SENTIMENT ANALYSIS OF RESTAURANT REVIEWS USING HYBRID CLASSIFICATION METHOD

M. GOVINDARAJAN

Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India
Email: 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. In this research work, new hybrid classification method is proposed based on coupling classification methods using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble was designed using Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA). In the proposed work, a comparative study of the effectiveness of ensemble technique is made for sentiment classification. The feasibility and the benefits of the proposed approaches are demonstrated by means of restaurant 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.

Keywords— Accuracy, Arcing classifier, Genetic Algorithm (GA). Naïve Bayes (NB), Sentiment Mining, Support Vector Machine (SVM)

I. INTRODUCTION

Yelp users give ratings and write reviews about businesses and services on Yelp. These reviews and rating help other yelp users to evaluate a business or a service and make a choice. The problem most users face nowadays is the lack of time; most people are unable to read the reviews and just rely on the business' ratings. This can be misleading. While ratings are useful to convey the overall experience, they do not convey the context that led users to that experience. For example, in case of a restaurant, the food, the ambience, the service or even the discounts offered can often influence the user ratings. This information is not conceivable from rating alone, however, it is present in the reviews that users write.

The classification of yelp restaurant reviews into one or more, "Food", "Service", "Ambience", "Deals/Discounts", and "Worthiness", categories is the problem in consideration. Inputs are the Yelp restaurant reviews and review ratings. The multi-label classifier outputs the list of relevant categories that apply to the given Yelp review. Consider a Yelp review: "They have not the best happy hours, but the food is good, and service is even better. When it is winter we become regulars". It is easily inferred that this review talks about "food" and "service" in a positive sentiment, and "deals/discounts" (happy hours) in a negative sentiment. Extracting classification information from the review and presenting it to the user, shall help the user understand why a reviewer rated the restaurant "high" or "low" and make a more informed decision, avoiding the time consuming process of reading the entire list of restaurant reviews.

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

There are large number of papers on related topics, for example, recommendation systems (Adomavicius, G and et al., 2005), informative peer-prediction method (Nolan Miller and et al., 2005), and rating prediction.

Adomavicius, G and et al., (2005) presents us an overview of recommend systems. Besides, it describes the current version of recommendation methods that are mainly divided into three categories, content-based, collaborative, and hybrid recommendation approaches. However, there are limitations on these approaches. This paper discusses several possible extensions that can improve recommendation capabilities, as well as make recommendation systems applicable to an broader range of application.

Michael J and et al., (2007) presents us a basic content-based recommendation system; it recommends an item based on the description of this item, as well as the profile of the user's interest. These two factors together determine the final

recommendation. Although the details of an item may differ in different recommendation systems, there are things staying in common. For example, the means to compare item features.

Gayatree Ganu and et al., (2009) gave us a more similar example. A free-text format review is difficult for computers to analyze, understand and aggregate. To identify the information in the text reviews, this paper presents new ad-hoc and regression-based recommendation methods that takes into consideration the textual component of user reviews.

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.

Ziqiong Zhang and et al., (2011) used standard machine learning techniques naive Bayes and SVM are incorporated into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative. The effects of feature presentations and feature sizes on classification performance are discussed.

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 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.

Freund and Schapire (1995,1996) proposed an algorithm the basis of which is to adaptively resample and **c**ombine (hence the acronym--arcing) so that the weights in the resampling are increased for those cases most often misclassified and the combining is done by weighted voting.

In this research work, proposes a new hybrid method for sentiment mining problem. A new architecture based on coupling classification methods (NB, SVM and GA) using arcing classifier adapted to sentiment mining problem is defined in order to get better results.

III. PROPOSED METHODOLOGY MATH

Several researchers have investigated the combination of different classifiers to from an ensemble classifier. 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 Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA) are constructed to predict classification scores. The reason for that choice is that they are representative classification methods and very heterogeneous 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 heterogeneous ensemble technique is 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 restaurant 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 hybrid method for sentiment mining problem. A new architecture based on coupling classification methods using arcing 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 out 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

Three classification methods are adapted for each training set. The most competitive classification methods are used for a given corpus. The results are evaluated using the cross validation method on restaurant review based on the classification accuracy.

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 $S = \left\{ x_i \in R^n | i = 1,2,...,N \right\}$, where each example belongs to one of the two classes, represented by $y_i \in \{\div 1, -1\}$, the SVM learning method seeks the optimal hyperplane **w . 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:

$$f(\mathrm{X}) = sign(W.X + b) \qquad (1)$$

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

$$f(x) = sign\left(\sum_{i=1}^{N} a_i y_i (x_i..x) + b \right) \qquad (2)$$

The function depends on the training examples for which a_i s 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 (A. Abbasi, et al., 2008) 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 <- 0
initialize P(t)
while (not termination condition)
t <- 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 of bi-grams enhance the information contained in the 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 Hybrid Method

Given a set D, of d tuples, arcing (Breiman. L, 1996) works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset *D* will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples *d* from training dataset 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 hybrid classifier (NB, SVM and GA), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid Method using Arcing Classifier
Input:

- D, a set of d tuples.
- k = 3, the number of models in the ensemble.
- Base Classifiers (NB, SVM and GA)

Output: Hybrid Method, M^*.

