IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Dictionary Based SVM Feature Selection for Sentiment Classification

K. Bhuvaneswari[1*], R. Parimala[2]

[1]Dep. of Computer Science, Government Arts College, Kulithalai, Tamilnadu, India
[2]Dept of Computer Science, Periyar E.V.R. College, Trichy, Tamilnadu, India

[*]*Corresponding Author: bhuvaneswarik27@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 15/Aug/2018, Published: 31/Aug/2018

Abstract- Sentiment Analysis (SA) is the computational study of opinions, sentiments and emotions expressed in text in order to determine the thoughts of people in the direction of certain objects and facts. The opinions of people have a major influence in our every day decision-making process. In recent days, the people are sharing their opinions in the form of blogs, tweets, face book messages, news groups, comments and reviews. The proposed Dictionary Based Support Vector Machine Feature Selection (DBSVMFS) model extracts sentiment features using Support Vector Machine (SVM) weight method to improve the performance of SA. Different levels of pre-processing methods are applied to reduce the features. A set of sentiment features Adjectives, Adverbs and Verbs are extracted by using WordNet based POS (Part-Of-Speech). Feature selection using SVM weight method is applied to select the most important features. SVM classifier is used for sentiment classification and the experimental results prove the effectiveness of the proposed model by improving sentiment classification accuracy.

Keywords: Sentiment Analysis, Classification, Support Vector Machine, Feature Selection, Part-Of- Speech.

I. INTRODUCTION

In recent years, there is fast growth in sentiment analysis of texts. SA is the part of natural language processing for information extraction and used to find out people's emotions expressed in form of positive and negative comments by analyzing many documents [1]. Opinion mining or SA is the sub division of web mining and computational methods for understanding, extracting, classifying and assessing the opinions expressed in various online news sources, social media comments, and other user-generated contents and also analyze polarity of the text document or sentence can be positive or negative. Researches in the field of SA began with the study of the problem of classification and also find how to improve accuracy of sentiment classification. It is also helpful in business intelligence applications and recommender systems where user input and feedback could be quickly summarized [2].

There are three level of sentiment analysis Sentence level, Document level and Feature/Aspect level [3]. Document level sentiment classification is used to classify the whole document is either positive or negative reviews. In Internet world, people use different websites to post their reviews, opinions for the movie and based on that reviews people took the decision. In this paper, the document level sentiment classification is done using movie reviews because special challenges are associated with it and the reviews are domain specific sentiment words.

In sentiment classification all the features in the document are not necessary. The feature or attribute with high relevancy are considered to be important for sentiment classification task. Feature selection plays an important role to select the subset of relevant features to achieve the higher accuracy in classification [4]. Feature selection using feature ranking or weighting is an important part in sentiment analysis and it passes the subset of features as input for sentiment classification and improves the performance of the model [5]. This paper proposes dictionary based feature selection technique and SVM feature weight for sentiment classification. Researchers have achieved better results in SVM classifier. The proposed model is experimented using polarity movie review dataset by applying SVM classifier for classifying sentiment reviews document and compared with existing literatures

The paper is organized as follows: Section I explains the introduction of sentiment analysis classification. Section II provides the details of related work in Sentiment Analysis. Section III describes the detailed methodology of the proposed model. Section IV discusses the experimental results of DBSVMFS model. Section V concludes the paper with future scope.

.

II. RELATED WORK

Nurul et al. [6] implemented a lexical based method for Term Counting and Term Counting Average methods for classifying sentiment of Facebook comments. They extracted verbs, adverbs and negations and construct a list of POS combination of words and scoring methods. The authors concluded that Term Counting method performed better for Adverb words. Oaindrila et al. [7] presented a novel approach for improving classification performance of online movie reviews using machine learning algorithms and POS method. In this paper the author created bigram matrix using adjective and noun combinations using SVM Lite classifier and achieved accuracy of 76.6% for Term Frequency.

