a genetic algorithm that incorporates various machine learning algorithms to improve the performance of feature selection. Generally, we applied the NB, LR, and SVM algorithms to the dataset in the training set and evaluate the resulting model using the dataset in the test set. Most of the existing work shows that support vector machine and Naïve Bayes are perfect methods in sentiment classification [2, 3, 22, 25–27]. So SVM and NB classifiers are used as base classifiers in our approaches.

The accuracy is measured using the SVM classification algorithm with and without information gain and an optimized feature reduction (GA). Figure 1 shows the performance of SVM algorithm. The accuracy is better with feature reduction IG and an optimal selection. Most of the work shows that SVM outperformed the other machine learning algorithms [2–5, 22, 26]. In our work SVM is one of the base classifiers.

4.1. Sentiment Analysis with Different Learning Tests. Generally, before building the sentiment analysis, we first need to decide which learning model should be used to construct feature selection model. We propose the use of genetic algorithm to improve the performance of opinion mining and to address the problems in sentiment analysis. The framework consists of two steps: learning type evaluation and sentiment analysis.

In the first test, learning type evaluation stage, the performance of the learning types is evaluated with multidomain review dataset and movie review dataset to decide which learning type performs better for sentiment analysis and we need to select the best learning type from a set of different learning types. The ten different learning types were considered according to the feature selection and machine learning algorithms:

(a) Two-feature selection is as follows:

(i) Information gain is one of the important feature selection measures used in sentiment classification, which outperformed the other feature selection methods [4, 10, 16]. It is based on the value or weight of information contained in reviews, which select important feature with respect to class attributes. The weight of each attribute with respect to the class is calculated by using information gain for each attribute which will vary from 0 to 1. The higher the weight of an attribute, the greater the information gain.

(ii) In this study, genetic algorithm uses heuristic method to assign weights to various sentiment words or attributes [25, 28].

(b) Three machine learning algorithms are as follows: NB, LR, and SVM.

(c) Two ensemble methods are as follows: bagging and Bayesian boosting.

(d) Ten learning types are as follows:

The combination of IG, GA, feature selection, three machine learning algorithms, and ensemble techniques gives a total of 10 different learning tests. In model I word vector model is represented by unigram and model II uses unigram and bigram attributes.

(i) Model I:

unigram + IG + GA + NB;
unigram + IG + GA + LR;
unigram + IG + GA + SVM;
unigram + IG + GA + BSVM;
unigram + IG + GA + BNB.

(ii) Model II:

joint feature + IG + GA + NB;
joint feature + IG + GA + LR;
joint feature + IG + GA + SVM;
joint feature + IG + GA + BSVM;
joint feature + IG + GA + BNB.

4.2. Improving the Efficiency of Hybrid Genetic Algorithm. Incorporating a local search into a genetic algorithm can increase the efficiency of the algorithm. The efficiency of the searching process is increased in terms of the time required to reach a global optimal solution and memory needed to process the population. The major steps in this study are as follows.

4.2.1. Initial Population. In GA, the initial populations of n strings are randomly generated and collection of such strings is called initial population [23]. The information gain feature weights are used as the final strings in the initial population. The information gain solution features are used as the solution string in the initial population. The solution features are represented using binary string character. Specifically 1 represents a selected attribute or feature and 0 represents the discarded one. Generate random population of n individual. Each attribute is switched on with the probability Pi. In this study, the population size is set to 50 and the Pi value is set to 0.1.

4.2.2. Selection. To evaluate the quality of each solution, classification accuracy is used as the fitness function. For each solution in the population, tenfold cross validation with classification algorithm is used to assess the fitness of that particular solution. Solution for the next iteration is selected probabilistically and in this study tournament is used as the selection scheme. The size of tournament specifies the fraction of the current population, which should be used as a tournament member. The size of the tournament is set to 0.05. There are several population replacement methods such as generational replacement method and steady-state method. In generational replacement method, the entire population is replaced in every iteration, but, in steady state, fraction of the population is replaced in every iteration.

4.2.3. Crossover. Crossover is the process of exchange of information between two parents to produce a new offspring. Choose two individuals from the population and perform crossover based on a crossover probability P_c. The probability is set to 0.6. Different crossover types such as single point, uniform, and shuffle crossover are used. We use uniform crossover by selecting two individuals and swapping substring at a randomly determined crossover point x. If the mixing ratio is 0.5, then half of the genes in the offspring will come from parent 1 and half will come from parent 2. Mutation is randomly mutated individual feature characters in a solution string based on a fixed probability P_m. The mutation probability is set to 0.01.

5. Methodology

Opinion mining is conducted at any of the three levels, the document level, the sentence level, and the attribute level. In this study, we applied supervised machine learning models for sentiment classification of reviews for the selected movie reviews and product reviews. The models are NB, LR, and SVM algorithm together with genetic algorithm which uses information gain as feature reduction technique. In this work NB, LR, SVM, and hybrid model are applied to classify the documents and find a set of opinion as positive or negative.

5.1. Naïve Bayes. The basic idea is to find the probabilities of the categories given a review document by using the joint probabilities of words and categories. It is based on the assumption of words being conditionally independent.

