International Journal of Pure and Applied Mathematics
Volume 118 No. 6 2018, 407-415
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Sentiment Classification using Correlation and Instance Feature Selection

K. Bhuvaneswari[1] and Dr. R. Parimala[2]

[1]*Department of CS, Government Arts College, Kulithalai, India*
bhuvaneswarik27@gmail.com
[2] *P.G and Research Department of CS, Periyar E.V.R. College, Trichy, India*
rajamohanparimala@gmail.com

Abstract

Sentiment Analysis is one of the most and contemporary research area for text analysis using web based data mining and Natural Language Processing techniques. In recent days, the people are sharing their sentiments and opinions in the form of blogs, tweets, face book messages, news groups, comments and reviews. There is a necessity to classify sentiment reviews either positive or negative because people always wish to hear others opinion before making decisions. The proposed model Sentiment Classification using Correlation and Instance Feature Selection(SCCIFS) that extracts only subset of sentiment features using Instance based Monte Carlo(MC) sampling coupled with Correlation feature selection method to improve the performance of sentiment classification using Movie reviews dataset. The Verbs, Adjectives and Adverbs are treated as sentiment words. A set of sentiment words are selected using WordNet based POS (Part Of Speech). The Correlation based feature selection method is applied to select the subset of sentiment features and the reduced subset is applied for Instance based selection using Monte Carlo sampling technique using different combinations. Finally Support Vector Machine (SVM) classifier is used for sentiment classification. The experimental results prove the effectiveness of the proposed model by improving classification accuracy

AMS Subject Classification: Data mining

Key Words: Sentiment Analysis, Correlation, Classification, POS, Support Vector Machine, Feature Selection

1 Introduction

Sentiment Analysis extracts customer's opinions from the web and classifies those opinion reviews using sentiment classification approach whether it is positive or negative. In recent days, the people are expressing their sentiments and opinions on different forums. Different sentiment classification techniques are used to classify sentiments from text data in their suitable class either positive or negative. In this paper, the movie reviews are used for document level sentiment classification because special challenges are associated with movie reviews and it is based on domain specific semantic words[3]. The proposed model used supervised learning techniques on a labelled movie reviews dataset is freely available on the Internet created by Pang and Lee[4]. Sentiment Analysis can be classified at three types namely Feature level, Sentence level and Document level[1]. The Feature level classification retrieves the most important features from text document and classifies whether it is positive or negative opinions. Sentence level classification is to classify the reviews at individual sentence. Document level sentiment classification is used to classify the whole document that comprises as positive or negative reviews.
The SCCIFS model used a WordNet dictionary to generate a set of sentiment words. A feature set is containing adverbs, adjectives and verbs. The reduced feature subset is created using Correlation feature selection and Monte Carlo Instance selection techniques. The main objective of this study is to improve sentiment classification accuracy using feature reduction. The SCCIFS model is experimented with publically available polarity Movie review dataset. The paper is organized as follows: Section 2 provides the details of related work in Sentiment Analysis. The methodology of proposed model is mentioned in Section 3. Section 4 discusses the results of SCCIFS model. Section 5 concludes the paper. The proposed model used Support Vector Machine (SVM) classifier for classifying sentiment reviews in document level and compared with the existing results.

2 Related Work

Pang et al., implemented document level sentiment classification using SVM, Naive Bayes and Maximum Entropy techniques for movie reviews and got 82.90% accuracy for SVM using cross validation for unigram features [2]. Pang and Lee obtained 86.4% of accuracy of document level sentiment classification of the movie reviews using text categorization techniques at document level by applying Naïve Bayes classifier[5]. Chaovalit and Zhou compared supervised and unsupervised algorithm for classification and obtained 83.54% of accuracy for supervised method and 77% of accuracy for unsupervised method[6]. Rushdi et al., explored the Sentiment Analysis task and performed cross validations in SVM using Pang Movie review corpus[7]. Isabella and Suresh used movie reviews for sentiment analysis and assessed a range of feature selectors to improve the performance of the classifiers systematically[8]. Kalaivani and Shanmuganathan applied SVM, NB and KNN algorithm for sentiment classification using movie reviews by applying SVM classifier 3-fold cross validation and obtained accuracy of 81.45%[9]. Mouthami et al., implemented a new algorithm called Sentiment Fuzzy Classification Algorithm to improve classi-

