Improving Sentiment Learning using AWQ Feature Selection (SLAWQFS)

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

[1]Assistant Professor, Department of Computer Science, Government Arts College, Kulithalai, Tamilnadu, India
[2]Assistant Professor, P.G & Research Dept of Computer Science, Periyar E.V.R. College, Trichy, Tamilnadu, India
[1]bhuvaneswarik27@gmail.com and [2]rajamohanparimala@gmail.com

Abstract

Opinions are cognitive, affective, behavioural perception of psychological attitudes. Public opinion plays a central role in this interactive society. People take decisions depends on other's opinion to buy a product, making investments, choosing school, watching movies, booking hotels, visiting tourist places etc., In recent days, the people are sharing their opinions in the form of blogs, tweets, face book, news groups, comments and reviews. Document level sentiment classification which classifies an opinion document as positive or negative. Opinion words are features that indicate desirable or undesirable state. The Adjectives and Adverbs are represented as Qualifiers (Q). Verbs are called Action Words (AW). The objective is to extract AWQ features from Review corpus and select top most correlated features using SVM weight. Support Vector Machine Learning (SVML) and Back Propagation based Deep Learning (BPDL) are used for sentiment classification and the results show that the SVML have the better accuracy.

Keywords: Feature Selection, Qualifiers, Correlation, Sentiment Learning, Deep Learning, Support Vector Machine Learning

1. Introduction

Sentiment learning trained to determine the judgment based on a textual comment. In the past days people would seek opinions from friends, relatives, or consumer reports. However, in the Internet era, it is much easier to collect and store various opinions from different people around the world. People look to review sites, e-commerce sites, online opinion sites and social media to get feedback on how a particular product or service may be perceived in the market. Similarly, organizations use surveys, opinion polls, and social media as a system to obtain feedback on their products and services.

Liu [1] stated sentiment analysis has attracted by both the academicians and industry persons because of various challenging research issues for a broad set of applications. The sentiment reviews which are available online as well as off line are containing mostly textual information given by the customer to provide relevant product feedback. Opinion is a subjective expression that describes sentiments of an individual, assessment of performance or emotions about entities, things and their attributes. Researchers studied [2] opinion mining on textual information and proposed method to improve the accuracy of sentiment classification. The subjectivity classification identifies whether a given text document contains opinionated information or not. Then the sentiment analysis is responsible for categorizing an opinion into either positive or negative from the set of opinionated documents because the user is interested to know what opinion have been expressed on certain product, service, or topic.

Sentiment analysis can be categorized into Document Level, Sentence Level and Aspect Level. The document level sentiment learning is trained with a specific set of labeled documents and classifies unlabeled document into one of label either positive or negative. Feature selection is a widely employed technique for reducing dimensionality of features and chooses a small subset of the relevant features from the original ones without any transformation. It usually leads to improve accuracy for classification, lower computational cost, and better model interpretability.

The proposed **SLAWQFS** algorithm extracts AWQ features using WordNet dictionary. The redundant and irrelevant features are removed using correlation feature selection and the top most sentiment features are selected using SVM feature weight.

The main objective of this study is to improve sentiment classification accuracy with dimensionality reduction using SVML and BPDL. The proposed model is experimented with publically available Movie Reviews dataset. The rest of the paper is organized as follows**:** Section 2 provides the details of related work in Sentiment Analysis. Section 3 explains the detailed proposed methodology. Section 4 discusses the experimental results. Section 5 concludes the research work with future scope.

2. Related Work

Techniques of supervised feed forward neural network include Support Vector Machine and Back Propagation. Accuracy is one of the issues in Sentiment Learning. Mustafa et al., [3] founded the relations between sentiments in the levels of document, sentence, and applied verb oriented sentiment classification approach for social domains. Stanford POS tagger is applied to determine the verbs and other elements of opinion structures. Jadav et al., [4] performed sentiment analysis using feature selection and semantic analysis. They performed preprocessing using Stop words removal, stemming, POS tagging and calculating sentiment score with help of SentiWordNet dictionary and applied SVM classification algorithm to classify sentiment reviews. Stephanie [5] explored the role of Parts-Of-Speech (POS) in feature selection using text categorization. The model used different features, namely nouns, verbs, adjectives and adverbs using WordNet based POS were collected. Chi Square and Information Gain feature selection methods were used and obtained the better results with nouns feature set.