Anitha and Bhargavi [8] performed document level sentiment analysis using adverbs and adjectives to enhance accuracy for classifying documents. They found that SVM and Naive Bayes classifiers performance are better than SentiWordNet methods. Bhuvaneswari and Parimala [9] proposed SCCIFS model using verbs, adverbs and adjectives by combining correlation feature weight and Monto Carling sampling selection methods. They used Movie reviews dataset and obtained 93.25% of classification accuracy. Pang and Lee [10] proposed a new machine learning system that applied text-categorization techniques to retrieve subjective sections of the document using minimum cuts in graphs and got 86.4% of accuracy of document level sentiment classification of the movie reviews using Naïve Bayes classifier.

Gautami and Naganna [11] explained various feature selection techniques for sentiment analysis. In their method, unigrams, bigrams, trigrams and four grams features are selected and the maximum classification accuracy of 84.75% is obtained for unigrams using Linear SVM with TF-IDF scheme. Anuj Sharma and Shubhamoy Dey [12] applied different feature selection methods Document Frequency, Information Gain, Gain Ratio, Chi Squared, Relief-F on labelled movie reviews dataset and the experimental results show that information gain gives better performance for different number of features.

Supaporn et al.[13] focused on Relief feature selection technique to improve the performance of sentiment classification using specifically the Naïve Bayes classifier and obtained 80.40% of accuracy using movie reviews. Rajwinder and Prince [14] implemented a new method for improving sentiment classification accuracy by integrating Random Forest algorithm with Gini Index feature selection method to predict sentiment analysis on Movie reviews. Gini Index is applied to select the features that are relevant to the task and Random Forest is used to classify the selected items as positive and negative.

Asha S Manek et al. [15] investigated sentiment analysis for movie reviews using various feature selection methods with Naive Bayes and SVM classifier. The proposed work performed pre-processing, feature selection techniques on movie reviews and the result proved that Gini Index method gave better performance with SVM classification for large amount of dataset and Correlation based feature selection with SVM for small amount of dataset. Shahana and Bini [16] used Mutual Information, Chi-Square, Information Gain and TF-IDF feature selection techniques to select high dimensional feature set. The authors considered unigram, bigram, POS tags of words and function words as feature set and found that unigram is the best method to extract sentiment features and Information Gain gives better accuracy of 83.1% using movie reviews.

Pramod et al., [17] created feature vector using dictionary words and applied Naïve Bayes classifier for sentiment and emotion analysis. They obtained high accuracy using product and movie sentiment reviews.

III. METHODOLOGY

This section presents the design and methodology of DBSVMFS model for sentiment classification using document level. The design of DBSVMFS model is given in Figure 1. The proposed model has Preprocessing, Feature Extraction, Feature Selection, Sentiment Classification and Evaluation. The primary step of DBSVMFS model is to collect sentiment reviews from movie corpus and classify the reviews are positive or negative.

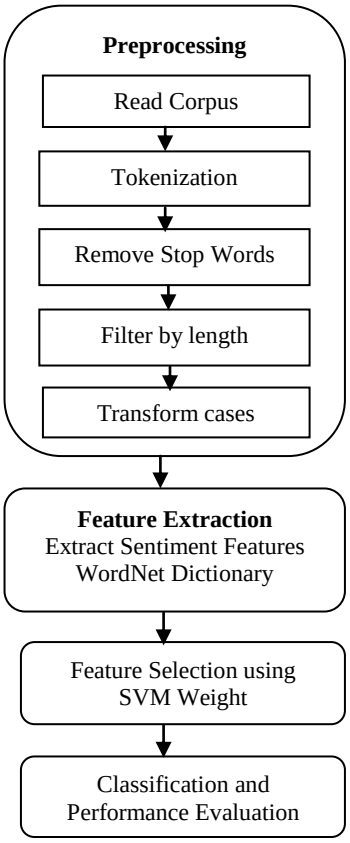

Figure 1. DBSVMFS Model

3.1 Preprocessing

The dataset consists of irrelevant and redundant information. Several preprocessing steps are applied to prepare raw data for further processing. Tokenization is used to split the text of a document into a sequence of tokens. The English stop words are removed from a document and then the reduced tokens are filtered by length. Transform cases is applied to convert all characters of tokens into lower cases.

