International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 02 - Issue 08 || August 2017 || PP. 16-22

Opinion Mining on Textual Reviews Using Feature Reduction Method

N. Sudha
Research Scholar
Department of Computer Science and Engineering
Annamalai University

Dr.M.Govindarajan
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

Abstract: Sentimental analysis is the task of finding sentiment or extracting opinion such as positive or negative from online text. Many datasets have a large number of features obtained the accuracy results might not good. Therefore reducing the features are important for efficient working of Machine Learning algorithms. Since it removes noise and irrelevant features from the feature vector. In this proposed work, we are using feature reduction technique such as gain ratio as an attribute evaluator along with ranker search to rank all the features over the datasets like movie, spam and restaurant review. The classification is performed using Naïve Bayes and SupportVector Machine then compared our proposed method with individual classifiers. Result shows that feature reduction method improved more accuracy of sentiment classification.

General Terms: Data mining, Classification, Data sets, Social network feature reduction and Algorithms
Keywords: Sentiment analysis, opinion extraction, reviews, Support vector machine, Naïve bayes

1. Introduction

Due to increasing number of social network contacts, online discussion forums, People are involving largely in internet to share their opinion, feelings, feedback about various subjects and events. Opinion mining refers to the use of natural language processing for tracking the opinion or attitude of how people feel about products and services. Sentiment analysis is also called as opinion mining [1]. It is a task that finds the opinion either a positive or negative from the review document like product reviews or movie reviews. These reviews not only used for individuals in taking informed decisions but organizations too for identifying customer attitudes or opinions about products and services [2]. Sentiment classification can be represented as a text classification problem. Bag Of Words (BOW) is represented commonly used for sentiment classification using machine learning algorithms.

The words exist in all documents generate feature vector. Commonly, this feature vector is a huge in dimensionality used by machine learning methods.

The Bag of Words is represented commonly for sentiment classification, results very high in dimensionality of the feature space. Because of larger growths in data storage and data accomplishment, data pre-processing techniques such as feature selection have become more popular in classification tasks. Machine learning algorithm can hold high-dimensional feature space by means of attribute selection methods in which remove the noise and irrelevant features [3]. Feature selection is a kind of dimensionality reductionand to decrease the dimensionality, eliminate irrelevant and redundant data for that the feature selection is used [4]. Various number of feature reduction techniques are Information Gain and Gain Ratio , Chi-square, among these gain ratio and information gain are the most popular methods since it provides better results in terms of accuracy and are consistent compared to other feature reduction technique [5]. This research is mainly focus on feature reduction technique with classifier used Naïve bayes and Support vector machine for sentiment classification.

The rest of the paper is organized as follows. Section 2 gives related work on sentiment analysis. The existing methods are described in section 3. Section 4 shows the datasets and the experimental results are discussed. Section 5 concludes the work done and future work.

2. Literature Review

Sentiment classification can be approximately divided into lexical based approach and machine learning approach [6]. A number of approaches such as Naive Bayes (NB), Maximum Entropy (ME), and

Support Vector Machines (SVM) are used to classify reviews. Some of the features can be used for sentiment classification are term presence, frequency, negation, n-grams and parts of speech [7]. The author presents a survey on the sentiment analysis challenges related to their approaches and techniques [8]. The supervised machine learning approaches like Naïve Bayes, Maximum Entropy and Support Vector Machines are discussed. The overall performance of these classifier models produces accuracy more than 80%. The SVM based classifier model perform better accuracy than the maximum entropy and Naïve Bayes [9]. While another study [10] proposed the unsupervised approach in which used aspect level fine-grained sentiment analysis using shallow semantic role labeling (SRL) through linguistic rules can extract aspect modifier sentiment word triples from the review sentences.