Hemalatha et al., [6] developed a new feature selection approach; Shuffled Frog Leaping Algorithm (SFLA) optimizes the process of feature selection and yields the best optimal feature subset which increases the predictive accuracy of the classifier using nouns. Oaindrila et al., [7] presented a novel approach for classification of online movie reviews using POS and machine learning algorithms. The authors employed unigrams, bigrams and POS bigrams tagged phrases using noun and adjective combinations to SVM light classifier. They reported better accuracy of 76.6% on movie reviews using term frequency.

Bhuvaneswari et al., [8] proposes Dictionary Based Support Vector Machine Feature Selection (DBSVMFS) model and selects the combinations of adjectives, adverbs and verbs as sentiment features. Support Vector Machine weight features improve the accuracy (96.95%) of sentiment classification for movie reviews. Gautami et al., [9] experimented various feature selection techniques for Linear SVM and Naïve Bayes sentiment Classifiers. Linear SVM outperformed with high accuracy of 84.75% using higher order n-gram features for movie review. Pinar [10] analysed various filter based feature selection algorithms to predict the risks of hepatitis disease.

Firuz [11] normalized and combined the scores of three major feature selection methods: inter-correlation, information gain and chi-squared statistic and inter-correlation to enhance the classification accuracy. Patrawut et al., [12] implemented semi supervised learning algorithm called Deep Belief Networks with Feature Selection (DBNFS) using chi squared filter based feature selection method. The results indicated that the model speed up the training time and achieved higher classification accuracy. Qurat et al., [13] discussed the advantage of deep learning in the field of sentiment analysis to improve the performance of classification. Arnav et al., [14] proposed a hybrid deep learning method using word embedding. It consists of Convolutional Neural Networks and Recurrent Neural Networks to achieve optimal accuracy using movie review dataset and Stanford Sentiment Treebank data set. Pang and Lee [15] extracted subjective portions of the document, performed document level sentiment classification using Naïve bayes classifier on movie review dataset and achieved 86.4% of accuracy.

Methodology

The goal of this paper is to make a system to classify text sentiment using BPDL and SVM. Figure 1. depicts the framework for Sentiment classification. The SLAWQFS algorithm obtains sentiment features subset and finds the classification accuracy through SVML and BPDL classifier. Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. This algorithm shows the efficiency of the model and it is analyzed in

terms of accuracy. Review corpus represents AWQ features as a vector space model; TF-IDF is used for sentiment learning to classify sentiment as positive or negative. WordNet Dictionary is used for extracting AWQ features; Correlation weight based filter technique combined with SVM weight feature selection selects the feature subset.