3.2 Feature Extraction

In order to extract the sentiment features, first perform POS tagging on the entire movie reviews data collection. WordNet is a lexical resource for sentiment analysis to do the tagging and selects all the Adjectives, Adverbs, Nouns and Verbs. Adjectives are mainly used to represent semantic orientations; Adverbs, Nouns and Verbs have also been used to express sentiments. In this study, the proposed DBSVMFS model chooses Adjectives, Adverbs, and Verbs as sentiment features since they play an important role in opinions and Nouns are not included because they are more contexts dependent. The Adjectives (AJ), Adverbs (AD), Verbs (V) and their combinations Adjectives + Adverbs (AJAD), Adjectives + Verbs (AJV), Adverbs + Verbs (ADV), Adjectives + Adverbs + Verbs (AJADV) are extracted and treated as sentiment features. The Term Frequency - Inverse Document Frequency (TF-IDF) word vector is created.

.

3.3 Feature Selection using SVM Weight

Feature selection is an essential part of machine learning and refers to the process of selecting a subset of relevant features and analysis the most meaningful features. Filter based feature selection methods use statistical measure to select the subset of features using the score. The features are ranked by score and features with best scores are used to classify the model. The filter based SVM weight feature selection method is applied to select most important features from the extracted sentiment words. This SVM weight method calculates the relevance of the attribute by computing for each attribute of the input dataset the weight with respect to the class attribute. The coefficients of a hyperplane calculated by an SVM are set as attribute weights. Different combinations of features (F) Adjectives + Adverbs + SVMF (AJADSVMF), Adverbs + Verbs + SVMF (ADVSVMF), Adjectives + Verbs + SVMF (AJVSVMF) and Adjectives + Adverbs + Verbs + SVMF (AJADVSVM) are selected by applying SVM weight which are having highest values.

3.4 Support Vector Machine Classifier

SVM is a supervised learning method applied to analyze the data and identify data patterns that can be used for classification and regression analysis. The aim of the SVM classifier is that finding the optimal hyperplane that maximizes the margin between the decisions using two classes positive and negative. In this study, SVM model represents each review in vectorized form as a data point in the space. SVM classifier find a hyperplane to separate the sentiment documents into classes. This method is used to analyze the complete vectorized data and find a hyperplane to train a model. SVM classifier is applied to the reduced dataset and k - fold cross validation is used to measure the performance of classification.

3.5 Performance Evaluation

Accuracy is one metric for evaluating classification models 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. For binary classification, accuracy can be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Algorithm DBSVMFS

i) Read sentiment corpus.
ii) To perform preprocessing such as tokenization, removing stop words and filter tokens by length.
iii) Apply WordNet dictionary to extract sentiment features containing Adjectives, Adverbs, Verbs and their combinations.
iv) Create TF-IDF word vector for extracted sentiment features, find feature weights using SVM weight method.
v) Rank all features using SVM weight and select top n% of ranked sentiment features.
vi) Apply SVM classifier to perform cross validation using k-fold cross validation
vii) Evaluate the classifier measures.

IV. RESULTS AND DISCUSSION

Dataset Used

The reviews are collected from Internet Movie Database (IMDB) review site available at http://www.cs. cornell.edu /people/pabo/movie-review-data. This study uses a data set of classified movie reviews prepared by Pang and Lee [18] contains 1,000 positive and 1,000 negative reviews and known as polarity dataset v2.0 or Cornell Movie Review Dataset.

Experimental Setup

The DBSVMSFS model uses Rapid Miner Studio software that contains a collection of machine learning algorithms with its text processing Extension and WordNet extension. First, the data set is preprocessed and Term Frequency – Inverse Document Frequency (TF-IDF) matrix is created. The WordNet dictionary is used to extract sentiment words.

Second the SVM weight based feature selection method is employed to assign weights for extracted sentiment features. The features are ranked using their weights and top most ranked sentiment features are selected.

Classification and Evaluation

The DBSVMSFS model is evaluated using movie reviews dataset by applying the SVM classifier. The reduced selected feature subset is applied for sentiment classification and 10 – fold cross validation is used to measure the performance of classification. The experiment results state that WordNet dictionary based feature extraction gives enhanced accuracy, using a combination of sentiment features than Adjectives, Adverbs, and Verbs alone. Table 1 summarizes the performance of sentiment classification accuracy. The top most 30% of sentiment features are selected by using SVM weight and the experiments obtained maximum accuracy for those features.