The starting point is the Bayes theorem for conditional probability, stating that, for a given data point w and class C, let R be the training dataset and their associated class labels and each dataset is represented by attribute space vector $W = (w_1, w_2, \ldots, w_n)$. The classification is to derive the maximum posteriori,

$$P\left(\frac{C_i}{W}\right) = P\left(\frac{W}{C_i}\right) P\left(C_i\right). \quad (1)$$

Class C_1 is positive and C_2 is negative. The probability of each of its attributes occurring in a given class is independent, when we estimate the probability of w as follows:

$$P\left(\frac{C}{w}\right) = P\left(C\right) \cdot \prod P\left(\frac{w}{C}\right). \quad (2)$$

Training a Naïve Bayes classifier, therefore, requires calculating the conditional probabilities of each attribute occurring in the classes, which can be estimated from the training dataset.

5.2. Logistic Regression. Logistic regression is one of the standard techniques used for applying statistics and discrete data analysis. It is based on maximum likelihood estimation. It is used to predict positive or negative class response from positive and negative attributes and predicting the outcome of the class label based on the positive or negative attributes.

5.3. Support Vector Machine. In this work, support vector machine classification algorithm is applied to classify the review documents and find a set of opinion as positive or negative. It has been shown that this is an effective classification algorithm and also is used in sentiment analysis. This algorithm outperformed the other classification algorithms [29]. The SVM finds hyper plane using support vectors. This approach was developed by Vladimir Vapnik, Bernhand Boser, and Isabelle Guyana in 1992.

5.4. Bagging Technique. Bagging technique is used to improve the classification model in terms of classification accuracy. The basic idea of this technique is to construct members from the training dataset. The bootstrap aggregating splits the training dataset into several new training datasets by sampling and model is built based on the new training dataset. For the training dataset T of size m, bagging generates n new training dataset by sampling with replacement. The classifier is trained on each training dataset and the new training dataset is equal to the original training dataset, so that the bagging technique produces better results than a single model [17, 24, 30]. We obtained best accuracy, using unigram, TF-IDF feature weighting scheme, and information gain feature selection with 10-fold cross validation. The idea of improving the supervised classification by randomly generated training dataset was proposed by Leo Breiman in 1994. The bagging is also referred as bootstrap aggregation.

We used 10-fold cross validation to measure the performance of sentiment classification. It has two subprocesses, one is a training subprocess and another one is testing subprocess. We considered movie review dataset and multidomain dataset D, of d documents. In the training subprocess for each iteration i ($i = 1, 2, \ldots, k$) a training dataset D of d document samples with replacement. Some of the original dataset D may not be included in D_i. This method generates set of classifier models $M_1, M_2, M_3, \ldots, M_k$. The bagging method separates training dataset into several new training datasets by random sampling. The training subprocess is used for training a model and the trained model is applied in the testing phase. In the first iteration $D_2, D_3, \ldots, D_k$ are

jointly served as the training set in order to obtain a first model, which is tested on D_1; the second iteration is tested on subsets $D_1, D_3, \ldots, D_k$, and tested on D_2 and so on. During the testing subprocess, the performance of the model is generated.

The bagged classifier M^* counts the vote and assign the class with the most votes to testing data set.

The bagging algorithm is as follows.

Input is as follows:

(i) The review dataset D, a set of d training review dataset.
(ii) Number of k models in the classifier.
(iii) Classifier is used as a learning scheme (in this we used SVM machine learning classifier).

Output is as follows:

(i) a composite model M^*.

Method is as follows:

(i) Training subprocess:

For $i = 1$ to k do//create k models.
Create bootstrap sample D_i by sampling original review dataset D with replacement use D_i to derive a model M_i.
End for

(ii) Testing subprocess:

If classification then

(i) Let each of the k models classify testing dataset and return the majority vote.

5.5. Bayesian Boosting. The Bayesian boosting algorithm is an iterative machine learning algorithm based on Bayes' theorem which is used to improve the performance accuracy. This method is an ensemble of classifiers for product review attributes. At each iteration the training data set is reweighted. We apply the Naïve Bayes algorithm several times, and all the models are combined into a single model. In this process, the number of iterations is set to be 10 [17, 31].

6. Performance Evaluations

In this study, we use movie review dataset and multidomain dataset and the evaluation involves splitting the available dataset into a training set and a testing set. We use a genetic algorithm that incorporates hybrid model to improve the performance of feature selection. Generally, we applied the NB, LR, and SVM algorithms to the dataset in the training set and evaluate the resulting model using the dataset in the test set.

The cross validation method involves partitioning the dataset randomly into 10-fold. We use one partition as a testing set and the remaining partitions to form training set. We repeat this process 10 times, each of the partitions as the testing dataset and the remaining partitions to form a training set. In this work, four evaluation measures, accuracy, overall error rates, type I error, type II error, sensitivity, and specificity are used to test the effectiveness of opinion mining. Once we selected an algorithm and an evaluation methodology, we need to select a performance metric. For two class problems, a test case will be either positive or negative. This yields four quantities that we can compute by applying a model to a set of test cases, as shown in Table 6.