fication accuracy of Movie review dataset[10]. Anitha and Bhargavi implemented a document level sentiment analysis to extract adverb and adjective features for improves the accuracy of classification by using SentiWordNet for calculating the score of a word[11]. The results show Naive Bayes and SVM classifiers are better than SentiWordNet approaches. Gautami Tripathi and Naganna investigated different feature selection methods to obtain the results for sentiment analysis using NB and Linear SVM classification algorithms for unigrams, bigrams, trigrams and four grams[13]. Siddhartha Ghosh et al., discussed the concept of polarity values in sentiment analysis using movie reviews. The Naïve Bayes classifier is applied and got the accuracy of 71% for 10 validations and 70.50% for 50 validations[3]. Abinash Tripathy et al., extracted all features and converted each feature to numerical vectors using Movie review dataset. The vector of features results 89.5% of accuracy using Naïve Bayes classifier[12]. Benito Alvares et al., used sentence level classification of reviews using POS tagging and feature pruning by extracting opinion words using opinion sentences and generate opinion summary using a clustering algorithm[14]. K. Bhuvaneswari and Dr. R. Parimala proposed a new feature selection model Sentiment Reviews Classification using Hybrid Feature Selection (SRCHFS) that extract sentiment feature set coupled with Correlation feature selection method using verbs, adverbs and adjectives. They obtained an accuracy of 92.25% using SVM classifier[15]. In this study, the SCCIFS model focuses to improve an accuracy of sentiment classification of movie reviews dataset using Correlation and MC Instance selection by applying the SVM classifier approach using only on combination of sentiment words (Verbs, Adverbs, Adjectives)

3 Proposed Model

This section presents the design and methodology of SCCIFS proposed sentiment classification. In this study, binary sentiment classification technique is applied to classify the movie into two classes either positive or negative using document level sentiment analysis. The proposed model consists of Feature Extraction, Feature Selection, Instance Selection and Classification. The first step of SCCIFS model is to collect sentiment reviews from movie corpus. The unstructured format of reviews is transformed into structured format. The reviews are preprocessed using tokenization, removing stop words and filter by length.

3.1 Feature Extraction

Feature Extraction is the method of extracting related features. In the existing research on sentiment analysis is measured as all words are features. The proposed SCCIFS model retrieves only three parts of sentiment words as features. The verbs, adverbs and adjectives show a significant role in opinions. The WordNet dictionary is used to tag all the Verbs, Adverbs, Adjectives as sentiment features using Movie reviews dataset. The Term Frequency - Inverse Document Frequency (TF-IDF) word vector is created.

3.2 Correlation based Feature Selection

Feature selection is the process of selecting a subset of features. The filter based correlation feature selection method is applied to select most important features from the extracted sentiment words. A correlation is a relationship between features or data attributes. The features are selected by applying correlation weight which are having highest values. A correlation is a number between -1 and +1 that measures the degree of association between two attributes

3.3 Instance based MC Sampling

Monte Carlo is a computational technique based on constructing a random process for a problem and carrying out a numerical experiment N-fold sampling from a random sequence of numbers with a prescribed probability distribution. This is a simple extension of the random selection, also known as the Monte Carlo algorithm. It repeats the random selection given a number of times (the number of iterations) and selects the best subset. In this algorithm the quality of the set of selected instances is determined by the accuracy of the 1-NN classifier. The selected feature subset is applied for instance based MC sampling technique to reduce the dataset.

3.4 Support Vector Machine Classifier

SVM are based on the concept of decision planes that defines decision boundaries. The aim of the SVM classifier is that finding the hyperplane that maximizes the margin between the two classes. The vectors that define the hyperplane are the support vectors. In this study, SVM model represents each review in vectorized form as a data point in the space. This method is used to analyze the complete vectorized data and find a hyperplane to train a model. The set of textual data vectors are said to be optimally separated by hyperplane only when it is separated without error and the distance between closest points of each class and hyperplane is maximum origin. With the hyperplane, the test reviews are predicted to a class based on which side of the hyperplane they fall on. Researchers have achieved better results in SVM classifier. SVM classifier is applied to the reduced dataset and cross validation is used to measure the performance of classification.

