Sentiment Analysis of Movie Reviews Using Hybrid Classification Models

Sung Min Lee[1], Hye Jin Kim[2], Jiwon Park[3]

[1]Department of Electrical Engineering, Pohang University of Science and Technology, South Korea
[2]Department of Civil Engineering, Korea Advanced Institute of Science and Technology, South Korea
[3]Department of Chemical Engineering, Seoul National University, South Korea

ABSTRACT

Internet has provided people a platform to express their opinions and thoughts. Sentiment analysis helps to analyse those opinions and categorize them. This research is done on the movie review dataset obtained from the Internet Movie Database (IMDb). The data is classified using some of the popular learning based classifiers like Naive Bayes, Decision Tree and Support Vector Machine (SVM) classifiers and their accuracies are compared. Finally, the three learning based classifiers are combined using the Majority vote ensemble classifier. It is found that the accuracy obtained from the above said ensemble is better than the individual classifiers and also better than the ensemble which uses the random forest as one of the classifiers.

KEYWORDS: Sentiment Analysis, Movie Review, Ensemble classifier, Majority Vote Classifier.

I. INTRODUCTION

The current digital age has transformed the internet as a huge database. Users of the Internet express their opinions across the various platforms such as Twitter, Facebook, IMDb etc. These opinions are very helpful in finding the overall taste of the social community. Lots of movies get released every day. The decision to watch a movie depends on its reviews. Movie reviews are largely of positive or negative polarity and contain a lot of sentiment based words. These words can be put to effective use in Sentiment Analysis. Sentiment analysis is "The process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention" [1]. By subjecting the movie review data to Sentiment Analysis, one can find whether the movie has more positive or negative reviews. Tajinder singh et al [2], used the SVM classifier for analysing the sentiment of twitter posts and observed the improvement in accuracy with proper pre-processing of text data. Isha Gandhi et al [3], proposed a hybrid ensemble classifier that combines the representative algorithms of Instance based learner, Naive Bayes and Decision Tree Algorithms using voting methodology and concluded that the ensemble provides better accuracy. Bin Lu et al [4], combined a sentiment lexicon and the SVM classifier for opinion analysis and observed that the lexicon classification when combined with SVM outperformed the individual machine learning techniques. Zamahsyari et al [5], implemented a majority vote ensemble that combined Decision Tree, Naive Bayes and Random Forest to analyse the economic news of Bahama. In the present work IMDb's large movie review dataset [6] has been analysed using some popular learning based classifiers like Decision tree, Naive Bayes and SVM. Finally the learning based classifiers are combined into an ensemble using a majority vote classifier. Ensemble is an effective technique which combines various learning algorithms so as to improve the overall prediction accuracy. The accuracy obtained from each classifier is compared. From this work, it was found that the ensemble classifier gives better results than individual classifiers and also better results than the ensemble which uses the random forest as one of the classifiers [5].

II. METHODOLOGY

Data collection

The Large Movie Review dataset from ImDb contains 50,000 reviews which are split evenly into 25000 training and 25000 test datasets. There are 25000 positive and 25000 negative reviews. There is also an additional 50,000 data that is unlabelled. In the present work, only the labelled data are considered for analysis. For any movie, there are no more than 30 reviews because that could affect the overall result of the classification. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels [7]. There are two directories named train and test corresponding to the training and test datasets. Each directory further has two more directories named "pos" and "neg" that contain the respective reviews. These reviews are stored in text files.

Data Pre-processing

The reviews often contain words and punctuations unnecessary for classification. These elements affect the accuracy of classification. Data pre-processing allows the removal of these unwanted elements. In this process, the following steps were

1. Tokenizing
2. Case Removal
3. Punctuation Removal
4. Stop words Removal
5. Stemming

The series of steps performed in pre-processing are demonstrated in Table 1.