In the supervised learning approach, a classifier is first trained based on a large feature set, which consists of labeled data. Then, this classifier is used to identify and classify unlabelled test data into two classes of positive and negative sentiments. Some researchers used several feature sets to attempt to improve the classification accuracy [11][12]. Kang et al. (2011) enhanced Naïve Bayes to use on restaurant reviews and yield 83.6% accuracy on more than 6000 documents [13]. Study on five feature selection methods (Document Frequency, Information Gain, Gain Ratio, Chi Squared, Relief-F) and three popular sentiment feature lexicons (HM, GI and Opinion Lexicon) are investigated on movie reviews corpus with a size of 2000 documents. The results show that Information Gain gave consistent results and Gain Ratio performs overall best for sentimental feature selection while sentiment lexicons gave poor performance [14].

Experiments are performed the behavior of two classifiers, Naive Bayes and SVM, is investigated in combination with using various feature selection schemes and the results SVM obtained 84.75 % accuracy than NB for unigram approach[15]. Based on Fisher's discriminant ratio, an effective feature selection method is proposed compared it with Information Gain method while Support Vector Machine used as a classifier and the results shows that the Fisher's discriminant ratio based on word frequency estimation has the best performance with accuracy 86.61% and 82.80%[16]. Using feature selection techniques such as Mutual information, Chi-Square, Information gain and TF-idf used for selecting features from high dimensionality of feature set.. The classifier of support vector machine provided and also investigate that which feature is best to extract sentiments from the reviews. We are considering unigram, bigram, POS tags of words and function words as a feature set[17]. Evaluating the feature reduction method PCA, with two classifiers are Support Vector Machine and Naive Bayes while using less number of features classification accuracy has been increased on product reviews[18].

3. Methodology

This section focuses on feature reduction method for improving the accuracy and performance of sentiment classification. This figure shows the proposed methodology of sentiment analysis.

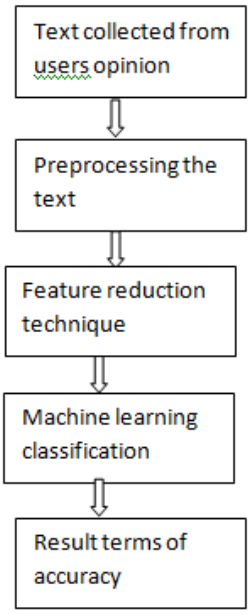

Figure 1: Proposed methodology

First, the user reviews are collected and pre-processed on the basis of natural language processing techniques includes word tokenization, stop word removal and stemming. The remaining tokens were arranged as per their frequencies in the documents set. Next apply feature selection methods are used to choose top n-ranked distinctive attributes for training the classifier. Gain ratio is one of the feature reduction method used for improving the accuracy. Naïve Bayes and Support Vector Machine are applied to evaluate the performance of sentiment classification.

3.1 Feature Reduction

The task of feature reduction is reduction dimension in feature space while verifying the minimum of accuracy. It causes the removal of irrelevant features and the outcomes are in more efficient categories such as easy analysis of sentiment after reduction, visualization of outcomes and better perception of low dimension. There are different popular methods of reduction, i.e. Document Frequency (DF) and Term Frequency-Inverse Document Frequency (TF-IDF) and Standard Deviation (SD). All of these Methods use score in terms of extraction and chosen the size of the predefined set of characters. In this work, the feature set dimension is reduced using Gain Ratio can be used to evaluating the attribute and ranker search for ranking all the attribute on the dataset and classification algorithm is used to classify the opinion.

3.1.1 Gain Ratio

GR is used for attributed evaluator when gain ratio chooses then default the ranker search method gets selected. The aim of Gain Ratio is to maximize the information Gain as it provides a normalized score of a feature's contribution to an optimal information gain based classification decision. Gain Ratio is used as an repetitive process where we choose smaller sets of variables in incremental fashion. These repetitions will ends when there is only predefined number of variables existing. It applies normalization to information gain score by utilizing a split information value.