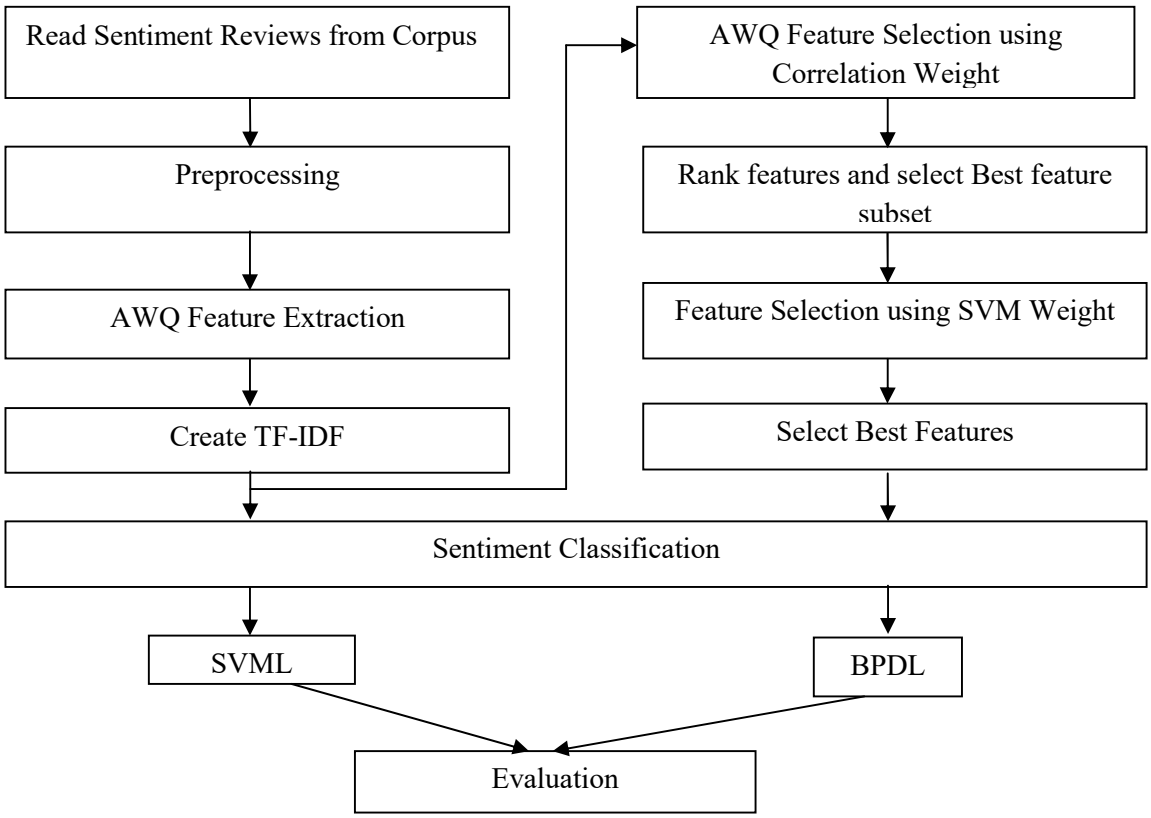

Figure 1. Framework for SLAWQFS Model

3.1 Preprocessing

The labeled review corpus consists of redundant and irrelevant information. Preprocessing steps are applied on the sentiment reviews to optimize it for further experimentations. Tokenization is used to split the reviews of text into a sequence of tokens using unigrams. The stop words in English dictionary are removed and then length based filtration scheme is applied for reducing the generated token set and tokens with less than 3 characters and more than 15 characters are discarded. Finally all the sentiment tokens are converted into lower case letter.

3.2. Feature Extraction using AWQ Features

In the existing study on sentiment analysis considered as all speech words are sentiment features. The proposed SLAWQFS model retrieves only the combinations of ADJ+ADV+VRB as features. The Adjective, Adverbs and Verbs play a major role in opinion mining. The WordNet dictionary is used to perform tagging and extracts all the Adjective, Adverbs and Verbs as sentiment features. The Term Frequency - Inverse Document Frequency (TF-IDF) word vector is created.

3.3 Feature Selection using Correlation Weight

Feature selection is the method for selecting a subset of relevant features. Filter based feature selection methods make use of a statistical measure to get a rank of all features by the score and the best features are selected from the dataset. The filter based correlation feature selection method is applied to select most important AWQ features from the extracted sentiment words. Features may be correlated with one another or redundant. The correlation weight is calculated for ADJ+ADV+VRB combinations of sentiment features. The top most features are selected by applying correlation weight which is having

highest threshold value ranges from 0.1% to 0.5% AWQ features. The correlation of an each attribute is computed with respect to the label attribute using Eqn. (1).

$$\text{Correlation} = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{(n-1)S(X)S(Y)} \quad \textbf{(1)}$$

where X and Y are two attributes with standard deviations S(X) and S(Y) and mean values $\overline{X}$ and $\overline{Y}$.

