I.J. Modern Education and Computer Science, **2016, 11, 58-64**
Published Online November 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.11.08

Review Length Aware Hybrid Approach to Sentiment Analysis

Babaljeet Kaur
Department of Computer Science Engineering, Punjabi University Regional Centre, Mohali, India
Email: babbaljeet001@gmail.com

Naveen Kumari
Department of Computer Science Engineering, Punjabi University Regional Centre, Mohali, India
Email: naveencse2k4@gmail.com

***Abstract*—**Sentiment analysis is a popular research problem to find out within the natural language processing that is dealing with identifying the sentiments or mood of people's towards elements such as product, text, services and the technology. However, there are few researches conducted on the sentiment analysis of technical article review, so to overcome this deficiency conducts the sentiment analysis over the technical article review and classifying the sentence by overall sentiments that is representing the review is positive or negative. The paper presents the combination of SVM and KNN and find out how much given article sound technically good. The proposed technique is compared with other existing techniques and results shows that the proposed technique is better as compared to the other technique**.**

***Index Terms*—**Sentiment Analysis, SuperFetch Reviews, Support Vector Machine, K-Nearest Neighbor.

I. Introduction

It is challenging conduct the public information about almost any product, service due to big diversity. The sentiment analysis is similar to opinion mining and it is a field of NLP that identifies the people's emotions, opinions, sentiments, evaluation towards entities such as services, products and organization, etc. Sentiments can be expressed by different rating points like 5 stars, 4 stars. With the enormous growth of social data on internet, sentiment analysis becomes a popular research problem to tackle [6]. People express their sentiments or feelings over the internet through the social sites, blogs, forums, rating and comments. Due to this increase in the textual data, there is a need to analyze the sentiments for exploring business [3].

The sentiments of the people can be positive or negative. The positive reviews express the great polarity than the negative reviews. Different machine learning techniques are used to predict whether a document represent positive and negative reviews. Among all machine learning techniques the Support Vector Machine is widely used technique in the field of NLP and give the highest accuracy in recent researches. Sentiment analysis depicts the mood of individual about various entities and their attributes [12]. Various NLP techniques are used to classify the reviews in which consist of SVM, NB and Lexicon etc.

With the rapid growth of various social networking sites like Facebook and Twitter, sentiment analysis becomes more popular in the research area [8]. The various challenges in sentiment analysis is one that the public don't always express their opinions or sentiments in same way means some express in the form of ratings and some in the form of comments and second involving sentences that don't express any sentiment.

The online review sites like www.osnews.com provides abundant information to researchers on how people express their views on these sites. It is a popular site where users express their opinions about different operating system relates articles like Windows 7, Windows 8, MS-DOS, they are interested in reviews, comments & opinion play a crucial role in determining whether users are satisfied with certain article features or event or not. Such reviews have high potential for knowledge discovery.

In proposed work, an attempt has been made to classify the technical article review using the hybrid approach, to find out whether the article is technically sound good or not. The various researches conducted in the field of social sites, product and website reviews and proposed study focuses on the review of SuperFetch to determine the people's expression towards it. After the evaluation of different performance parameters, the result obtained are compared with other existing approach.

This paper is represented as follows: In Section II discusses some previous work done (i.e. Literature survey) in sentiment analysis. Section III represents the methodology along with explains the results and analysis obtained in Section IV. In Section V presents the conclusion and future work for our proposed method.

II. Retated Work

The lot of researches has been conducted in the past in the field of NLP of online review sentiments, upon which our research is based. But very few researches have been conducted for sentiment analysis of article reviews.

Appel and Chiclana [1] presented a combined approach to the sentiment analysis problem and approach is applied at sentence level. The hybrid approach used different natural language processing (NLP) techniques, a sentiment lexicon within SentiWordNet and fuzzy sets used to determine the semantic orientation polarity and its strength for sentences. This method is tested on different type of datasets including twitter and movie review dataset and results obtained are compared with Naïve Bayes and Maximum Entropy techniques. The accuracy achieved by using twitter dataset is 0.8655 and a precision is 0.8406 and hybrid method has been used to test dataset in two types of sub-methods HSC and HAC. The given hybrid approach works best at sentence level and higher level of accuracy obtained than Naïve Bayes and Maximum Entropy techniques. This approach improved the results achieved using Naïve Bayes, Maximum Entropy and also using NLP techniques, sentiment lexicon and fuzzy sets capable to perform best.