For a set of test cases, let A be the number of times the model predicted positive when the reviews label is positive, let B be the number of times the model predicted negative when the reviews label is positive, let C be the number of times the model predicted positive when the reviews label is negative, and let D be the number of times the model predicted negative when the reviews label is negative. Given these counts, we can define a variety of common performance metrics. The accuracy or recognition rate of a classifier on a test review is the percentages of the test dataset that are correctly classified by the classifier as explained in the following:

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}. \tag{3}$$

An overall error rate or misclassification refers to the number of wrongly classified reviews by the total number of sample review. Type I error refers to negative sample reviews that were wrongly classified as positive reviews. Type II error refers to positive sample reviews that were wrongly classified as negative reviews:

Overall error rate (%) = $(C + B)$/total number of samples.

Type I error rate (%) = C/total number of positive samples.

Type II error rate (%) = B/total number of negative samples.

7. Results and Discussion

To evaluate our model, we used Cornell movie review datasets and the multidomain dataset which are frequently used in the sentiment classification. The multidomain dataset contains 1000 positive and 1000 negative documents. The movie review dataset contains 1000 positive reviews and 1000 negative reviews. It is a challenging task because the reviewers use a lot of comparisons and sometimes used an unclear language. The performance results are shown in Tables 3, 4, 5, 7, and 8.

7.1. Comparison of Classifiers. Many classification algorithms are available for sentiment classification such as SVM, NB, KNN, maximum entropy, and decision tree. In this study, we used three classification algorithms NB, LR, and SVM and ensemble method along with IG and optimized feature reduction method; among all these methods bagged SVM is shown to perform better. The performance for each classification is shown in Tables 3–5, 7, and 8. The best accuracy value compared to the baseline accuracy is shown with an up arrow.

Table 3: Classification performance of book domain reviews.

Class/method	NB			LR			SVM			BSVM			BNB		
	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)
Unigram															
Predicted neg.	57	23	57.00	80	38	80.00	64	21	64.00	74	16	74.00	70	17	70.00
Predicted pos.	34	77	84.61	11	62	68.13	27	79	86.81	17	84	92.30	21	83	91.20
Avg. accuracy		70.21%			74.47%			74.95%			82.68% (↑)			80.18%	
Error rate		29.84%			25.65%			25.13%			17.27% (↓)			19.89%	
Unigram + bigram															
Predicted neg.	55	24	55.00	75	32	75.00	65	20	65.00	73	14	73.00	72	21	79.00
Predicted pos.	36	76	83.51	16	68	74.72	26	80	87.91	18	86	94.50	19	79	86.81
Avg. accuracy		68.66%			74.97%			75.97%			83.21% (↑)			79.11%	
Error rate		31.41%			25.13%			24.00%			16.75% (↓)			20.94%	

Table 4: Classification performance of DVD domain review (unigram).

Class/method	NB			LR			SVM			BSVM			BNB		
	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)
Unigram															
Predicted neg.	82	21	82.00	86	22	86.00	75	21	75.00	86	20	86.00	88	17	88.00
Predicted pos.	18	78	78.78	14	77	77.77	25	78	78.78	14	79	79.79	12	82	82.82
Avg. accuracy		80.42%			81.95%			76.87%			82.92%			85.39% (↑)	
Error rate		19.59%			18.09%			23.11%			17.08%			14.57% (↓)	
Unigram + bigram															
Predicted neg.	81	22	81.00	91	33	91.00	79	23	79.00	80	20	80.00	86	19	86.00
Predicted pos.	19	77	77.77	09	66	66.66	21	76	76.76	20	79	79.79	14	80	80.80
Avg. accuracy		79.42%			78.29%			77.89%			79.92%			83.47% (↑)	
Error rate		20.60%			21.10%			22.11%			20.10%			16.58% (↓)	

Table 5: Classification performance of electronics domain reviews.

Class/method	NB			LR			SVM			BSVM			BNB		
	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)
Unigram															
Predicted neg.	83	17	83.00	88	34	88.00	75	23	75.00	82	14	82.00	87	13	87.00
Predicted pos.	17	83	83.00	12	66	66.00	25	77	77.00	18	86	86.00	13	87	87.00
Avg. accuracy		83.00%			74.47%			76.00%			84.00%			87.00% (↑)	
Error rate		17.00%			23.00%			24.00%			16.00%			13.00% (↓)	
Unigram + bigram															
Predicted neg.	85	19	85.00	89	29	89.00	75	19	75.00	79	11	79.00	87	11	87.00
Predicted pos.	15	81	81.00	11	71	71.00	25	81	81.00	21	89	89.00	13	89	89.00
Avg. accuracy		83.00%			80.00%			78.00%			84.00%			87.00% (↑)	
Error rate		17.00%			20.00%			22.00%			16.00%			12.00% (↓)	

7.2. Performance of Individual Classifier. In this study, we use accuracy and overall error rate to evaluate our proposed approach on the movie review data set and multidomain dataset. Information gain feature selection is used to reduce feature vector space and TF-IDF feature weighting schemes were utilized and selected top p percent attributes with higher weights are selected for training the classifier where the p value is set to 0.7. All the experiments were validated using 10-fold cross validation. Tables 3–5, 7, and 8 show the experimental results when using the classifier together

Table 6: Quantities computed from a test set for a two-class problem.