3.5 Performance Evaluation

Accuracy is one of the most common performance evaluation parameter and it is calculated as the ratio of number of correctly predicted reviews to the number of total number of reviews present in the corpus.

The algorithm is given below.

Algorithm SCCIFS

1. Read sentiment review corpus; perform pre-processing such as tokenization, removing stop words and filter tokens by length.

2. Extract sentiment words consists of Verbs, Adverbs and Adjectives

3. Create TF-IDF word vector for extracted sentiment features.
4. Apply Correlation on document term matrix.
5. Sort the features in descending order of correlation value.
6. Select top n% of features with highest correlation values.
7. Apply the selected feature subset into instance selection based on MC sampling.
8. The reduced feature subset obtained from Step 7 is given into SVM classifier to perform cross validation.
9. Evaluate the classifier measures.

4 Experimental Results and Discussions

4.1 Dataset Used

This dataset was prepared by Pang and Lee in order to classify movie reviews [4]. The reviews are collected from Internet Movie Database (IMDB) review site available at http://www.cs. cornell.edu /people/pabo/movie-review-data. The reviews are equally partitioned into 1000 positive and 1000 negative reviews.

4.2 Experimental Set up

The SCCIFS model uses Rapid Miner Studio software with its text processing extension, web mining and WordNet extension. This model is implemented using SVM classifier by grouping different sentiment features using feature and instance selection. First, the data set is preprocessed and Term Frequency –Inverse Document Frequency (TF-IDF) matrix is created. The WordNet dictionary is used to extract three sentiment features verbs, adverbs and adjectives. Second the Correlation based feature selection method is employed to select top most sentiment features using correlation value and the selected feature subset is applied for instance based MC sampling technique. Finally, SVM classifier is applied to the reduced dataset and cross validation is used to measure the performance of classification.

4.3 Results

The SCCIFS model is evaluated using Movie reviews dataset by applying the SVM classifier. The experiment shows that the proposed model gives better accuracy using verbs, adverbs and adjectives by combining correlation feature weight and MC sampling selection methods. 1 summarizes the performance of classification accuracy of Movie review dataset. The top most n% of sentiment features are selected by using correlation value and instance based MC sampling method. In MC technique the proposed model obtained better accuracy by using mixed measure type, Euclidean distances and 200 iterations for sampling to select subset of features.

Table 1: Classification Accuracy using Correlation and MC selection

Top n%	Number of features	Accuracy (%)
0.1	1107	91.75
0.2	2214	93.00
0.3	3321	**93.25**
0.4	4428	93.20
0.5	5535	93.00
0.6	6641	93.10
0.7	7748	91.05
0.8	8855	88.00
0.9	9962	83.75

From the table, the SCCIFS model gives the better accuracy of 93.25% using top most 30% of sentiment features by applying the SVM classifier.

4.4 Comparative Analysis

The results are compared with other similar works on the same dataset; the results of proposed SCCIFS model are promising. 2 shows the results of proposed model with existing literatures of datasets and graphical representation are shown in 1.

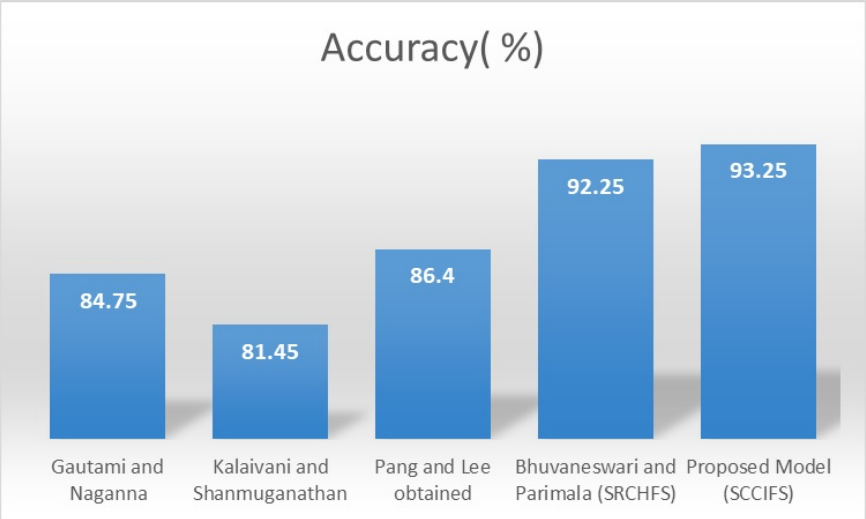

Figure 1: Comparison Results with different models

5 Conclusion

The proposed SCCIFS model presents an approach for sentiment classification by combining correlation feature selection and MC instance selection. The Experimental results show that by combining top most correlation feature weight with MC sampling method achieves the best feature subset for classification and gives better