Vikash Nandi and Suyash Agrawal [2] applied hybrid approach to political sentiment analysis and in this research twitter is used as data source. The hybrid approach that combines the Lexical dictionary Based technique with features of the SVM learning classifier. In Dictionary Based approach data set words are matched with the dictionary words and Support Vector Machine (SVM) used for exact features which are helpful in classification. When using lexical approach overall 295 tweets, after implementation they get 122 positive and 81 negative tweets. The results demonstrated that Support Vector Machine with LinearSVC raise the classification efficiency and it is 91 and 93% and LinearSVC perform well than the Support Vector Machine Classifier kernel=linear.

Zainuddin and Selamat [7] applied Support Vector Machine approach (SVM) for classification of reviews. The feature extraction methods like term-weighting scheme in which consists of Term Frequency Inverse Document Frequency, Binary Occurrences (BO), Term Occurrences (TO) and the feature selection methods like Chi-Squared were used. They have used two types of dataset; one is a movie review dataset and another SFU review dataset for their experiment. The Support Vector Machine worked well as the classifier because of its capability to handle large features. The result indicated that the comparison of both test and training in terms of the Area under the curve (AUCs) of Support Vector Machine is greater with unigram model rather than weighting scheme. The highest accuracy obtained is 73.21% and 71.05% and Term Frequency Inverse Document Frequency (TFIDF) performs well as compared to Binary Occurrences and Term Occurrences. Then accuracy obtained by using Support Vector Machine classifier is highest when the small number of features is selected. Shoukry [13] comparing the results of both obtained in machine learning experiment and semantic orientation, the accuracy (0.806) obtained in SVM learning algorithm greater than achieved using semantic orientation (SO) technique's accuracy (0.719).

Pang [14] has considered the classification of documents, e.g. identifying whether a review is positive or negative. The three machine learning techniques (Support Vector Machine, Maximum Entropy, and Naïve Bayes) used to classify the movie reviews and discover decisively that these techniques outperform human based baselines.

Grandi [9] has observed that the present sentiment analysis technique is satisfactory for a single entity, but can produce wrong results when with the arrangement of various items. Paper is importing techniques with the help of voting theory and according to the preference aggregation to collect a set of multiple items with high accuracy. The Paper proposed notion of Borda count, which joins public'sentiment, according to similar comparative preference information and demonstrate this class of standards fulfills various properties which a characteristics understanding in sentiment area. The SP (sentiment preference structure) is used over a set of candidates. Borda always behaves better than random procedure in identifying the winner in the complete profile.

Medhat and Hassan [10] have included sophisticated categorization of a many recent articles and illustration of different recent algorithms that are used in sentiment analysis. The fifty four articles considered with their objectives and divided into six categories (SA, SC, ED, FS, BR and TL). After the contribution of number articles to these six categories across several years indicate that still SA and SC attract various researchers more constantly. Machine learning algorithms are mainly used to solve the problem of SC for its capability to use the training data which gives it the right of domain adaptability. The lexicon based algorithms are commonly used to solve the SA problem because they are computationally more efficient. The overall work done in previous years indicates that among ML and Lexicon algorithms, the lexicon based technique used most frequently. Vohra [11] also considered two popular approaches for sentiment analysis: machine learning based and lexicon based approaches. The various classification techniques come under the machine learning and these are mostly used to classify the text. The lexicon based approaches used sentiment dictionary within the different opinion words and match them with data to compute the polarity.

III. Methodology

The proposed study has designed a hybrid approach for developing sentiment analysis process. This section provides detailed design of the proposed approach in which combination of Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) is used.

A. Data Set

The data set is a collection of technical reviews regarding the memory management feature namely SuperFetch introduced by Microsoft in Windows operating system. The reviews are collected from

different resources like websites (www.osnews.com and www.anandtech.com) and collected the reviews from various IT professionals and experts. All the reviews are collected in the format of excel sheet and fetch into the Microsoft SQL server. Collected reviews are important in an experimental environment for the further processing.

