

Bagged Ensemble Classifiers for Sentiment Classification of Movie Reviews

M.Govindarajan

Assistant Professor

Department of Computer Science and Engineering

Annamalai University Annamalai Nagar

Tamil Nadu, India

e-mail: govind_aucse@yahoo.com

Abstract—The area of sentiment mining (also called sentiment extraction, opinion mining, opinion extraction, sentiment analysis, etc.) has seen a large increase in academic interest in the last few years. Researchers in the areas of natural language processing, data mining, machine learning, and others have tested a variety of methods of automating the sentiment analysis process. The feasibility and the benefits of the proposed approaches are demonstrated by means of movie review that is widely used in the field of sentiment classification. In the proposed work, a comparative study of the effectiveness of ensemble technique is made for sentiment classification. Bagging and boosting are two relatively new but popular methods for producing ensembles. In this work, bagging is evaluated on movie review data set in conjunction with Naive Bayes (NB), Support Vector Machine (SVM), Genetic Algorithm (GA) as the base learners. The proposed bagged NB, SVM, GA is superior to individual approaches for movie review in terms of classification accuracy.

Keywords- Accuracy, Bagging, Genetic Algorithm (GA), Naive Bayes (NB), Sentiment Mining, Support Vector Machine.

I. INTRODUCTION

Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Sentiment analysis is a kind of text classification that classifies text based on the sentimental orientation of opinions they contain. It is also known as opinion mining, opinion extraction and affects analysis in the literature.

Recently, many web sites have emerged that offer reviews of items like books, cars, snow tires,

vacation destinations, etc. They describe the items in some detail and evaluate them as good/bad, preferred/not preferred. So, there is motivation to categorize these reviews in an automated way by a property other than topic, namely, by what is called their 'sentiment' or 'polarity'. That is, whether they recommend or do not recommend a particular item. One speaks of a review as having positive or negative polarity.

Now, such automated categorization by sentiment, if it worked effectively, would have many applications. First, it would help users quickly to

classify and organize on-line reviews of goods and services, political commentaries, etc. Secondly, categorization by sentiment would also help businesses to handle 'form free' customer feed-back. They could use it to classify and tabulate such feedback automatically and could thereby determine, for instance, the percentage of happy clientele without having actually to read any customer input. Not only businesses but governments and non-profit organizations might benefit from such an application. Thirdly, categorization by sentiment could also be used to filter email and other messages. A mail program might use it to eliminate so-called 'flames'. Finally, perhaps a word processor might employ it to warn an author that he is using bombastic or other undesirable language. In this light, there is suitable motivation to look at the possibility of automated categorization by sentiment.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents proposed methodology and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

II. RELATED WORK

Sentiment analysis of movie reviews is considered to be very challenging since movie reviewers often present lengthy plot summaries and also use complex literary devices such as rhetoric and sarcasm.

Previously used techniques for sentiment classification can be classified into three categories.

These include machine learning algorithms, link analysis methods, and score based approaches. The effectiveness of machine learning techniques when applied to sentiment classification tasks is evaluated in the pioneering research by Pang et al, 2002.

Many studies have used machine learning algorithms with support vector machines (SVM) and Naïve Bayes (NB) being the most commonly used. SVM has been used extensively for movie reviews (Pang et al, 2002; Pang and Lee, 2004; Whitelaw et al., 2005) while Naïve Bayes has been applied to reviews and web discourse (Pang et al, 2002; Pang and Lee, 2004; Efron, 2004). In comparisons, SVM has outperformed other classifiers such as NB (Pang et al., 2002).

Genetic algorithms are search heuristics that are similar to the process of biological evolution and natural selection and survival of the fittest. Genetic Algorithms (GAs) are probabilistic search methods. GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions (S Chandrakala et al, 2012). The experiments with GA's start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation.

The ensemble technique, which combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains (T. Ho,

1994; J. Kittler., 1998). In topical text classification, several researchers have achieved improvements in classification accuracy via the ensemble technique. In the early work (L. Larkey et al, 1996), a combination of different classification algorithms (k-NN, Relevance feedback and Bayesian classifier) produces better results than any single type of classifier.