The split information value corresponds to the important information obtained by dividing the training dataset D into v divisions, resulting to v outcomes on attribute A

$$\text{SplitInfo}_A(D) = -\sum_{j=l}^{v} \frac{|D_j|}{|D|} \text{x} log_2 \frac{|D_j|}{|D|}$$

Gain ratio is the ratio between the information gain and intrinsic value can be defined by the following equation,

Gain Ratio (A) = Information Gain(A)/Split Info(A)

3.1.2 Ranker Search

Ranker method is ranked attributes by their individual evaluations Use in conjunction with attribute evaluators (ReliefF, Gain Ratio, Entropy etc.) with the parameter generate ranking (true or false), number to select, and threshold values is set threshold by which attributes can be discarded. Default value results in no attributes are discarded. Use either this option or number to select to reduce the attribute set. The classification, variable ranking is a filter method: it is a preprocessing step, independent of the choice of the predictor]. The ranker method generally performs the rank which attributes should be obtain high or low rank according to the selected attribute in the given datasets. Ranker is providing a rating of the attributes, orderly by their score to the evaluator.

3.2 SENTIMENT CLASSIFICATION

Sentiment classification is the binary classification task of labeling an opinionated document as expressing either an positive or negative opinion. For sentiment classification, we selected two machine learning algorithms like support vector machine (SVM), Naive Bayes (NB) are used.

3.2.1 Support Vector Machine

This is a supervised machine learning approach used to analyze the data and recognize data patterns used for classification and regression analysis. It constructs a set of hyper plane in a high dimensional space, and finding the largest margin to divide the objects of different classes. The significant property of SVMs is that they simultaneously decrease the empirical classification error and increase the geometric margin; therefore it is called maximum margin classifiers. In the binary classification case, the simple idea behind the training

procedure is to find a hyper plane expressed by vector and not only divides the document vectors in one class from those in the other, but for which the division (margin), is possible in greater.

3.2.2 Naïve Bayes

Naïve Bayesian Classification is based on Naïve Bayes theorem as well as the concepts of maximum likelihood and Bayesian probability with strong independence assumptions between the variables. It requires a number of parameters linear in the number of variables in a learning problem. Maximum likelihood training can be done by examining a closed form expression, which takes linear time, rather than by expensive iterative approximation as used for other types of classifiers.

In order to perform classification, we need to compute the posterior probability, P(c|d) Based on the Bayesian probability and the multinomial model, we have

$$P(c|d) = \frac{P(d\backslash c)P(c)}{P(d)}$$

4. Experimental Results

4.1 Dataset description

All the experiments carried out in this section are computed using open source tool Weka. We used three various domain in the experiments are used to investigate the performance of the proposed method using machine learning algorithms such as Naïve Bayes and Support Vector Machine.

4.1.1 Moview Review Dataset

The basic data set consist of 2000 movie reviews, 1000 labeled positive and 1000 labeled 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-eview-data

4.2.1 Spam Review Dataset

This dataset consists of 4601 instances. For our supervised methods, we need to divide the data set into training set and test set. We conduct 10- fold cross-validation.the dataset is randomly split into ten folds where nine folds are selected for training and the tenth fold is selected for test. It is downloaded from—UCI Machine Learning
Repository web page: https://archive.ics.uci.edu/ml/datasets/Spambase

4.1.3 Yelp Review Dataset

This dataset consists of 1960 instances. They were downloaded from https://yelpdataset challenge.com.

4.2 RESULTS AND DISCUSSION

Table 1: Classification accuracy for NB and SVM classifiers

Approach		Existing	Proposed
NB	Movie	87.9	95,4
	Spam	79.2	92.9
	Yelp	75.51	93.4
SVM	Movie	86.5	94.3
	Spam	90.4	92.6
	Yelp	90.1	93.1

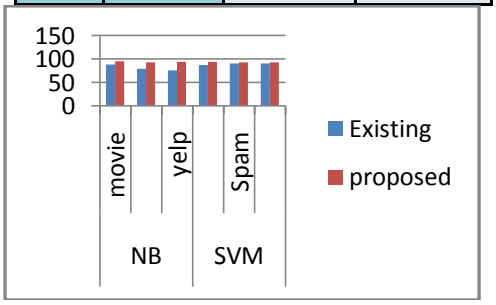

Figure1: Accuracy for Movie, Spam and yelp review.