B. Preprocessing

Preprocessing is the process to remove those words which are not important to analyze the collected data. In other words, it represents to remove those words whose significance is not important to analyze data. In the proposed work, pre-processing includes the elimination of stop words and the words which represent the noun. Stop Words include the words like it, is, the etc [10]. These words are not useful to identify the positivity and negativity of sentences. Some nouns like name tags in the reviews, words like SuperFetch etc. These words are not necessary to identify the positivity and negativity.

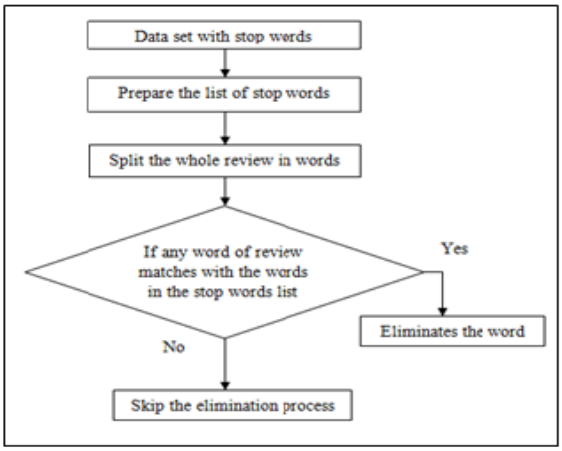

Fig.1. Flow chart of removal of stop words

C. Proposed Approach

The proposed hybrid approach is combined approach of SVM and KNN which classify the SuperFetch reviews. According to the previous related results and study the K-Nearest Neighbor (KNN) is a better approach to analyze the small sized sentence, and Support Vector Machine (SVM) is an approach with better results in the large sized sentences. So the proposed hybrid approach is sentence length size aware approach. If the length of a review is small then the review will be passed to the KNN method for evaluation and if the length of a review is large than the review is evaluated by SVM approach.

A brief working of proposed hybrid approach:

- Collect the data from online websites and from various IT professionals.
- Load the data into Microsoft SQL server.
- Remove the words of review that are found in a STOPWORDS list during the preprocessing step.
- Classify the reviews after the preprocessing of data.
- Classify the reviews accordingly:
  a. If length<threshold
  KNN will evaluate the review
  b. If length>threshold
  SVM will evaluate the review

D. Evaluate the Result

The parameters used for the evaluation of the proposed hybrid approach of article review data set are selected in a way such that the effectiveness of the proposed method can be measured. The result of the proposed approach is evaluated on the based of selected parameters, these parameters are Accuracy, Precision, Recall and F-Measure. These parameters are evaluated the results in three cases in which first case consists of 50 reviews, second consists of 100 reviews and in third case consists of 120 reviews. All parameters are described in detail in next section.

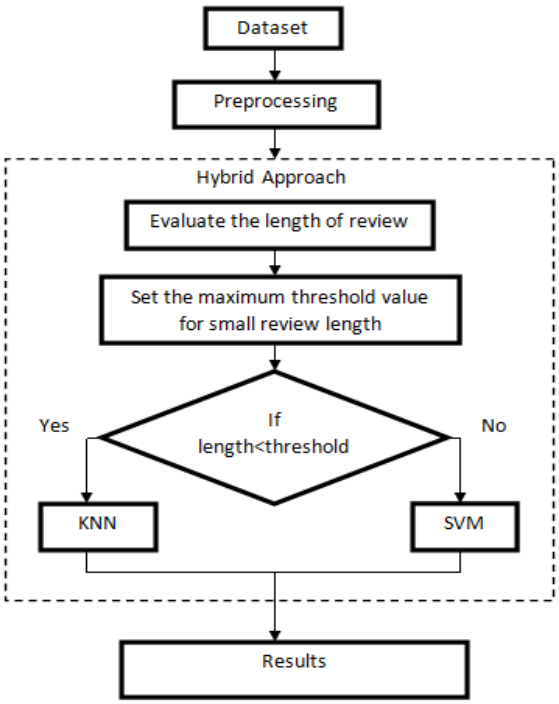

Fig.2. Proposed Work

The above figure represents the working of our proposed work. The proposed hybrid approach is working according to the given condition.

IV. Result and Discussion

The proposed work is carried out in Visual Studio 2010. The LIBSVM is software for Support Vector Machine (SVM) in Visual Studio was utilized for the classifying the text sample and presents the experiment analysis on article review available in Microsoft SQL server. The result is evaluated on the basis of following parameters:

A. Accuracy (A)

Accuracy is a parameter utilized to measure the classification task and it is proportional of correctly classified instances to the total number of instances [4].