	Actual negative reviews	Actual positive reviews
Predict negative	*D*	*B*
Predict positive	*C*	*A*

with genetic algorithm. It is an optimized feature reduction technique. Information gain is selected as a feature reduction method, because it outperforms the other feature reduction methods [23]. The classification results of NB show that accuracy result is comparatively lesser than all other individual classifiers and hybrid model. The overall error rate is higher than all other results. NB is not an efficient algorithm on unigram and joint feature. The reason for higher error rate performance is that all features are independent. Type I error of NB is higher than other classifiers. This shows that this model predicts negative reviews that were incorrectly classified as positive reviews for unigram and joint feature. Type II error is lesser than LR but higher than the other classifiers.

The classification results of LR show that accuracy result is higher than NB model but lesser than other classifier. Type I error rate is comparatively lesser than all other classifiers, which indicates that positive reviews were correctly classified as positive reviews, but type II error rates are higher than all other classifiers which indicates that positive reviews were incorrectly classified as negative reviews. Tables 9 and 10 show the results of type I error and type II error in percentage. The classification results obtained from book reviews are given in Table 3. In Table 3, the accuracy results of bagged SVM show that it is comparatively higher than other classifiers. The overall misclassification rate is comparatively lesser than all four classifiers. This indicates that the bagged SVM model predicts positive reviews more accurately than negative reviews for unigram feature and joint feature.

The performance results are compared and bagged SVM classifier result is much better than other classifiers. The bagged SVM achieves best accuracy value of 83.21% for book review using joint feature. The hybrid model Bayesian achieves best accuracy value of 85.39% for DVD reviews, 87.00% for electronics reviews, 81.58% for kitchen reviews with bagged SVM, and 89.50% for movie reviews using unigram. We give the results of five classifiers (NB, LR, SVM, bagged SVM, and Bayesian NB) and observe the performance of different classifiers.

7.3. Statistical Significance Test. We applied McNemar's statistical test to compare the performance of classifiers [31–33]. The comparisons of statistical test result show that the bagging method performs better than other classifiers. In Table 11, the C_{ff} denotes the count of the number of times that both classifiers failed.

The C_{ff} denotes the count of the number of times that both classifiers failed. Both classifiers predict positive reviews as negative reviews and vice versa. The C_{sf} denotes the count of the number of times that classifier A succeeded but classifier B failed; that is, classifier A predicts positive reviews as positive reviews and negative reviews as negative reviews, but classifier B predicts positive reviews as negative reviews and vice versa. The C_{fs} denotes the count of the number of times that classifier B succeeded but classifier A failed; that is, classifier B predicts positive reviews as positive reviews and negative reviews as negative reviews but classifier A predict positive reviews as negative reviews and vice versa. The C_{ss} denotes the count of the number of times that both classifiers succeeded. Both classifiers predict positive reviews as positive reviews and vice versa.

The null hypothesis and alternative hypothesis are

H_0: both classifiers perform similarly,

H_1: one of the classifier performs differently.

The McNemar test statistics is

$$z \text{ score} = \frac{\left(\left|C_{sf} - C_{fs}\right| - 1\right)}{\sqrt{C_{sf} + C_{fs}}}. \tag{4}$$

When z value is zero, two classifiers perform similarly; when the z value is increased one of the classifiers performs differently. Based on the z value, we should accept H_0 or reject H_1 or vice versa. We need to decide which classifier performed better based on the C_{sf} and C_{fs} values of the two classifiers. If C_{fs} value is smaller than C_{sf}, classifier B is said to perform better than classifier A.

In Tables 12–16, the symbol "$*$" denotes that classifier A performed better than classifier B because C_{sf} value is smaller than C_{sf} value. In Tables 12–16, the symbol "$**$" denote that classifier B performed better than classifier A because C_{sf} value is smaller than C_{sf}.

By looking at the McNemar's test result for the multidomain reviews and movie reviews (see Tables 12–16) it can be observed that BSVM has produced significantly better results than NB, LR, SVM, and BNB. H_1 is accepted with level of significance at 5% right tailed test. SVM and LR classifiers performed better than NB classifier. BNB classifier performed better than NB, LR, and SVM classifiers. For DVD, electronics and movie reviews LR classifier works better than SVM. For book reviews and movie reviews BNB classifier works better than BSVM.

To compare effectiveness of ensemble technique and other classifiers for opinion mining, the performance of SVM was considered as a baseline. An improvement of different models was calculated as

$$\text{Improvement} = \frac{\text{Average Value}_{OC} - \text{Average Value}_{SVM}}{\text{Average Value}_{SVM}}. \tag{5}$$

As shown in Figure 2, the hybrid model gives best result using SVM as base classifier. The positive value indicates that a hybrid model has an increasing average accuracy with respect to SVM. As shown in Figures 3–5, the negative value indicates that a hybrid model has a decreasing error rate with respect to baseline model SVM.

Table 7: Classification performance of kitchen domain review.