Table 2: Precision, Recall and F-Measure using NB

NB						
	Movie		Spam		Yelp	
	Before	After	Before	After	Before	After
Precision	0.862	0.96	0.901	0.915	0.558	0.94
Recall	0.879	0.95	0.69	0.922	0.774	0.93
F-measure	0.844	0.95	0.801	0.922	0.649	0.93

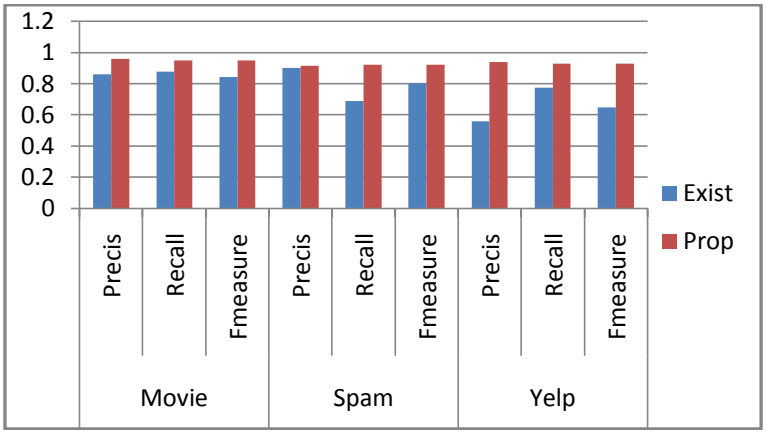

Figure 2: Naïve Bayes Performance Measures

In this experiment we use three different review datasets are selected. It consists of review documents contains various number of features ranging from 27 to 676. They are movie, spam and restaurant review (attributes types are numeric or nominal). Details of these datasets are explained in section 4.1. The weka tool can be used for implementation. Initially the documents are preprocessed in an efficient way to get good results in accuracy. After preprocessing, the feature vector is generated. Further the features are taken into classification algorithms including Naïve Bayes and Support vector machine are applied to the dataset. Before applying classifiers, the data reduction technique such as gain ratio is used as an attribute evaluator. By default the gain ratio uses ranker search method for ranking all the features using parameter values(evaluation using . In order to identifying the optimum number of features for the dataset used, the number of features are varied from min n= 13 to max n= 50. Among the different values, the number of features used and it is observed that the accuracy is max when the value of n is more than 35.

Table 3: Precision, Recall and F-measure using SVM

SVM			
Existing			
	Precision	Recall	F-measure
Movie	0.749	0.866	0.803
Spam	0.896	0.835	0.923
Yelp	0.812	0.86	0.835

SVM			
proposed			
	Precision	**Recall**	**F-measure**
Movie	0.95	0.94	0.94
Spam	0.926	0.817	0.926
Yelp	0.94	0.93	0.93

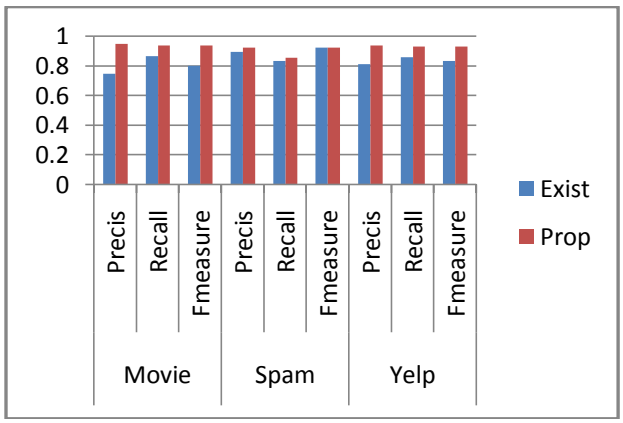

Figure 3: SVM Performance Measure

Then the accuracy is measured using the 2 classifiers. For all domain we are getting accuracy such as 92%, 93% and 100% respectively. The performance of SVM and Naive bayes classifier before feature reduction and after feature reduction is shown in Figure 1, 2 and 3. Thus the classification accuracy obtained through Gain ratio based feature reduction method with classifiers is better than individual classifiers on different review datasets.