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where;

- TP is the number of true positives
- TN is the number of true negatives
- FP is the number of false positives
- FN is the number of false negatives

B. Precision (P)

Precision measure is the ratio of the number of correct positive results and number of all positive results [15]. It measures the exactness of any classifier. The higher the precision means that less false positives (FP), whereas the lower precision means that more the false positives are.

$$\frac{TP}{TP+FP} \quad (2)$$

C. Recall (R)

Recall is the ratio of the number of correct positive results and number of positive results that should have been returned [15]. It measures the completeness or the sensitivity of the sentiment classifier. Higher the recall means that small false negatives (FN), whereas lower the recall is more false negatives it leads to.

$$\frac{TP}{TP+FN} \quad (3)$$

D. F-Measure (F1 Score)

F-Measure or F1 Score is the harmonic mean of the Precision and Recall. F-measure reaches it's best worth at 1 and worst at 0 [5].

$$\frac{2\times Precision\times Recall}{Precison+Recall} \quad (4)$$

The proposed work is divided into three cases.

- Case 1: First case evaluates the described parameters using 50 reviews.
- Case 2: Second case evaluates the all parameters using 100 reviews.
- Case 3: Third case evaluates the parameters using 120 reviews.

Table 1. Results of proposed hybrid approach in all cases

Parameters	Proposed Hybrid Approach		
	50 Reviews	100 Reviews	120 Reviews
Precision	0.84	0.89	0.94
Recall	0.97	0.96	0.96
Accuracy	84.31	87.13	90.74
F-Measure	0.90	0.92	0.95
Positive %	62.00	70.00	68.33
Negative %	8.00	5.00	5.00

The above table shows that the result evaluations in all cases. The maximum accuracy achieved by the proposed hybrid approach is 90.74% which proved the effectiveness of proposed approach. All other parameters also produced better results in all cases.

The highest recall is 0.97 in case of 50 reviews; the highest precision is 0.94 in case of 120 reviews and the highest F- measure is 0.95 also in case of 120 reviews. Positivity of reviews come greater in 100 reviews than the other cases. Now all the parameters result are presenting in graphical form.

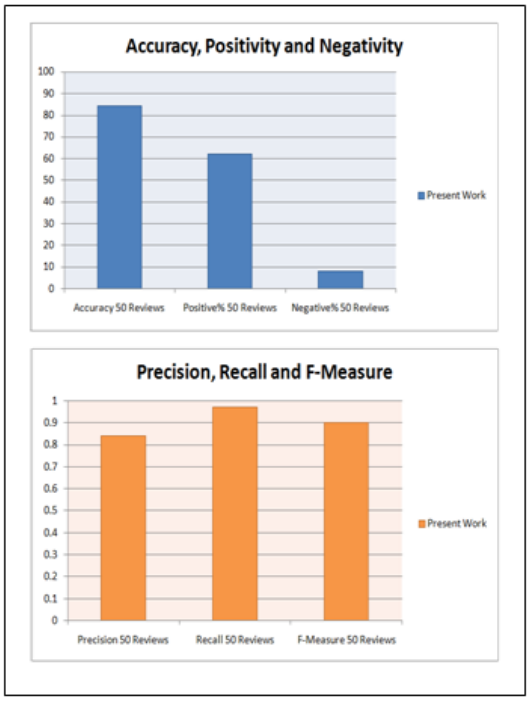

Fig.3. Result evaluations in proposed work using 50 reviews

The above figure represents two different plots using the 50 reviews.