Class/Method	NB			LR			SVM			BSVM			BNB		
	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)
Unigram															
Predicted neg.	65	33	71.42	75	49	82.41	59	16	64.84	68	12	74.73	69	23	75.82
Predicted pos.	26	66	66.66	16	50	50.00	32	83	83.83	23	87	87.87	22	76	76.76
Avg. accuracy		73.03%			65.79%			74.74%			81.58% (↑)			76.32%	
Error rate		36.31%			36.84%			24.73%			18.42% (↓)			23.68%	
Unigram + bigram															
Predicted neg.	67	30	73.62	72	51	79.12	63	19	69.23	64	10	70.32	74	24	81.31
Predicted pos.	24	69	69.69	19	48	48.48	28	80	80.80	27	89	89.89	17	75	75.75
Avg. accuracy		75.26%			63.16%			75.26%			80.53% (↑)			78.42%	
Error rate		28.42%			36.84%			24.73%			19.47% (↓)			21.57%	

Table 8: Classification performance of movie domain reviews.

Class/method	NB			LR			SVM			BSVM			BNB		
	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)	Actual neg.	Actual pos.	Acc. (%)
Unigram															
Predicted neg.	66	24	66.00	87	11	87.00	86	20	86.00	89	10	89.00	76	15	76.00
Predicted pos.	34	76	76.00	13	89	89.00	14	80	80.00	11	90	90.00	24	85	85.00
Avg. Accuracy		71.00%			88.00%			83.00%			89.50% (↑)			80.50%	
Error rate		29.00%			12.00%			17.00%			10.50% (↓)			19.50%	
Unigram + bigram															
Predicted neg.	65	20	65.00	86	10	86.00	84	20	84.00	89	10	89.00	73	13	73.00
Predicted pos.	35	80	80.00	14	90	90.00	16	80	80.00	11	90	90.00	27	87	87.00
Avg. Accuracy		72.50%			88.00%			82.00%			89.50% (↑)			80.00%	
Error rate		27.50%			12.00%			18.00%			10.50% (↓)			20.00%	

Table 9: Result of type I error in percentage.

Dataset/method	Unigram, type I error (%)					Joint features, type I error (%)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
Book	34.00	11.00	27.00	17.00	21.00	36.00	16.00	26.00	18.00	19.00
DVD	18.18	14.14	25.25	14.14	12.12	19.19	9.09	21.21	20.20	14.14
Electronics	17.00	12.00	25.00	18.00	13.00	15.00	11.00	25.00	21.00	13.00
Kitchen	26.26	16.16	32.32	23.23	22.22	24.24	19.19	28.28	27.27	17.17
Movie	34.00	13.00	14.00	11.00	24.00	35.00	14.00	16.00	11.00	27.00

Table 10: Result of type II error in percentage.

Dataset/method	Unigram, type II error (%)					Joint features, type II error (%)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
Book	25.27	41.75	23.07	17.58	18.68	26.37	35.16	21.97	15.38	23.07
DVD	21.00	22.00	21.00	20.00	17.00	22.00	33.00	23.00	20.00	19.00
Electronics	17.00	34.00	23.00	14.00	13.00	19.00	29.00	19.00	11.00	11.00
Kitchen	41.75	53.84	17.58	13.18	25.27	32.96	56.04	20.87	10.98	26.37
Movie	24.00	11.00	20.00	10.00	15.00	20.00	11.00	20.00	10.00	13.00

Table 11: Possible result of two classifiers.

	Classifier A failed (actual neg.)	Classifier A succeeded (actual pos.)	Row total
Classifier B failed (Actual neg.)	C_{ff}	C_{sf}	$C_{ff} + C_{sf}$
Classifier B succeeded (actual pos.)	C_{fs}	C_{ss}	$C_{fs} + C_{ss}$
Column total	$C_{ff} + C_{fs}$	$C_{sf} + C_{ss}$	n

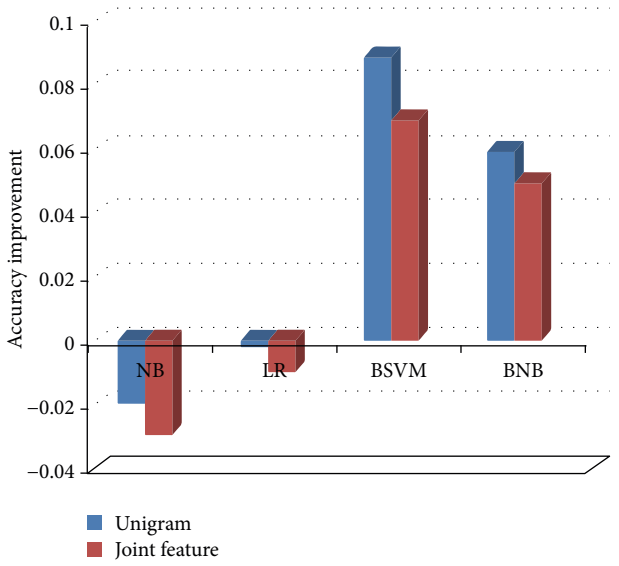

Figure 2: Accuracy improvement.

The confusion matrix of sentiment classification of the two-class multidomain dataset and movie reviews using genetic bagging is tabulated in Table 17. The accuracy is estimated by means of the confusion matrices. The accuracy and an overall error rate are calculated for genetic bagging with different attribute weight values for two-class review dataset. In the sentiment classification exercise, feature weight relations greater than 0.100, 0.200, 0.300, 0.400, and 0.500 are used. Here, the number of individuals in the population pool for the GA algorithm is 50. The convergence of accuracy value reaches in 150 generations. So, the maximum number of generations is 150. The resulting multidomain review dataset classification was more accurate with the increase in attribute weight value. The proposed hybrid method results using information gain and genetic incorporated baggingtechniqueusing SVM as a base classifier performs better in terms of accuracy.