3.4 Feature Selection using SVM Weight

Feature selection and ranking is helpful to identify relevant features. The proposed model uses linear SVM for both feature ranking and classification and conducts 10 - fold cross validation on the corpus, and choose the parameters leading to the highest accuracy. Feature ranking based on SVM weight require a SVM classifier with dot kernel. The combination of ADJ+ADV+VRB features are selected which are having highest correlation weight and gives as an input for SVM to find the weight for sentiment features. This calculates the relevance of the attributes by computing for each attribute of the input feature set with respect to the class attribute. The features are ranked based on SVM weight and top most features are selected using threshold value ranges from 0.1% to 0.5%.

3.5 Support Vector Machine Learning

SVM is a supervised learning method used for classification or regression. A support vector machine constructs a hyperplane in a high-or infinite dimensional space, which can be used for classification with two class labels. The decision function is of the form

$$y(x) = w^T x + b \quad (2)$$

where x is input vector, Y be the class labels, w and b are optimized weight and bias.. The Linear kernel is represented as

$$K(x,y) = x^T y + c \quad (3)$$

where x and y are vectors in input space, and c is a free parameter.

3.6 Deep Learning Classification

Deep Learning is based on a multi-layer feed-forward artificial neural network which creates non – linear interactions among the features gives a better solution for classification that is trained with stochastic gradient descent using back-propagation. The network can hold a large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation functions. A multilayered feed forward neural network comprises a chain of interconnected neuron which creates the neural architecture starting with an input layer to match the feature space, followed by multiple layers of nonlinearity, and ending with a classification layer to match the output space. Figure 2. shows the Deep Learning architecture along with input and output layers consists of multiple hidden layers more than two.

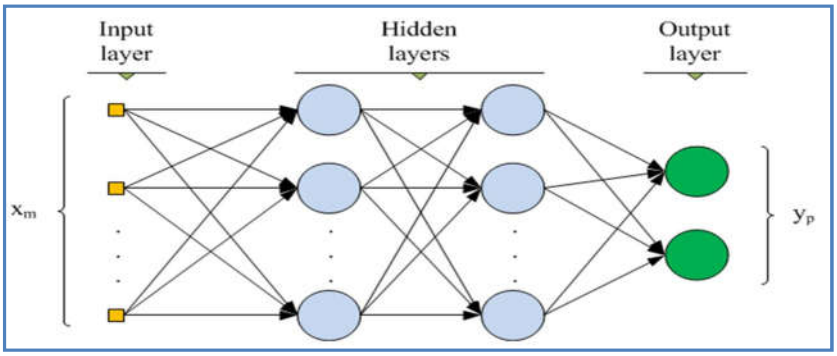

Figure 2. Deep Learning Architecture

The input layer consists of neurons equal to the number of input feature X_m and the output layer consists of two class variables Y_p. The cross validation strategy is applied to find the optimum number of features

in the hidden layer. Multilayered neural networks are ideal when the data set has more number of features.

The objective of the back propagation algorithm is to optimize the weights related with features so that the network can learn to predict the output more accurately. Once the predicted value is computed, it transmits back layer by layer and re-calculates error associated with each feature. The proposed model uses five layers, one input layer, two hidden layers and two output layers with tanh activation function and number of iterations are 200. The hidden layers have 50 features each.

3.7 Classification and Evaluation

The proposed model uses SVML and BPDL classifier for sentiment classification.. A confusion matrix is used to evaluate the performance of a classifier on a set of test data. Accuracy is one of the measures for classification models and it is calculated as the ratio of number of correctly predicted reviews and the total number of reviews present in the corpus. For binary classification, accuracy can be calculated in terms of positives and negatives as follows

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

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

3.8 Algorithm SLAWQFS

The various steps of proposed algorithm as follows

1. Read sentiment corpus.
2. To perform preprocessing such as tokenization, removing stop words, filter tokens by length and t transform cases.
3. Apply WordNet dictionary to extract AWQ sentiment features.
4. Create TF-IDF word vector.
5. Find correlation feature weights using specified threshold value and rank all the features.
6. Select the best subset using top most ranked AWQ sentiment features.
7. The SVM weight is calculated for the features that are selected by step (6).
8. Apply SVML and DL classifier to perform k-fold cross validation.
9. Evaluate the classifier measures.