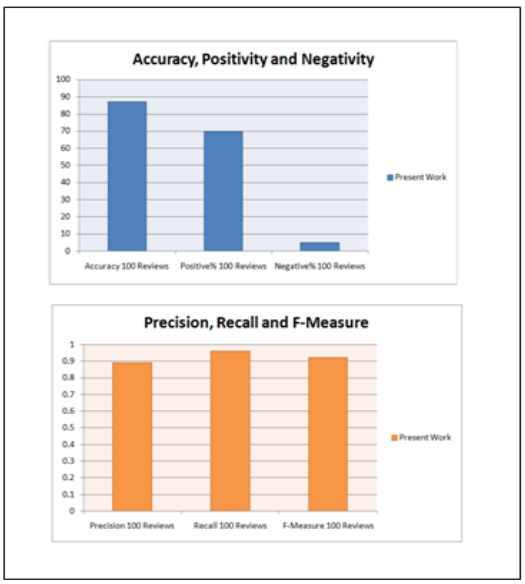

Fig.4. Result evaluations in proposed work using 100 reviews

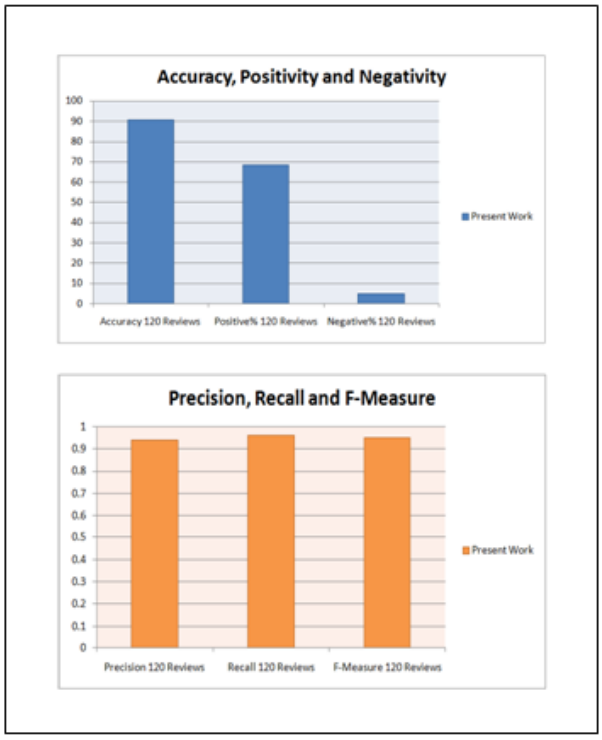

Fig.5. Result evaluations using 120 reviews

The figure 4 and 5 represents the result using 100 and 120 reviews of SuperFetch and in both cases positivity is greater than the negativity. The accuracy achieved is maximum in case of 120 reviews which indicate that the proposed hybrid approach performs well for a large number of reviews. The above provided results were for our proposed hybrid method. Now compared our results with a Lexicon approach.

Table 2. Comparison of proposed hybrid approach with lexicon approach

	Proposed Hybrid Approach			Lexicon Approach		
	50 Reviews	**100 Reviews**	**120 Reviews**	**50 Reviews**	**100 Reviews**	**120 Reviews**
Precision	0.84	0.89	0.94	0.76	0.78	0.81
Recall	0.97	0.96	0.96	0.97	0.96	0.96
Accuracy	84.31	87.13	90.74	76.79	77.88	80.33
F-Measure	0.90	0.92	0.95	0.85	0.86	0.88
Positive %	62.00	70.00	68.33	52.00	58.00	56.67
Negative %	8.00	5.00	5.00	8.00	5.00	5.00

The above table shows that the proposed method is more efficient as compared to the lexicon approach because the highest accuracy achieved by the lexicon approach is 80.33% among all the cases, but highest accuracy achieved by the proposed method is 90.74% using 120 reviews which is much greater than the lexicon approach. The Recall and F-measure of the proposed approach is above the 0.90 which making the proposed method highly effective for sentiment analysis of article reviews.