To compare effectiveness of genetic bagging technique for opinion mining the performance of genetic algorithm without bagging was considered as baseline. As shown in Figure 6, the hybrid model gives the best result using SVM as base classifier. The positive value indicates that hybrid model has increasing average accuracy with respect to GA.

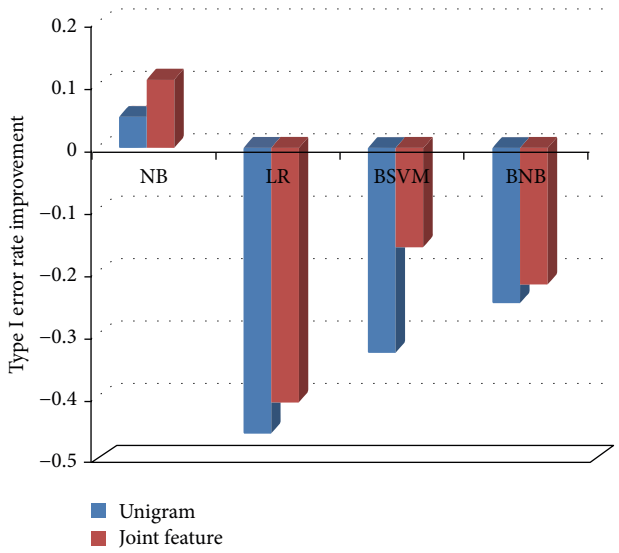

Figure 3: Type I error improvement.

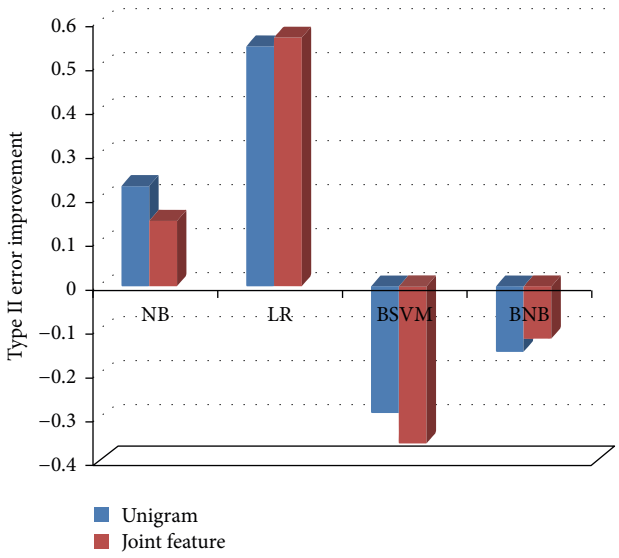

Figure 4: Type II error improvement.

7.4. Threats to Validity. In this work, we used only bag-of-words, information gain, and optimized feature reduction that incorporates ensemble methods of machine learning approaches that used to improve classification performance. It considers only positive reviews and negative reviews and does not consider neutral reviews for sentiment classification. Multidomain dataset and movie review datasets are imbalanced; large datasets should be considered for further validate the result of the study. In future, attribute construction based on other feature reduction methods should be considered.

8. Conclusions

The main aim of the study is to improve the performance of opinion mining. We proposed an optimized feature reduction

Table 12: McNemar's test results for book reviews.

Classifier B	Unigram (classifier A)					Joint feature (classifier A)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
NB		0.767*	0.870*	2.400*	1.468*		1.177*	1.381*	2.184*	1.995*
LR			0.051*	1.586*	1.074*			0.756*	2.167*	1.360*
SVM				1.484*	1.025*				1.233*	0.562*
BSVM					0.460**					0.616**
BNB										

Table 13: McNemar's test results for DVD reviews.

Classifier B	Unigram (classifier A)					Joint feature (classifier A)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
NB		0.250*	0.651*	0.451*	0.952*		0.050**	0.251**	0.050*	0.752*
LR			0.972**	0.150*	0.651*			0.150**	0.150*	0.852*
SVM				1.153*	1.654*				0.351*	1.053*
BSVM					1.037*					0.857*
BNB										

Table 14: McNemar's test results for electronics reviews.

Classifier B	Unigram (classifier A)					Joint feature (classifier A)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
NB		1.150**	1.350**	0.150*	0.750*		0.550**	0.950**	0.150*	0.950*
LR			0.150**	1.350*	1.950*			0.350**	0.750*	1.550*
SVM				1.450*	2.150*				1.150*	1.950*
BSVM					0.550*					0.750*
BNB										

Table 15: McNemar's test results for kitchen reviews.

Classifier B	Unigram (classifier A)					Joint feature (classifier A)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
NB		0.564*	1.077*	2.411*	1.385*		1.590**	0.307*	1.693*	1.282*
LR			1.693*	3.027*	2.001*			2.308*	3.335*	2.924*
SVM				1.282*	0.256*				0.974*	0.564*
BSVM					0.974**					0.204*
BNB										

Table 16: McNemar's test results for movie reviews.