Ensemble machine learning techniques increase classification accuracy by combining arrays of specialized learners. Bootstrap aggregating also called Bagging was one of the first ensemble machine learning techniques (N. Anitha et al, 2013). Saraswathi et al, (2012) proposed Inverse Document Frequency and classify the opinions by using bagging algorithms.

In this work, bagging is evaluated on movie review in conjunction with NB, SVM, GA as the base learners. The performance of the proposed bagged (NB, SVM, GA) classifiers are examined in comparison with standalone classifiers.

III. PROPOSED METHODOLOGY

Several researchers have investigated the combination of different classifiers to form an ensemble classifier (D. Tax et al, 2000). An important advantage for combining redundant and complementary classifiers is to increase robustness, accuracy, and better overall generalization. This research work aims to make an intensive study of the effectiveness of ensemble techniques for sentiment classification tasks. In this work, first the base classifiers such as Naive Bayes (NB), Support Vector Machine (SVM), Genetic Algorithm (GA) are constructed to predict classification scores. The

reason for that choice is that they are representative classification methods and very homogeneous techniques in terms of their philosophies and strengths. All classification experiments were conducted using 10×10 -fold cross-validation for evaluating accuracy. Secondly, well known homogeneous ensemble techniques are performed with base classifiers to obtain a very good generalization performance. The feasibility and the benefits of the proposed approaches are demonstrated by means of movie review that is widely used in the field of sentiment classification. A wide range of comparative experiments are conducted and finally, some in-depth discussion is presented and conclusions are drawn about the effectiveness of ensemble technique for sentiment classification.

This research work proposes new combined methods for sentiment mining problem. A new architecture based on coupling classification methods using bagging classifier adapted to sentiment mining problem is defined in order to get better results. The main originality of the proposed approach is based on five main parts: Preprocessing phase, Document Indexing phase, feature reduction phase, classification phase and combining phase to aggregate the best classification results.

A. *Data Pre-processing*

Different pre-processing techniques were applied to remove the noise from our data set. It helped to reduce the dimension of our data set, and hence building more accurate classifier, in less time.

The main steps involved are i) document pre-processing, ii) feature extraction / selection, iii) model selection, iv) training and testing the classifier.

Data pre-processing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop-word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example, „a“, “the”, “an”, “of” etc. in English language), so that they are not useful for classification. Stemming is the action of reducing words to their root or base form. For English language, the Porter"s stemmer is a popular algorithm, which is a suffix stripping sequence of systematic steps for stemming an English word, reducing the vocabulary of the training text by approximately one-third of its original size. For example, using the Porter"s stemmer, the English word “generalizations” would subsequently be stemmed as “generalizations → generalization → generalize → general → gener”. In cases where the source documents are web pages, additional pre-processing is required to remove / modify HTML and other script tags.

Feature extraction / selection helps identify important words in a text document. This is done using methods like TF-IDF (term frequency-inverse document frequency), LSI (latent semantic indexing), multi-word etc. In the context of text classification, features or attributes usually mean

significant words, multi-words or frequently occurring phrases indicative of the text category.

After feature selection, the text document is represented as a document vector, and an appropriate machine learning algorithm is used to train the text classifier. The trained classifier is tested using a test set of text documents. If the classification accuracy of the trained classifier is found to be acceptable for the test set, then this model is used to classify new instances of text documents.

B. Document Indexing

Creating a feature vector or other representation of a document is a process that is known in the IR community as *indexing*. There are a variety of ways to represent textual data in feature vector form, however most are based on word co-occurrence patterns. In these approaches, a vocabulary of words is defined for the representations, which are all possible words that might be important to classification. This is usually done by extracting all words occurring above a certain number of times (perhaps 3 times), and defining your feature space so that each dimension corresponds to one of these words.

When representing a given textual instance (perhaps a document or a sentence), the value of each dimension (also known as an attribute) is assigned based on whether the word corresponding to that dimension occurs in the given textual instance. If the document consists of only one word, then only that corresponding dimension will have a value, and every other dimension (i.e., every other attribute) will be zero. This is known as the “bag of words”

approach. One important question is what values to use when the word is present. Perhaps the most common approach is to weight each present word using its frequency in the document and perhaps its frequency in the training corpus as a whole. The most common weighting function is the *tfidf* (term frequency-inverse document frequency) measure, but other approaches exist. In most sentiment classification work, a binary weighting function is used. Assigning 1 if the word is present, 0 otherwise, has been shown to be most effective.