3. Experimental Results and Discussions

The **SLAWQFS** model uses Rapid Miner Studio software that contains a more number of machine learning algorithms with text processing extension and WordNet extension. The proposed algorithm is validated with real world movie reviews dataset. It has become one of the standard benchmark corpora for sentiment classification. The Classification accuracy is evaluated from each classifier and tabulated in subsequent sections. The proposed model investigated the effectiveness of the choice of combining SVM weight with Correlation feature weight using SVML classification and compares it with BPDL for classifying the sentiment text documents.

4.1 Dataset Used

Movie Reviews: This dataset was prepared by Pang and Lee in order to classify movie reviews collected from Internet Movie Database (IMDB) review site consists of 1000 positive and 1000 negative sentiment reviews available at http://www.cs. cornell.edu /people/pabo/movie-review-data. Each review consists of a plain text file and labeled as positive or negative..

4.2 Research Findings

The **SLAWQFS** model is evaluated using movie reviews dataset by applying the SVM with dot kernel and BPDL classifier. The AWQ features are applied for sentiment classification and 10 – fold cross validation is used to measure the performance of classification. The BPDL classifier uses two

hidden layers with 20 neurons each and one output layer for two class labels positive and negative respectively. The experiment results state that **SLAWQFS** method gives enhanced accuracy, using a combination of AWQ sentiment features than Adjectives, Adverbs, and Verbs alone. .Table 1 summarizes the performance of sentiment classification accuracy for all sentiment features and their combinations using SVML and BPDL classification.

Table 1. Sentiment Classification Accuracy for all AWQ Features

AWQ Features	NF	Accuracy in %	
		SVML	BPDL
ADJ	6325	78.20	80.25
ADV	1421	72.10	69.45
VB	4943	72.95	73.45
ADJ+ ADV	7523	78.60	81.35
ADJ+ VRB	9792	77.70	80.60
ADV+ VRB	6292	75.90	77.70
ADJ+ADV+VRB	10973	79.30	**81.45**

From Table 1, the combination of all AWQ features (ADJ+ADV+VRB) gives maximum sentiment classification accuracy of 81.45% using Deep Learning classifier for 10973 features. The SVM classifier gives 79.30% of accuracy. The comparison is shown in Figure 3.

Figure 3. Sentiment Classification Accuracy using SVM and DL

The Table 2 shows the number of features and classification accuracy for various ranges of threshold values from 0.1% to 0.5% of top most AWQ (ADJ+ADV+VRB) sentiment features using correlation weight by applying SVM and **BP**DL classifiers.

Table 2. Sentiment Classification Accuracy using Correlation Weight

Threshold Value (Top p %)	Number of Features	Accuracy in %	
		SVM	DL
0.1	1097	91.90	91.75
0.2	2195	92.35	91.30
0.3	3292	**93.90**	**92.40**
0.4	4389	92.75	91.35
0.5	5487	92.20	91.10

From Table 2, the proposed model gets maximum sentiment classification accuracy of 93.90% for 3292 features using SVM classification. BPDL classification gives 92.40% of accuracy.