Classifier B	Unigram (classifier A)					Joint feature (classifier A)				
	NB	LR	SVM	BSVM	BNB	NB	LR	SVM	BSVM	BNB
NB		3.350*	2.350*	3.650*	1.850*		3.050*	1.850*	3.350*	1.450*
LR			0.950**	0.250*	1.450**			1.150**	0.250*	1.550**
SVM				0.740*	0.938**				1.450*	0.350**
BSVM					1.606**					1.850**
BNB										

that incorporates ensemble methods of machine learning approaches that uses information gain and genetic algorithm as feature reduction techniques to improve classification performance. The results show that feature selection based on genetic algorithm along with an ensemble approach outperformed the other approaches. We conducted comparative study experiments on multidomain dataset and movie review dataset in opinion mining. The effectiveness of single classifiers, Naïve Bayes, logistic regression, support vector machine, and ensemble technique for opinion mining, is

Table 17: Confusion matrix of sentiment classification of the two-class reviews using GA algorithm based bagging technique.

Class/attribute weight	Attribute weight ≥ 0.510		Attribute weight ≥ 0.400		Attribute weight ≥ 0.300		Attribute weight ≥ 0.200		Attribute weight ≥ 0.100	
	Actual neg.	Actual pos.	Actual neg.	Actual pos.	Actual neg.	Actual pos.	Actual neg.	Actual pos.	Actual neg.	Actual pos.
Book										
Predicted neg. (type II error)	43	08 (8.79)	61	05 (5.49)	68	09 (9.89)	78	12 (13.18)	82	09 (9.89)
Predicted pos. (type I error)	48 (48.00)	92	30 (30.00)	95	23 (23.00)	91	13 (13.00)	88	09 (9.00)	91
Overall error rate (%)	29.31		29.31		16.75		13.08		09.42	
Average accuracy (%)	70.76		70.76		83.29		86.92		90.63	
DVD										
Predicted neg. (type II error)	71	09 (9.00)	70	07 (7.00)	72	07 (7.00)	80	13 (13.00)	87	11 (11.00)
Predicted pos. (type I error)	29 (29.29)	90	30 (30.30)	92	28 (28.28)	92	20 (20.20)	86	13 (13.13)	88
Overall error rate (%)	19.09		18.59		17.58		16.58		12.06	
Average accuracy (%)	80.95		81.42		82.45		83.47		88.00	
Electronics										
Predicted neg. (type II error)	48	05 (5.00)	68	05 (5.00)	73	02 (2.00)	87	08 (8.00)	90	07 (7.00)
Predicted pos. (type I error)	52 (52.00)	95	32 (32.00)	95	27 (27.00)	98	13 (13.00)	92	10 (10.00)	93
Overall error rate (%)	28.50		18.50		14.00		10.50		08.50	
Average accuracy (%)	71.50		81.50		85.50		89.50		91.05	
Kitchen										
Predicted neg. (type II error)	79	45 (49.45)	77	31 (34.06)	81	25 (27.47)	80	21 (23.07)	72	10 (10.98)
Predicted pos. (type I error)	12 (12.12)	54	14 (14.14)	68	10 (10.10)	74	11 (11.11)	78	19 (19.19)	89
Overall error rate (%)	30.00		23.68		18.42		16.84		15.26	
Average accuracy (%)	70.00		76.32		81.58		83.16		84.74	
Movie										
Predicted neg. (type II error)	62	20 (20.00)	76	16 (16.00)	91	9 (9.00)	94	4 (4.00)	97	3 (3.00)
Predicted pos. (type I error)	38 (38.00)	80	24 (24.00)	84	9 (9.00)	91	6 (6.00)	96	3 (3.00)	97
Overall error rate (%)	29.00		20.00		07.00		05.00		03.00	
Average accuracy (%)	71.00		80.00		91.00		95.00		97.00	

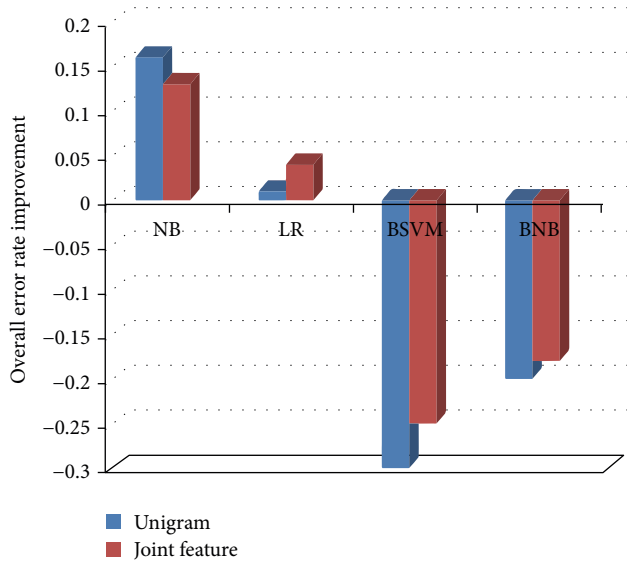

Figure 5: Overall error rate improvements.