C. Dimensionality Reduction

Dimension Reduction techniques are proposed as a data pre-processing step. This process identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis.

Steps:

- ✓ Select the dataset.
- ✓ Perform discretization for pre-processing the data.
- ✓ Apply Best First Search algorithm to filter out redundant & super flows attributes.
- ✓ Using the redundant attributes apply classification algorithm and compare their performance.
- ✓ Identify the Best One.

1) Best first Search

Best First Search (BFS) uses classifier evaluation model to estimate the merits of attributes. The attributes with high merit value is considered as potential attributes and used for classification Searches the space of attribute subsets by augmenting with a backtracking facility. Best first

may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions.

D. Existing Classification Methods

1) Naive Bayes (NB)

The Naïve Bayes assumption of attribute independence works well for text categorization at the word feature level. When the number of attributes is large, the independence assumption allows for the parameters of each attribute to be learned separately, greatly simplifying the learning process.

There are two different event models. The multi-variate model uses a document event model, with the binary occurrence of words being attributes of the event. Here the model fails to account for multiple occurrences of words within the same document, which is a more simple model. However, if multiple word occurrences are meaningful, then a multinomial model should be used instead, where a multinomial distribution accounts for multiple word occurrences. Here, the words become the events.

2) Support Vector Machine (SVM)

The support vector machine (SVM) is a recently developed technique for multi dimensional function approximation. The objective of support vector machines is to determine a classifier or regression function which minimizes the empirical risk (that is the training set error) and the confidence interval (which corresponds to the generalization or test set error).

Given a set of N linearly separable training examples $\mathbf{S} = \{\mathbf{x}_i \in \mathbf{R}^n | i = 1, 2, \dots, N\}$, where each example belongs to one of the two classes, represented by $y_i \in \{+1, -1\}$, the SVM learning method seeks the optimal hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, as the decision surface, which separates the positive and negative examples with the largest margins. The decision function for classifying linearly separable data is:

$$f(\mathbf{X}) = \text{sign}(W \cdot X + b)$$

Where \mathbf{w} and b are found from the training set by solving a constrained quadratic optimization problem. The final decision function is

$$\mathbf{f}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \mathbf{a}_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + \mathbf{b}\right)$$

(2)

The function depends on the training examples for which a_i is non-zero. These examples are called support vectors. Often the number of support vectors is only a small fraction of the original data set. The basic SVM formulation can be extended to the non linear case by using the nonlinear kernels that maps the input space to a high dimensional feature space. In this high dimensional feature space, linear classification can be performed. The SVM classifier has become very popular due to its high performances in practical applications such as text classification and pattern recognition.

The support vector regression differs from SVM used in classification problem by introducing an alternative loss function that is modified to include a distance measure. Moreover, the parameters that control the regression quality are the cost of error C , the width of tube ϵ and the mapping function ϕ .

In this research work, the values for polynomial degree will be in the range of 0 to 5. In this work, best kernel to make the prediction is polynomial kernel with $\epsilon = 1.0E-12$, parameter $d=4$ and parameter $c=1.0$.

3) Genetic Algorithm (GA)

The genetic algorithm is a model of machine learning which derives its behaviour from a metaphor of some of the mechanisms of evolution in nature. This done by the creation within a machine of a population of individuals represented by chromosomes, in essence a set of character strings.

The individuals represent candidate solutions to the optimization problem being solved. In genetic algorithms, the individuals are typically represented by n -bit binary vectors. The resulting search space corresponds to an n -dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function.

Genetic algorithms use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation. Mutation and crossover are two of the most commonly used operators that are used with genetic algorithms that represent individuals as binary strings. Mutation operates on a single string and generally changes a bit at random while

crossover operates on two parent strings to produce two offsprings. Other genetic representations require the use of appropriate genetic operators.

The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found. In practice, the performance of genetic algorithm depends on a number of factors including: the choice of genetic representation and operators, the fitness function, the details of the fitness-dependent selection procedure, and the various user-determined parameters such as population size, probability of application of different genetic operators, etc. The basic operation of the genetic algorithm is outlined as follows:

Procedure:

begin

$t \leftarrow 0$

initialize $P(t)$

while (not termination condition)

$t \leftarrow t + 1$

select $P(t)$ from $p(t - 1)$

crossover $P(t)$

mutate $P(t)$

evaluate $P(t)$

end

end.