Table 1. Proposed Classification Accuracy for 10-Fold Cross Validation

Top n% Features	Sentiment Features													
	AJSVMF		ADSVMF		VSVMF		AJADSVMF		AJVSVMF		ADVSVMF		**AJADVSVMF**	
	NF	ACC	NF	ACC	NF	ACC	NF	ACC	NF	ACC	NF	ACC	NF	ACC
0.1	633	92.20	142	79.35	484	89.10	752	92.90	979	93.90	629	92.65	1097	94.90
0.2	1265	94.10	284	81.50	989	91.35	1505	95.65	1958	95.66	1258	94.15	2195	96.05
0.3	1898	94.40	426	82.10	1483	92.10	2257	95.60	2938	96.05	1888	94.60	3292	**96.95**
0.4	2530	94.40	568	81.95	1977	90.95	3009	95.30	3917	95.95	2517	93.50	4389	96.25
0.5	3163	93.15	711	81.80	2472	90.65	3762	94.25	4896	95.10	3146	93.20	5487	95.70
0.6	3795	92.20	853	80.60	2966	88.85	4514	93.20	5875	93.25	3775	91.85	6584	94.20
0.7	4428	90.05	995	78.55	3460	86.40	5266	90.50	6854	91.00	4404	88.70	7681	91.60
0.8	5060	87.85	1137	76.15	3954	91.85	6018	88.75	7834	87.45	5034	84.90	8778	88.55
0.9	5693	82.25	1279	73.70	4449	77.30	6771	84.25	8813	82.25	5663	80.35	9876	84.10

*The bold-faced values indicate better performance. NF – Number of Features ACC – Accuracy From the Table 1, the proposed model gives the better sentiment classification accuracy of 96.95% for AJADVSVMF subset by applying the SVM classifier.

Comparative Analysis

The results are compared with other similar works on the same dataset; the results of DBSVMSFS model are promising. Table 2 shows the results of proposed model with existing literatures of the dataset and graphical representation is shown in Figure 2.

Table 2. Comparative Results among different existing literatures

Existing Literature	Accuracy
Gautami and Naganna [11]	84.75%
Supaporn et al.[15]	80.40%
Pang and Lee [10]	86.4%
Bhuvaneswari and Parimala [9]	93.25%
Proposed DBSVMFS model	**96.95%**

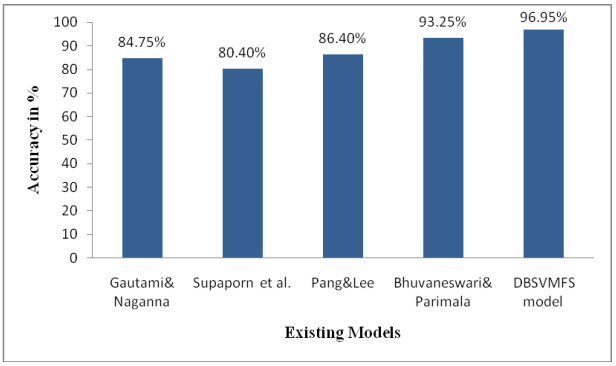

Figure 2.Comparison Results

V. Conclusion

Different feature selection methods are available for sentiment classification. The proposed DBSVMFS model presents an approach for document level sentiment classification by applying SVM weight feature selection

method. This model is evaluated by the combination of sentiment words using WordNet opinion lexicon with SVM value. The results show that AJADVSVMF combination improves the accuracy of sentiment classification. DBSVMFS model is implemented using SVM classifier for single domain using Verbs, Adverbs, Adjectives and their combinations. The future work must focus on to improve the accuracy of sentiment classification by combining different feature selection techniques.

REFERENCES

[1]. Rohini S. Rahate, Emmanuel M, "Feature Selection for Sentiment Analysis by using SVM", International Journal of Computer Applications, Volume 84, No 5, December 2013.

[2]. Supriya B. Moralwar1 , Sachin N. Deshmukh, "Different Approaches of Sentiment Analysis", International Journal of Computer Sciences and Engineering, Volume-3, Issue-3, 2015.