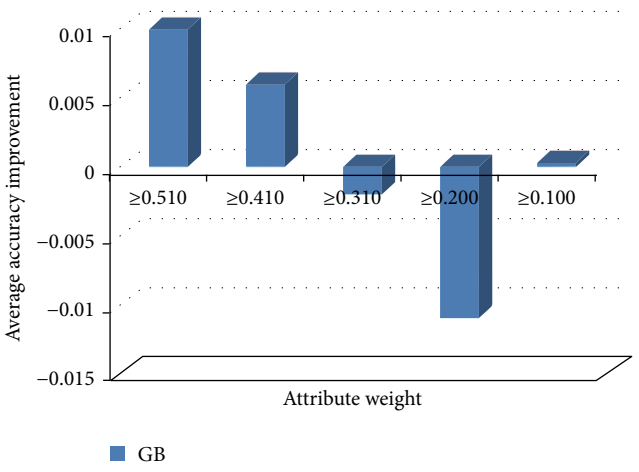

Figure 6: Average accuracy improvement (GA versus proposed GB).

compared on five datasets. The proposed hybrid method is evaluated and experimental results using information gain and genetic algorithm with ensemble technique perform better in terms of various measures for movie, book, DVD, electronics, and kitchen reviews. Five classification algorithms are evaluated using McNemar's test to compare the level of significance of the classifiers. A direction for future work is to study the performance of feature selection methods on different machine learning classifiers and to evaluate the model for sentiment analysis with other domain reviews.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pa, USA, 2002.

[2] B. Pang, L. Lee, and S. Vaithyanathain, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 79–86, 2002.

[3] B. Pang and L. Lee, "A sentimental education: sentimental analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 271–278, 2004.

[4] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373, Association for Computational Linguistics, 2004.

[5] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 412–418, Barcelona, Spain, 2004.

[6] J. Wiebe, R. Bruce, M. Martin, T. Wilson, and M. Bell, "Learning subjective language," *Computational Linguistics*, vol. 30, no. 3, pp. 277–308, 2004.

[7] C. Zhang, W. Zuo, T. Peng, and F. He, "Sentiment classification for Chinese reviews using machine learning methods based on string kernel," in *Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology*, pp. 909–914, November 2008.

[8] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, pp. 519–528, May 2003.

[9] M. Chau and J. Xu, "Mining communities and their relationships in blogs: a study of online hate groups," *International Journal of Human Computer Studies*, vol. 65, no. 1, pp. 57–70, 2007.

[10] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6527–6535, 2009.

[11] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of internet restaurant reviews written in cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, 2011.

[12] L.-S. Chen and H.-J. Chiu, "Developing a neural network based index for sentiment classification," in *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, pp. 744–749, Hong Kong, China, 2009.

[13] J. Tao and T. Tan, "Emotional chinese talking head system," in *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04)*, pp. 273–280, ACM Press, New York, NY, USA, October 2004.

[14] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 168–177, ACM, August 2004.

[15] Y. Zhang, Y. Dang, and H. Chen, "Gender classification for web forums," *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 41, no. 4, pp. 668–677, 2011.

[16] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[17] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011.

[18] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 3, pp. 397–407, 2012.

[19] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 623–634, 2014.

[20] P. Kalaivani and K. L. Shunmuganathan, "Sentiment classification of movie reviews by supervised machine learning approaches," *Indian Journal of Computer Science and Engineering*, vol. 4, no. 4, pp. 285–292, 2013.

[21] P. Kalaivani and K. L. Shunmuganathan, "Performance evaluation of feature selection method for sentiment classification of online reviews using machine learning techniques," *International Review on Computers and Software*, vol. 8, no. 8, pp. 1769–1775, 2013.

[22] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for opinion analysis," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, pp. 625–631, Bremen, Germany, 2005.

[23] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums," *ACM Transactions on Information Systems*, vol. 26, no. 3, article 12, 2008.

[24] M. Whitehead and L. Yaeger, "Sentiment mining using ensemble classification models," in *Innovations and Advances in Computer Sciences and Engineering*, pp. 509–514, Springer, Dordrecht, The Netherlands, 2010.

[25] F. Salvetti, S. Lewis, and C. Reichenbach, "Automatic opinion polarity classification of movie reviews," *Colorado Research in Linguistics*, vol. 17, no. 1, 2004.

[26] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[27] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceeding of the 20th International Conference on Computational Linguistics*, 2004.

[28] P. V. Balakrishnan, R. Gupta, and V. S. Jacob, "Development of hybrid genetic algorithms for product line designs," *IEEE Transactions on Systems, Man, and Cybernetics. Part B. Cybernetics*, vol. 34, no. 1, pp. 468–483, 2004.

[29] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. A. Ureña-López, "Experiments with SVM to classify opinions in different domains," *Expert Systems with Applications*, vol. 38, no. 12, pp. 14799–14804, 2011.

[30] R. Prabowo and M. Thelwall, "Sentiment analysis: a combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009.

[31] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Proceedings of the 8th Pacific-Asia Conference (PAKDD '04)*, pp. 3–12, Sydney, Australia, May 2004.

[32] B. Bostanci and E. Bostanci, "An evaluation of classification algorithms using MC Nemar's test," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications*, vol. 201 of *Advances in Intelligent Systems and Computing*, Springer, New Delhi, India, 2013.

[33] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.