Our contribution relies on the association of all the techniques used in our method. First the small selection in grammatical categories and the use of³ bi-grams enhance the information contained in the⁴ vector representation, then the space reduction

allows getting more efficient and accurate computations, and then the voting system enhance the results of each classifier. The overall process comes to be very competitive.

E) Proposed Bagged Ensemble Classifiers

Given a set D , of d tuples, bagging (Breiman, L. 1996a) works as follows. For iteration i ($i = 1, 2, \dots, k$), a training set, D_i , of d tuples is sampled with replacement from the original set of tuples, D . The bootstrap sample D_i , by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i . A classifier model, M_i , is learned for each training set, D_i . To classify an unknown tuple, X , each classifier, M_i , returns its class prediction, which counts as one vote. The bagged (NB, SVM, GA), M^* , counts the votes and assigns the class with the most votes to X .

Algorithm: Bagged ensemble classifiers using bagging

Input:

- D , a set of d tuples.
- $k = 1$, the number of models in the ensemble.
- Base Classifier (NB, SVM, GA)

Output: A Bagged (NB, SVM, GA), M^*

Method:

1. for $i = 1$ to k do // create k models
2. Create a bootstrap sample, D_i , by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i
3. Use D_i to derive a model, M_i ;
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i , based

on the accuracies of percentage of correctly classified example in training data D_i .

5. endfor

To use the bagged ensemble models on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

IV. PERFORMANCE EVALUATION MEASURES

A. Cross Validation Technique

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation the folds are selected so that the mean response value is approximately equal in all the folds.

B. Criteria for Evaluation

The primary metric for evaluating classifier performance is classification Accuracy - the percentage of test samples that are correctly classified. The accuracy of a classifier refers to the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

V. EXPERIMENTAL RESULTS

A. Dataset Description

The basic data set consist of 2000 movie reviews, 1000 labelled positive and 1000 labelled negative (so they have a uniform class distribution). These were downloaded from Bo Pang's web page: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>.

B. Results and Discussion

TABLE 1: THE PERFORMANCE OF BASE AND PROPOSED BAGGED NB CLASSIFIER FOR MOVIE REVIEW DATA

Dataset	Classifiers	Accuracy
Movie-Review Data	Existing NB Classifier	91.15 %
	Proposed Bagged NB Classifier	92.55 %

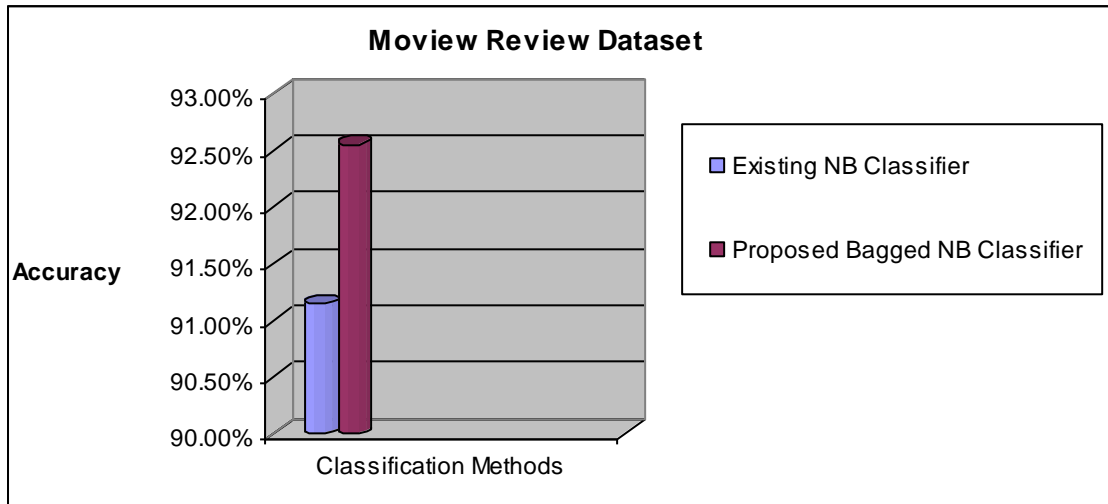


Figure 1: Classification Accuracy of Existing and Proposed Bagged NB Classifier using Movie Review Data