Table 1. Steps performed in pre-processing the data

Step	Before	After
Case Removal	Once again Mr. Costner has dragged out a movie for far longer than necessary!!	once again mr. costner has dragged out a movie for far longer than necessary!!
Punctuation Removal	once again mr**.** costner has dragged out a movie for far longer than necessary**!!**	once again mr costner has dragged out a movie for far longer than necessary
Stop words Removal	once again mr costner has dragged out a movie for far longer than necessary	mr costner dragged movie far longer necessary
Stemming	mr costner dragged movie far longer necessary	mr costner drag movie far longer necessary

However, the text data cannot be given as direct inputs to the classification algorithms. Hence, the data is coverted into TF-IDF matrix representation. TF-IDF stands for :term frequency-inverse document frequency. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [8].

The Term Frequency is computed as follows

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document) (1)

The Inverse Document Frequency is computed as follows

IDF(t) = log_e(Total number of documents / Number of documents with term t in it) (2)

Classification

Naive Bayes Classifier

The Naive Bayes classification is a probability based classifier based on the Bayesian theorem. It predicts the membership probability of a given record or data. Bayes' theorem is stated mathematically as,

$$P(A|B) = \{P(B|A)\ P(A)\} / \{P(B)\} \quad (3)$$

where,

P(A) is the probability of class.

P(B) is the probability of predictor.

P(A|B) is the posterior probability of B (Probability of class given predictor)

P(B|A) is the likelihood that is the probability of predictor given class.

The multinomial Naive Bayes has been implemented in this work. This variation estimates the conditional probability of a particular word/term/token given a class, as the relative frequency of term t in documents belonging to a class [9].

Decision Tree Classifier

The Decision Tree classifier creates a tree model from the training data which is then used to predict the class of target variables. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. In this paper, the Gini Index is used as the attribute selection measure. It helps to determine the root node at each level. Gini Index is widely used in Classification and Regression Trees (CART).

Support Vector Machine Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane [10]. It is one of the most effective classification algorithms. Linear SVM is employed here.

Ensemble

The learning based classifiers Naive Bayes, Decision Tree and SVM are combined using an ensemble vote classifier. The Ensemble Vote Classifier is a meta-classifier for combining similar or conceptually different machine learning classifiers for classification via majority voting [11]. There are two types of voting namely hard and soft. In hard voting, the frequently predicted class is considered as the final class. In soft voting, an average of class probabilities is used for class prediction. In this ensemble, the soft voting method has been used.

Experimental setup

Out of the 50000 movie reviews, 25000 are used in training the classifiers. The Naïve Bayes, Decision tree and the SVM classifiers are implemented individually and their accuracies are calculated using the 10 fold cross validation method. Finally, the classifiers are combined using the majority voting ensemble and the accuracy is obtained through 10 fold cross validation method.

III. RESULTS AND DISCUSSION

Prediction Accuracy of the classifiers is considered as the evaluation parameter. It was noted that the accuracy of prediction improved after the pre-processing steps were applied. The accuracy is calculated using 10 fold cross validation. The classifiers along with the accuracies obtained after the pre-processing are listed in Table 2.

Table 2. Comparison of Accuracies

Classifier	**Accuracy(%)**
Decision Tree (Gini Index)	68.44
Naive Bayes	86.53
Support Vector Machine	89.33
Majority Vote Classifier	95.00