5. Conclusion and Future work

This research work focused on feature reduction method with Naïve Bayes and SVM as classifiers used to carry out the sentiment analysis in different datasets such as Movie, Spam and Restaurant reviews. Extracting features from the document (review) through the attribute evaluator gain ratio and ranker search method. The extracted features and the classification algorithms obtained the accuracy greater.

We increased the accuracy using ranker method to rank all the features on the dataset, we are having list of features that have some ranks given by ranker algorithm in association with attribute evaluator. In future, the experiments can be performed with other domain such as web discourse and news articles with a combination of different feature reductions can be implemented.

References

[1]. Pooja Kawade, Nitin Pise, Pradnya Kulkarni, "A Case Study on Sentiment Analysis from Social Big Data", International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, Issue 7, pp. 13152-13156, July 2016.

[2]. R. Piryani a , D. Madhavi b , V.K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015", Information Processing and Management, ELSEVIER, vol. 53, pp. 122-150, July 2016.

[3]. Rasheed M.Elawady, Sherif Barakat, Nora M.Elrashidy, "Different Feature Selection for Sentiment Classification", International Journal of Information Science and Intelligent System, vol. 3, pp. 137-150, July 2014.

[4]. Brian Dickinson, Michael Ganger, Wei Hu, "Dimensionality Reduction of Distributed Vector Word Representations and Emoticon Stemming for Sentiment Analysis" , Journal of Data Analysis and Information Processing, vol.3 no.4, 153-162, November 2015.

[5]. Anuj Sharma , Shubhamoy Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", Special Issue of International Journal of Computer Applications on ACCTHPCA, pp.15-20, June 2012.

[6]. Chetashri Bhadane , Hardi Dalal , Heenal Doshi , "Sentiment Analysis: Measuring Opinions", Procedia Computer Science, ELSEVIER, vol. 45, pp. 808-814, 2015.

[7]. Z. Niu, Z. Yin, X. Kong, "Sentiment classification for microblog by machine learning", Fourth International Conference on Computational and Information Sciences (ICCIS), IEEE, pp. 286–289, 2012.

[8]. Doaa Mohey, El-Din Mohamed Hussein, " A Survey on Sentiment Analysis Challenges", Journal of King Saud University–Engineering Sciences, pp. 1-9, April 2016.

[9]. G. Vaitheeswaran, Arockiam, "Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets", International Journal of Advance Research in Computer Science and Management Studies, vol. 4 Issue 5, pp. 72-82, May 2016.

[10]. P. P. Xu, H. L. Jin, H. X. Shi, and W. Chen, "An Unsupervised Sentiment Information Identification Approach," Applied Mechanics and Materials, vol. 263, pp. 3330-3334, 2013.

[11]. B. Liu, "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies", vol. 5 no.1, pp. 1-167, 2012.

[12]. S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semi-supervised sentiment classification", Neurocomputing, vol.120, pp. 536-540, November 2013.

[13]. Hanhoon Kang, Seong joon Yoo, Dongil Han, " Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications,vol.39, pp. 6000-6010, 2011.

[14]. Anuj Sharma ,Shubhamoy Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", Special Issue of International Journal of Computer Applications (0975 – 8887) on ACCTHPCA, pp. 15-20, June 2012.

[15]. Gautami Tripathi and Naganna S, "Feature Selection And Classification Approach For Sentiment Analysis", An International Journal on Machine Learning and Applications (MLAIJ) , vol.2 no.2, pp. 1-16, June 2015.

[16]. Suge Wang, Deyu Li, Xiaolei Song, Yingjie Wei, and Hong Xia Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification", Expert Systems with Applications, ELSIEVIER, vol. 38, pp. 8696-8702, 2011.

[17]. Shahana P.H, Bini Omman, " Evaluation of Features on Sentimental Analysis" , International Conference on Information and Communication Technologies (ICICT 2014), ELSEVIER, vol. 46 pp. 1585 – 1592, 2015.

[18]. G.Vinodhini, RM.Chandrasekaran, "Effect of Feature Reduction in Sentiment analysis of online reviews", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 2, pp.2165-2172, June 2013.