TABLE 2: THE PERFORMANCE OF BASE AND PROPOSED BAGGED SVM CLASSIFIER FOR MOVIE REVIEW DATA

Dataset	Classifiers	Accuracy
Movie-Review Data	Existing SVM Classifier	91.35 %
	Proposed Bagged SVM Classifier	93.60 %

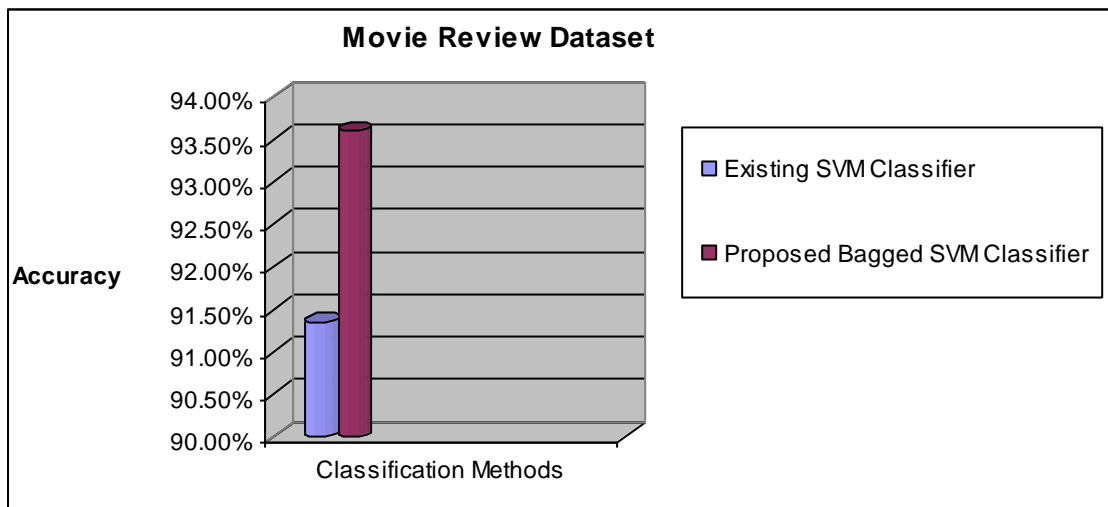


Figure 2: Classification Accuracy of Existing and Proposed Bagged SVM Classifier using Movie Review Data

TABLE 3: THE PERFORMANCE OF BASE AND PROPOSED BAGGED GA CLASSIFIER FOR MOVIE REVIEW DATA

Dataset	Classifiers	Accuracy
Movie-Review Data	Existing GA Classifier	91.25 %
	Proposed Bagged GA Classifier	92.40 %

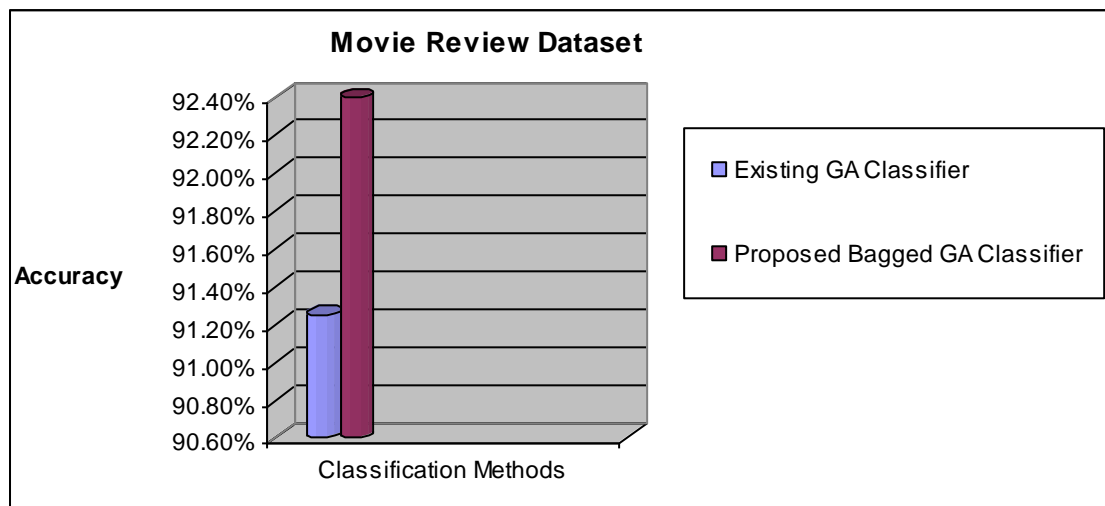


Figure 3: Classification Accuracy of Existing and Proposed Bagged GA Classifier using Movie Review Data