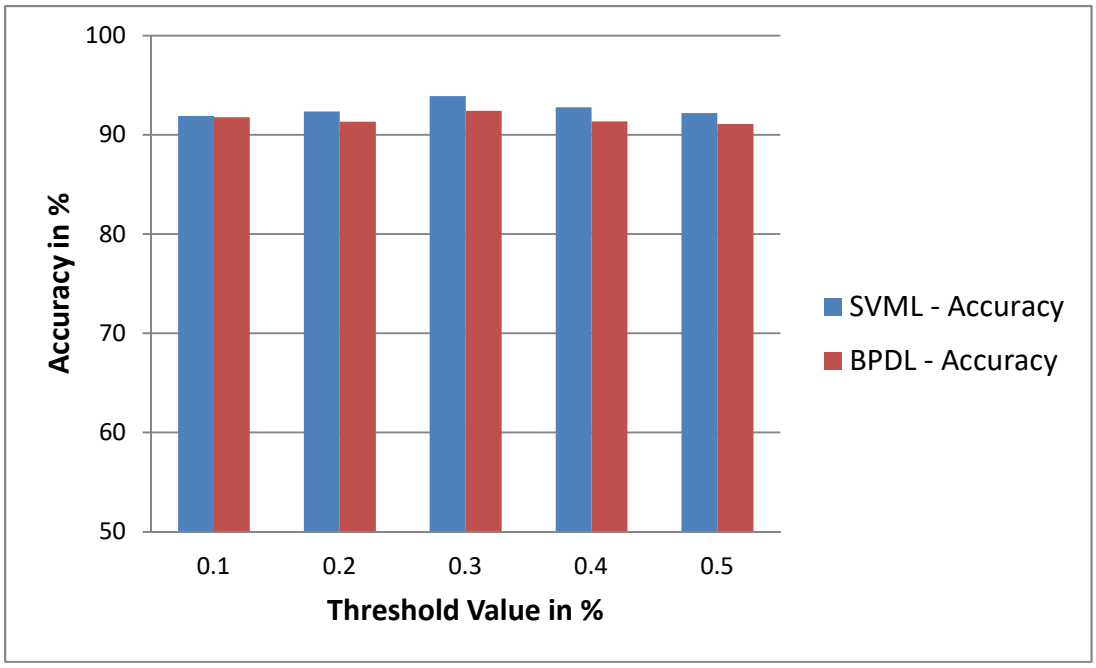

Figure 4. Sentiment Classification Accuracy using SVM and BPDL using Correlation Weight

Figure 4. shows the maximum accuracy 93.90% is obtained for 0.3% of top most combination of all AWQ (ADJ+ADV+VRB) features using SVM classification. The Table 3 shows the experimental results of SL-SVMCAWQFS model using SVM and BPDL classifiers. The SVM gives maximum accuracy of 98.45% for top most 30% of sentiment features and BPDL gives maximum accuracy of 95.30% for top most 40% of features. From Table 3 and Figure 5, the results prove that SVM classifier obtains better accuracy than BPDL classifier.

Table 3. Sentiment Classification Accuracy using SVM Correlated Weight

Threshold Value (Top p %)	Number of Features	Accuracy in %	
		SVM	DL
0.1	329	88.90	90.80
0.2	658	95.65	93.95
0.3	988	**98.45**	94.45
0.4	1317	97.25	95.30