Table 2: Comparative Results among Different Literatures

Existing Literature	Accuracy(%)
Gautami and Naganna[13]	84.75
Kalaivani and Shanmuganathan[9]	81.45
Pang and Lee obtained[5]	86.40
Bhuvaneswari and Parimala (SRCHFS)[15]	92.25
Proposed Model (SCCIFS)	93.25

accuracy of 93.25% for sentiment movie review data set using SVM classifier. In this paper, the SCCIFS model is implemented using SVM classifier for single domain using only verbs, adverbs and adjectives sentiment words. In future, this model can be extended by applying different classification algorithm by combining with various feature selectors and multi domain data set.

References

[1] Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan and Claypool Publishers, California(2012).

[2] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques, In: *Empirical Methods in Natural Language Processing*, Association for Computational Lingustics (2002), 79-86.

[3] Dr.Siddhartha Ghosh, Sujata M.Thamke, U.R.S Kalyani, Sentiment Analysis using Rapid Miner for Polarity Dataset, *J. Rece. Inno. Trend. Compu. Commu.*, **3**,No.8(2015), 5167–5172.

[4] Polarity dataset version 2.0, Sentiment Analysis Dataset, `http://www.cs.cornell.edu/people/pabo/moviereview-data/reviewpolarity.tar.gz`.

[5] B. Pang, L. Lee, Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts, In: *The 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics (2004), 271–278.

[6] P. Chaovalit and L. Zhou, Movie Review Mining: A Comparison between Supervised and Unsupervised Classification Approaches, In: *The 38th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society Washington(2005), 1-9.

[7] M. Rushdi Saleh, M.T. Martín Valdivia, A. Montejo-Ráez, L.A Urena-Lopez, Experiments with SVM to Classify Opinions in Different Domains,*J. Expe. Syst. Appl.*,**38**, No.12 (2011),14799-14804.

[8] J. Isabella, R.M. Suresh, Analysis and Evaluation of Feature Selectors in Opinion Mining, *J. Comp. Scie. Engi.*, **3**, No.6(2013), 757-762.

[9] P. Kalaivani and K.L. Shanmuganathan, Sentiment Classification of Movie Reviews By Supervised Machine Learning Approaches, *J. Comp. Scie. Engi.* , **4**, No.4(2013), 285-292.

[10] K. Mouthami, K.N. Devi, V.M. Bhaskaran, Sentiment Analysis and Classification based on Textual Reviews, In: *Information Communication and Embedded Systems*,IEEE(2013), 271-276.

[11] B.M. Anitha, B.R. Bhargavi, Opinion Classification Based on Verb, Adverb and Adjectives: Using Various Supervised Machine Learning Algorithms, In: *Multimedia Processing, Communication and Information Technology*,ACEEE(2013), 236-242.

[12] Abinash Tripathy, A. Agrawal and S. K. Rath, Classification of Sentimental Reviews using Machine Learning Techniques, In: *Recent Trends in Computing*,Procedia Computer Science(2015), 821-829.

[13] T. Gautami, S. Naganna, Feature Selection and Classification Approach for Sentiment Analysis, *J. Mach. Lear. Appl.*, **2**, No.2(2015), 1-16.

[14] B. Alvares, N. Thakur and S. Patil, Sentiment Analysis using Opinion Mining, *J. Engi. Rese. Tech.*, **5**, No.4(2016), 88-91.

[15] K. Bhuvaneswari and Dr.R. Parimala, Sentiment Reviews Classification using Hybrid Feature Selection, *Jour. Data. Theor. Appli.*, **10**, No.7(2017), 1-12.