Procedure:

1. For i = 1 to k do // Create k models
2. Create a new training dataset, D_i, by sampling D with replacement. Same example from given dataset *D* may occur more than once in the training dataset 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 hybrid model 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;

The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

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

This research is performed with the data from the Yelp Dataset Challenge [18]. This dataset includes business, review, user, and checkin data in the form of separate JSON objects. A business object includes information about the type of business, location, rating, categories, and business name, as well as contains a unique id. A review object has a rating, review text, and is associated with a specific business id and user id.

B. Results and Discussion

Table 1: The Performance of Base and Hybrid Classifier for Restaurant Review Data

Dataset	Classifiers	Accuracy
Restaurant Review Data	Naïve Bayes	85.00 %
	Support Vector Machi	85.20 %
	Genetic Algorithm	85.30 %
	Proposed Hybrid Meth	92.44 %

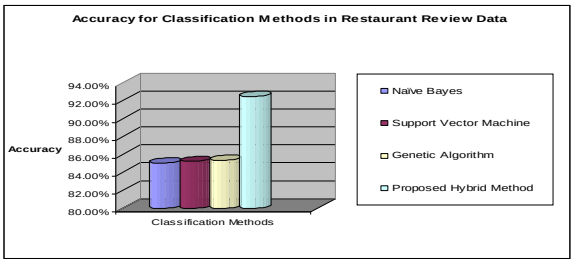

Figure 1: Classification Accuracy of Restaurant Review

The data set described in section 2 is being used to test the performance of base classifiers and hybrid classifier. Classification accuracy was evaluated using 10-fold cross validation. In the proposed approach, first the base classifiers Naïve Bayes, SVM and GA are constructed individually to obtain a very good generalization performance. Secondly, the ensemble of Naïve Bayes, SVM and GA is designed. In the ensemble approach, the final output is decided as follows: base classifier's output is given a weight (0–1 scale) depending on the generalization performance as given in Table 1. According to Table 1, the proposed hybrid model shows significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant. The proposed ensemble of Naïve Bayes, SVM and GA are shown to be superior to individual approaches for Restaurant review data in terms of Classification accuracy.

CONCLUSION

In this research, a new hybrid technique is investigated for Restaurant reviews and evaluated their performance based on the Restaurant review data and then classifying the reduced data by NB, SVM and GA. Next a hybrid model and NB, SVM, GA models as base classifiers are designed. Finally, a hybrid system is proposed to make optimum use of the best performances delivered by the individual base classifiers and the hybrid approach. The hybrid model shows higher percentage of classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction.

The experiment results lead to the following observations.

- GA exhibits better performance than SVM and NB in the important respects of accuracy.
- Comparison between the individual classifier and the hybrid classifier: it is clear that the hybrid classifier show the significant improvement over the single classifiers.

ACKNOWLEDGEMENT

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] A. Abbasi, H. Chen and A. Salem, (2008), "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums", ACM Transactions on Information Systems, Vol. 26, No. 3, PP.

[2] Adomavicius, G. ,Tuzhilin, A. (2005), "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 6.

[3] L. Brieman. (1996), "Bias, Variance, and Arcing Classifiers", Technical Report 460, Department of Statistics, University of California at Berkeley, CA 94720.

[4] S Chandrakala and C Sindhu, (2012), "Opinion Mining and sentiment classification a survey", ICTACT journal on soft computing, volume: 03, issue: 01, pp. 420-427.

[5] Freund, Y. and Schapire, R. (1995), "A decision-theoretic generalization of on-line learning and an application to boosting", In proceedings of the Second European Conference on Computational Learning Theory, pp. 23-37.

[6] Freund, Y. and Schapire, R. (1996), "Experiments with a new boosting algorithm", In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, pp.148-156.

[7] Gayatree Ganu, No´emie Elhadad, Am´elie Marian, (2009), "Beyond the Stars: Improving Rating Predictions using Review Text Content", Twelfth International Workshop on the Web and Databases, Providence, Rhode Island, USA

[8] 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.

[9] Kim S., Han K., Rim H., and Myaeng S. H. (2006), "Some effective techniques for naïve bayes text classification", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 11, pp. 1457-1466.

[10] J. Kittler, (1998), "Combining classifiers: a theoretical framework", Pattern Analysis and Applications, 1, pp.18–27.

[11] L. Larkey, W. Croft, (1996), "Combining classifiers in text categorization", in: Proceeding of ACM SIGIR Conference, ACM, New York, NY, USA, pp. 289–297.

[12] S. Li, C. Zong, X. Wang, (2007), "Sentiment classification through combining classifiers with multiple feature sets", Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 07), pp. 135–140.

[13] Michael J. Pazzani and Daniel Billsus, (2007), "Content-Based Recommendation Systems", Lecture Notes in Computer Science Volume 4321, pp 325-341

[14] Mullen, Tony and Nigel Collier. (2004), "Sentiment analysis using Support Vector Machines with diverse information sources", In Dekang Lin and Dekai Wu, editors, Proceedings of EMNLP, PP. 412-418, Barcelona, Spain. Association for Computational Linguistics.

[15] Nolan Miller, Paul Resnick, Richard Zeckhauser, (2005), "Eliciting Informative Feedback: The Peer-Prediction Method" management science Vol. 51, No. 9, pp. 1359–1373

[16]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.

[17]Rui Xia, Chengqing Zong, Shoushan Li , (2011), "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences, 181, pp. 1138–1152.

[18]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.

[19]http://www.yelp.com/dataset_challenge/

[20]Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li (2011), **"**Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications: An International Journal, Volume 38 Issue 6, Pages 7674-7682.

★★★