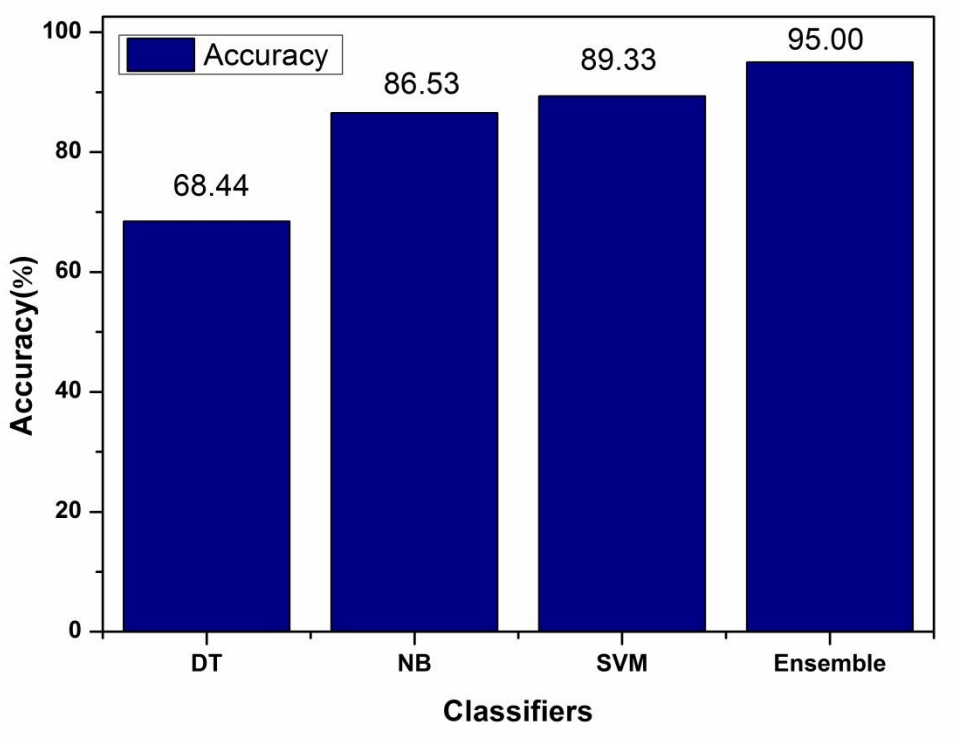

Figure 1. Comparison of Accuracies

From Figure 1 and Table 2, it is observed that the Majority Vote Classifier has the best accuracy. Zamahsyari and Arif Nurwidyantoro [5] combined Decision tree, SVM and Random forest classifiers using the voting classifier. The accuracy of the voting classifier largely depends on the individual classifiers that are combined together. It is observed that rather than the Random forest classifier, the Naive bayes classifier gives better accuracy in the ensemble when combined with decision tree and SVM classifiers.

IV. CONCLUSION

Sentiment Analysis is an important field according to business perspectives. In this work, the movie review data from IMDb was subjected to analysis. Prediction of movie review sentiment helps the concerned people to measure the success of a movie. The machine learning classification methods are known for their efficiency in classification. However, a selective combination of classifiers (Naive Bayes, Decision Tree and SVM) was observed to provide better results than their individual implementation.

V. REFERENCES

[1] https://www.brandwatch.com/blog/understanding-sentiment-analysis/
[2] Tajinder singh, Madhu Kumari, “Role of Text Pre-Processing in Twitter Sentiment Analysis”, Procedia Computer Science 89 (2016), pp.549-554.
[3] Isha Gandhi, Mrinal Pandey, “Hybrid Ensemble of Classifiers using Voting”, Green Computing and Internet of Things (ICGCIoT), 2015, DOI: 10.1109/ICGCIoT.2015.7380496.
[4] Bin Lu, K.T. Benjamin,” Combining A Large Sentiment Lexicon And Machine Learning For Subjectivity Classification”, Machine Learning and Cybernetics (ICMLC), 2010,DOI: 10.1109/ICMLC.2010.5580672.
[5] Zamahsyari, Arif Nurwidyantoro, “Sentiment Analysis of Economic News in Bahasa Indonesia Using Majority Vote Classifier”, Data and Software Engineering (ICoDSE), 2016, DOI:10.1109/ICODSE.2016.7936123.
[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Pos, “Learning Word Vectors for Sentiment Analysis”,2011 In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015
[7] https://github.com/jalbertbowden/large-movie-reviews-dataset/tree/master/acl-imdb-v1
[8] http://www.tfidf.com/
[9] http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/
[10] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
[11] https://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/