[3]. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, California, 2012.

[4]. Ciurumelea, Adelina, "Analyzing Reviews and Code of Mobile Apps for Better Release Planning", IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER),Austria, 2017.

[5]. Saeys, Yvan, Thomas Abeel, and Yves Van de Peer, "Robust Feature Selection using Ensemble Feature Selection Techniques." Machine Learning and Knowledge Discovery in Database, pp. 313-325, 2008.

[6]. Nurul Fathiyah Shamsudin, Halizah Basiron, Zurina Saaya, "Lexical Based Sentiment Analysis – Verb, Adverb & Negation", Journal of Telecommunication, Electronic and Computer Engineering, ISSN: 2180 – 1843 e-ISSN: 2289-8131 , Vol. 8 No. 2, pp.161-166, 2017.

[7]. Oaindrila Das, Rakesh Chandra Balabantaray, "Sentiment Analysis of Movie Reviews using POS Tags and Term Frequencies", International Journal of Computer Applications (0975 – 8887), Volume 96– No.2, June 2014.

[8]. 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, pp. 236-242, 2013.

[9]. K. Bhuvaneswari and R. Parimala, "Sentiment Classification using Correlation and Instance Feature Selection", International Journal of Pure and Applied Mathematics, Volume 118, No. 6, pp. 407-415. Special Issue, 2018

[10]. B. Pang, L. Lee, "Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts", In the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 271–278, 2004.

[11]. T. Gautami, S. Naganna, "Feature Selection and Classification Approach for Sentiment Analysis", Journal of Machine. Learning Applications, No.2, pp. 1-16, 2015.

[12]. 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 Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, pp. 15-20, June 2012.

[13]. Supaporn, Lonapalawong, Jun, Zhang Le, "Applying Relief Algorithm for Feature Selection in Sentiment Classification for Movie Reviews" Journal of Computational and Theoretical Nano Science, Volume 14, Number 11, pp. 5418-5423(6), November 2017.

[14]. Rajwinder Kaur , Prince Verma, "Sentiment Analysis of Movie Reviews: A Study of Machine Learning Algorithms with Various Feature Selection Methods", International Journal of Computer Sciences and Engineering, Volume-5, Issue-9 E-ISSN: 2347-2693,.pp.113-121, 2017.

[15]. A.S. Manek, P.D. Shenoy,M.C. Mohan and Venugopal K R, "Aspect term extraction for Sentiment Analysis in Large Movie Reviews using Gini Index Feature Selection Method and SVM Classifier", World Wide Web Internet and Web Information Systems Springer, Volume 20, Issue 2, pp 135–154, 2016.

[16]. Shahana Bini Omman, "Evaluation of Features on Sentimental Analysis", In the Proceedings of the International Conference on Information and Communication Technologies (ICICT 2014), Procedia Computer Science 46, pp. 1585 – 1592, 2015.

[17]. Pramod M. Mathapati , A.S. Shahapurkar , K.D.Hanabaratti, "Sentiment Analysis using Naïve bayes Algorithm", International Journal of Computer Sciences and Engineering, Volume-5, Issue-7, 2017.

[18]. Pang, B. & Lee. L, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales", In the Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL), pp. 115–124, University of Michigan, USA, June 25–30, 2005.

Authors Profile

K. Bhuvaneswari is a Research Scholar and currently working as Assistant Professor in Government Arts College, Kulithalai. She received her Master of Computer Applications (MCA) in 2000 and M.Phil Computer Science in 2005 from Bharathidasan University, Tiruchirappalli. Her area of research focuses on Web Mining.

.

R. Parimala graduated with M.Sc. Applied Science at the National Institute of Technology (formerly Regional Engineering College) Tiruchirappalli in 1990. She received her M.Phil Computer Science at Mother Teresa University, Kodaikanal in 1999. She started teaching in 1999 at National Institute of Technology and is currently working as Assistant Professor in Department of Computer Science, Periyar E.V.R. College (Autonomous), Tiruchirappalli. She completed her Ph.D. at National Institute of Technology, Tiruchirappalli. Her area of research interests includes Neural Networks, Data Mining and Optimization Techniques.