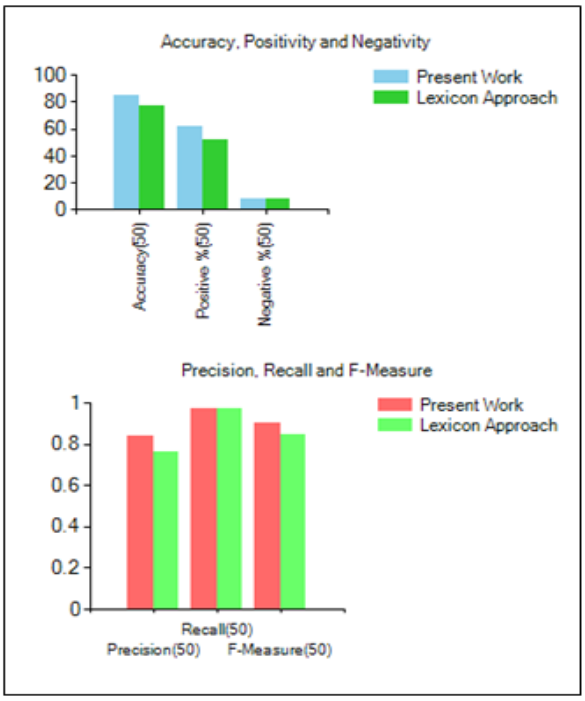

Fig.6. Comparison of the results of two approaches using 50 reviews

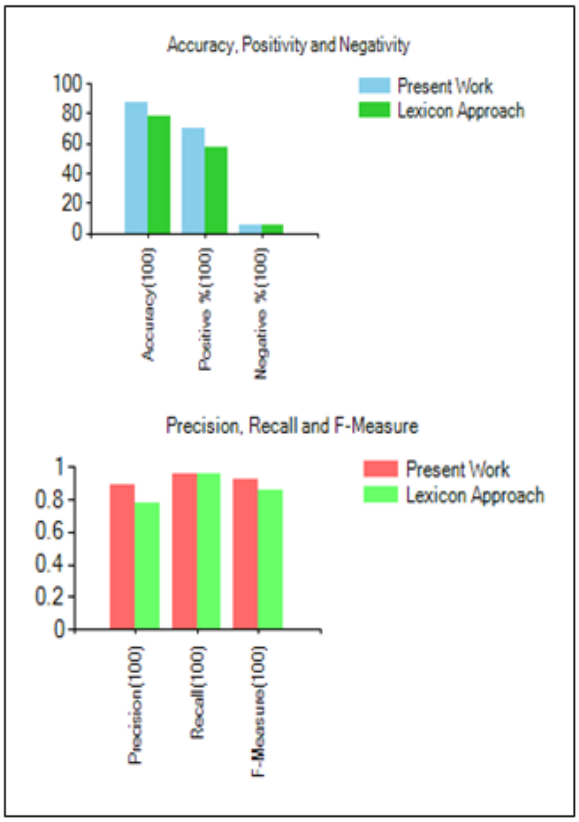

Fig.7. Comparison of the results of two approaches using 100 reviews

The figures 6 and 7 represent the result evaluation of both approaches in case of 50 and 100 reviews. As we can see Accuracy, Precision, F-Measure has higher values in the proposed method than the lexicon approach.

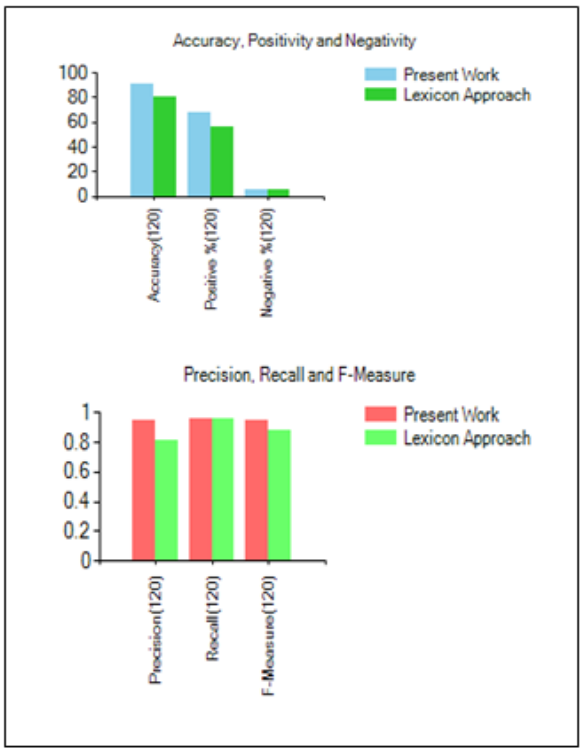

Fig.8. Comparison of the results of two approaches using 120 reviews

The above figure shows the two graphs which provide the comparison of proposed hybrid method with lexicon approach. The proposed method improves the performance in all cases based on the various parameters.

V. Conclusion and Future Work

Sentiment analysis has been absolutely more popular in order to understanding the public's opinions about article, technology and product. The paper considered the combination of SVM and KNN approaches and tested on the SuperFetch dataset. The proposed hybrid approach has been performed better constituting supervised machine learning approach both SVM and KNN to technical reviews. Support Vector Machine classifier increased the performance in case of large reviews and K-Nearest Neighbor increased the performance in case of small reviews. The proposed hybrid approach performance has been measured in terms of Precision, Recall, F-Measure and Accuracy and these parameters have been produced more satisfactory results. So results produced have been proved that the SuperFetch is a good feature in a memory management system.