0.5	1646	97.25	94.95

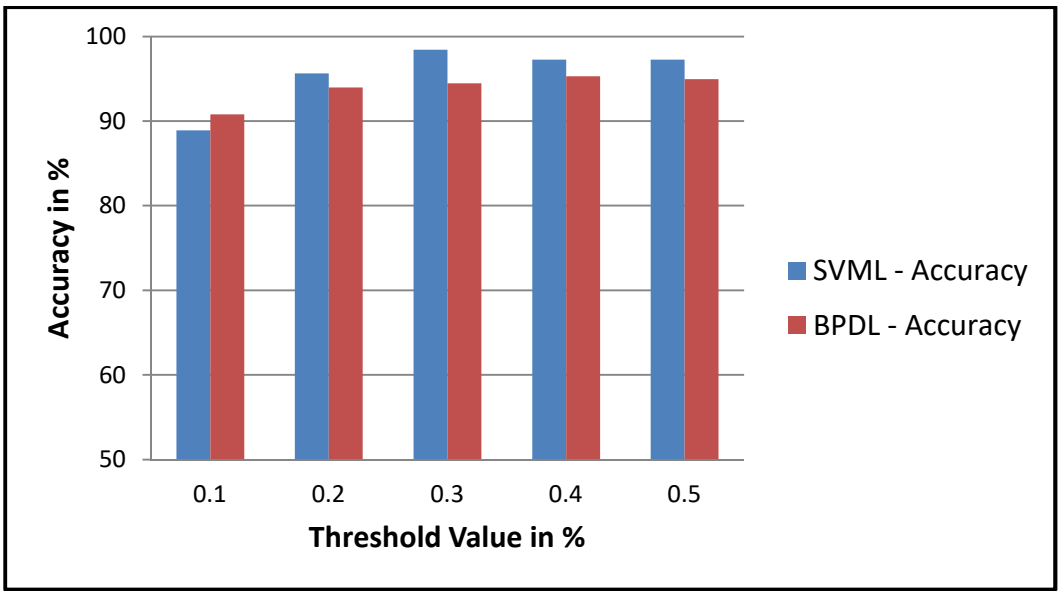

Figure 5. Sentiment Classification Accuracy using SVM and BPDL using SVM Correlated Weight

4.3 Comparative Analysis

The proposed technique showed improved classification accuracy compared to existing methods on the same dataset. Table 4 shows the comparison results of SL – SVMCAWQFS model with existing literatures of the movie reviews dataset and graphical representation is shown in Figure 6. The results show that the SL – SVMCAWQFS algorithm gives improved sentiment classification accuracy.

Table 4 . Comparative Results

Existing Literature	Accuracy in %
Oaindrila et al., [7]	76.60
Gautami et al., [9]	84.75
Pang and Lee [15]	86.4
Bhuvaneswari et al., [8]	96.95
Proposed SL – SVMCAWQFS	98.45