In this research work, new ensemble classification method is proposed using bagging classifier in conjunction with NB, SVM, GA as the base learner and the performance is analyzed in terms of accuracy. Here, the base classifiers are constructed using NB, SVM, GA. 10-fold cross validation (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated classification accuracy. Bagging is performed with NB, SVM, GA to obtain a very good classification performance. Table 1 to 3 shows classification performance for movie review using existing and proposed bagged NB, SVM, GA. The analysis of results shows that the proposed bagged NB, SVM, GA are shown to be superior to individual approaches for movie review in terms of classification accuracy. According to Fig. 1 to 3 proposed combined models show significantly larger improvement of classification accuracy than the base classifiers. This means that the combined methods are more accurate than the individual methods for the movie reviews.

VI. CONCLUSIONS

In this research work, new combined classification methods are proposed using bagging classifier in

conjunction with NB, SVM, GA as the base learners and the performance comparison has been demonstrated using movie reviews in terms of accuracy. This research has clearly shown the importance of using ensemble approach for movie reviews. An ensemble helps to indirectly combine the synergistic and complementary features of the different learning paradigms without any complex hybridization. Since all the considered performance measures could be optimized, such systems could be helpful in several real world sentiment mining applications. The high classification accuracy has been achieved for the ensemble classifiers compared to that of single classifiers. The proposed bagged NB, SVM, GA are shown to be significantly higher improvement of classification accuracy than the base classifiers. The movie reviews could be detected with high accuracy for homogeneous model. The future research will be directed towards developing more accurate base classifiers particularly for the sentiment mining applications.

ACKNOWLEDGMENT

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and

encouragement to carry out this work. This work is supported by DST-SERB Fast track Scheme for Young Scientists by the Department of science and technology, Government of India, New Delhi.

REFERENCES

- [1] N. Anitha, B. Anitha, S. Pradeepa, (2013), Sentiment Classification Approaches – A Review, International Journal of Innovations in Engineering and Technology (IJET), 3 (1), pp. 22-31.
- [2] Breiman, L. (1996a). Bagging predictors. Machine Learning, 24(2), pp.123–140.
- [3] S Chandrakala and C Sindhu, (2012), Opinion Mining and sentiment classification a survey, ICTACT journal on soft computing.
- [4] Efron, M. (2004), Cultural orientations: Classifying subjective documents by cocitation analysis, In Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design, pp. 41-48.
- [5] T. Ho, J. Hull, S. Srihari, (1994), Decision combination in multiple classifier systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, pp. 66–75.
- [6] J. Kittler, (1998), Combining classifiers: a theoretical framework, *Pattern Analysis and Applications*, 1, pp.18–27.
- [7] Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of International Joint Conference on Artificial Intelligence, pp.1137–1143.
- [8] L. Larkey, W. Croft, (1996), Combining classifiers in text categorization, in: Proceeding of ACM SIGIR Conference, ACM, New York, NY, USA, pp. 289–297.
- [9] B. Pang, L. Lee, S. Vaithyanathan, (2002), Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.
- [10] Pang, B., and Lee, L. (2004), A sentimental education: Sentimental analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 271-278.
- [11] Saraswathi.K ,Tamilarsi.A, (2012), A Modified Metaheuristic Algorithm for Opinion mining, International Journal of Computer Applications, Volume 58.
- [12] D. Tax, M. Breukelen, R. Duin, and J. Kittler, (2000), Combining multiple classifiers by averaging or by mutliplying?, *Pattern Recognition*, Vol 33, pp. 1475-1485.
- [13] Whitelaw, C., Garg, N., and Argamon, S. (2005), Using appraisal groups for sentiment analysis, Proceedings of the 14th ACM Conference on Information and Knowledge Management, pp. 625-631.