In the future work, the proposed hybrid approach will be applied for multiple article review like a news article, Window based article etc. together to sentiment analysis. The work on technical review, which can be extended using machine learning and lexicon or both combined. It also includes analyzing and improves the performance of the proposed hybrid approach based on the time taken by same.

Acknowledgment

I am thankful to my guide Assistant Professor Mrs. Naveen Kumari for all help and valuable suggestion provided by her throughout the work.

References

[1] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A Hybrid Approach to Sentiment Analysis with Benchmarking Results," *In Proceedings of 29th International Conference on Industrial Engineering and Application of Artificial Intelligence and Expert Systems, IEA/AIE, LNAI 9799*, 2016, pp. 242-254.

[2] V. Nandi, and S. Agrawal, "Political Sentiment Analysis using Hybrid Approach," *International Research Journal of Engineering and Technology*, vol. 3, issue. 5, pp. 1621-1627, May. 2016.

[3] A. Tripathi, A. Agrawal, and S. K. Rath, " Classification of Sentimental Reviews using Machine learning Techniques," *in Proceedings of 3rd International Conference on Recent Trends in Computing, ICRTC*, 2015, pp. 821-829.

[4] Nagamma, et al., "An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction," *in Proceedings of the International Conference on Computing, Communication and Automation (ICCCA2015*), IEEE, pp. 933-937.

[5] Y. Sharma, V. Mangat, and M. Kaur, "Sentiment Analysis & Opinion Mining," *International Journal of Soft Computing and Artificial Intelligence*, vol.3, issue. 1, pp. 59-62, 2015.

[6] C. Li, et al., "Recursive Deep Learning for Sentiment Analysis over Social Data," *in Proceedings of International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM*, 2014, pp. 180-185.

[7] N. Zainuddin, and A. Selamat, "Sentiment Analysis using Support Vector Machine," *in Proceedings of IEEE International Conference on Computer, Communication and Control Technology, I4CT*, 2014, pp. 333-337.

[8] C. Wang, et al., " SentiView: Sentiment Analysis and Visualization for Internet Popular Topics," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, Nov. 2013, pp. 620-630.

[9] U. Grandi, A. Loreggia, F. Rossi, and V. Saraswat, "A Borda count for collective sentiment analysis," *Annals of Mathematics and Artificial Intelligence*, 22 October 2015, DOI: 10.1007/s10472-015-9488-0.

[10] E. Haddi, X. Liu, and Y. Shi, "The Role of text pre-processing in sentiment analysis," *Information Technology and Quantitative Management*, pp.26-32, 2013.

[11] S. M. Vohra, and PROF. J. B. Teraiya, "A Comparative Study of Sentiment Analysis Techniques," *Journal of Information, Knowledge and Research in Computer Engineering*, vol. 2, issue. 2, pp. 313-317, Nov-Oct. 2013.

[12] S. Modha, G. S. Pandi, and S. J. Modha, "Automatic sentiment analysis for unstructured data," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol.3, pp.91-97, December -2013.

[13] A. Shoukry, and A. Rafea, "A Hybrid Approach for Sentiment Classification of Egyptian Dialect Tweets," *in*

Proceedings of First International Conference on Arabic Computational Linguistics, 2015, pp. 78-85.

[14] B. Pang, et al., "Thumps up? Sentiment Classification using Machine Learning Techniques," *in Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP, Philadelphia*, July. 2002, pp. 79-86.

[15] http://machinelearningmastery.com

Authors' Profiles

Babaljeet Kaur has completed her M.Tech in Computer Science Engineering from Punjabi University Regional Centre for Information Technology and Management, Mohali, India. Her research interests include Natural Language Processing, Digital Image Processing and Big Data.

How to cite this paper: Babaljeet Kaur, Naveen Kumari,"Review Length Aware Hybrid Approach to Sentiment Analysis", International Journal of Modern Education and Computer Science(IJMECS), Vol.8, No.11, pp.58-64, 2016.DOI: 10.5815/ijmecs.2016.11.08