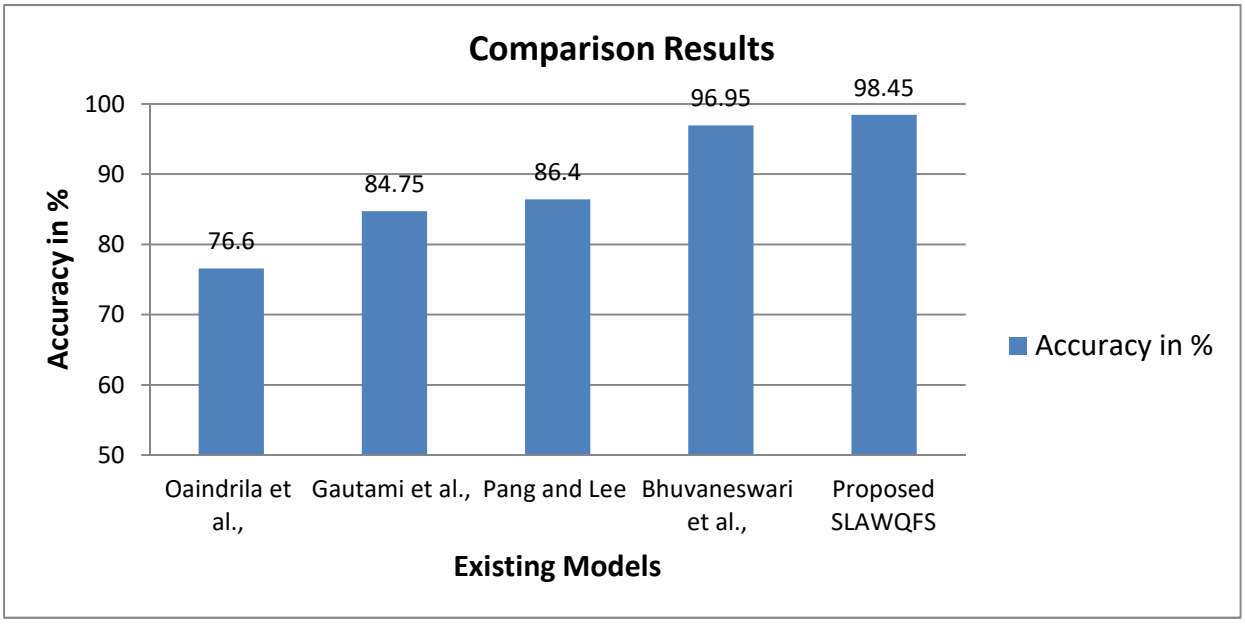

Figure 6. Comparison Result Analysis

CONCLUSION

In sentiment classification different feature selection algorithms are proposed. The proposed SL – SVMCAWQFS algorithm presents an approach for document level sentiment classification by combining SVM weight with correlation feature selection. In this proposed model two classifiers SVM and BPDL are used for assessing the new feature selection algorithm. The results show that ADJ+ADV+VRB combination improves the accuracy of sentiment classification using SVM classifier. The proposed model concludes that the SVM classification gives better accuracy for minimum number of sentiment features and BPDL helps to obtain maximum accuracy for more number of features.SVM Classifier are not useful while working with high dimensional data, which is where we have a large number of inputs. The future work must focus on to improve the accuracy of sentiment classification using Deep Learning of SVM.

References

[1] B. Liu, "Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing", 2nd Edition, (2012).
[2] B.Liu, "Sentiment Analysis: A Multi-faceted Problem", Proceedings of the IEEE Intelligent Systems,(2010), pp. 1–5.
[3] Mostafa Karamibekr, Ali and A. Ghorbani, "Verb Oriented Sentiment Classification", Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology , Vol. 01, (2012).
[4] Jadav, Bhumika M., Vimalkumar B. Vaghela, "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis", International Journal of Computer Applications (0975 – 8887), Vol. 146 – No.13, (2016).
[5] Stephanie Chua, "The Role of Parts-of-Speech in Feature Selection", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008,Vol. I,(2008)
[6] S.M.Hemalatha, C.S. KanimozhiSelvi, "Feature Selection for Opinion Mining Using Shuffled Frog Leaping Algorithm", International Journal Of Engineering And Computer Science, ISSN: 2319-7242 , Volume 7, Issue 2 , (2018), pp. 23656-23662.
[7] Oaindrila Das, Rakesh Chandra Balabantaray, "Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies", International Journal of Computer Applications (0975 – 8887), Vol. 96– No.25, (2014).
[8] K. Bhuvaneswari, R. Parimala, "Dictionary Based SVM Feature Selection for Sentiment Classification", International Journal of Computer Sciences and Engineering, E-ISSN: 2347-2693, Vol.-6, Issue-8, (2018).pp. 603 – 607.
[9] T. Gautami, S. Naganna, "Feature Selection and Classification Approach for Sentiment Analysis", Journal of Machine. Learning Applications, No.2, (2015), pp. 1-16.
[10] Pinar Yildirim , "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease ", International Journal of Machine Learning and Computing, Vol. 5, No. 4, (2015).
[11] Firuz Kamalov, Fadi Thabtah, "A Feature Selection Method Based on Ranked Vector Scores of Features for Classification", Springer, (2017).

[12] Patrawut Ruangkanokmas, Tiranee Achalakul and Khajonpong Akkarajitsakul, "Deep Belief Networks with Feature Selection for Sentiment Classification", 7th International Conference on Intelligent Systems, Modelling and Simulation, (2016), pp. 9-14.

[13] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat and A. Rehman, "Sentiment Analysis Using Deep Learning Techniques: A Review", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, (2017).

[14] Arnav Chakravarthy, Prssanna Desai, Simran Deshmukh and Surbhi Gawande, "HYBRID ARCHITECTURE FOR SENTIMENT ANALYSIS USING DEEP LEARNING", International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Vol. 9, No. 1, (2018), pp.735 – 738.

[15] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts", Proceedings of 42nd ACL, (2004), pp. 271–278.

.