ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

SVM and KNN based Hybrid Approach to Sentiment Analysis

Babaljeet Kaur*, Naveen Kumari**
Department of Computer Science and Engineering, Punjabi University Regional Centre, Mohali (Punjab), India

Abstract- Sentiment analysis is a popular research problem to find out in the field of natural language processing. The largest online review of product and services are created due to the rapid development of technology, which is an important source for people to gather information. Sentiment analysis allocates positive and negative polarity to an entity or items by using different natural language processing tools and also predicted high and low performance of various sentiment classifiers. The paper presents the hybrid approach to identify the reviews as positive and negative. The combination of SVM and KNN is used, in which SVM classifier is working best for large length reviews and KNN for small length reviews. This approach is tested on SuperFetch review. The SuperFetch reviews are showing that whether the hybrid approach gives an effective performance of sentiment analysis or not.
Index Terms - Sentiment analysis, SuperFetch review, SVM, KNN.

1. INTRODUCTION

The sentiment analysis found in the form of comments, reviews and feedback and provides necessary information for various purposes. These opinions or sentiments can be divided into two categories: positive and negative; or also categories of different rating points (e.g. 5 stars, 4 stars and 3 stars, etc.). The polarity of sentiments like “good” and “bad” also identify the sentiments either positive or negative [2].
With the rapid growth of various social networking sites such as Facebook and Twitter, sentiment analysis becomes more popular in the research area. The various challenges in sentiment analysis is one that the public don’t always express 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 sentiment analysis process is shown in figure 1. The text preparation step performs required text preprocessing and cleaning on the dataset which including removal of stop words. Sentiment identification step determines the sentiment of people expressed in the text and analyzes it. Finally, sentiment classification is conducted to get the results [3].

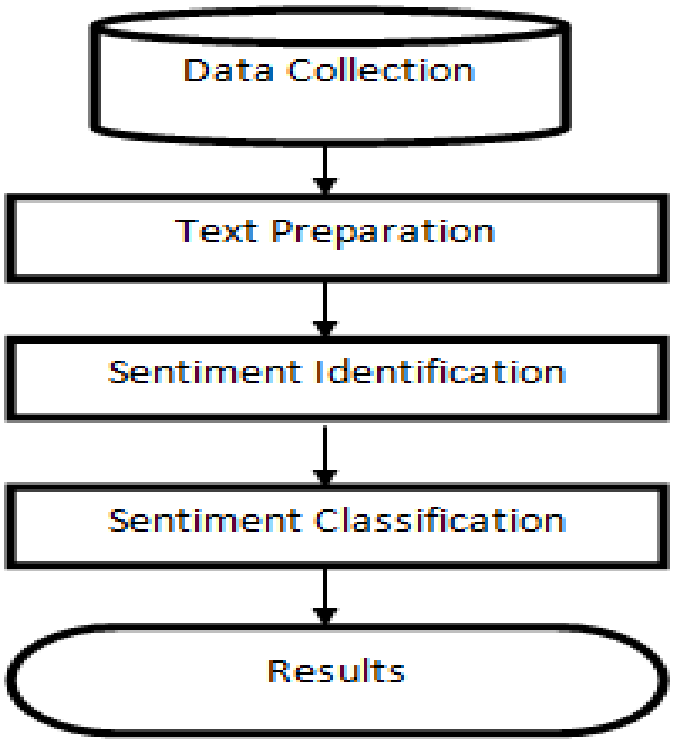

Fig. 1.1 Sentiment Analysis Model

Sentiment analysis concerns with express opinion based on either document and sentence level. In addition, the strength of the sentiments or opinions and target of the sentiments to find out. It also facilitates politician in order to analyse the public opinions with respect to political issue and policies. The paper presents an approach to determine how sentiments can be classified using Support Vector Machine and K-Nearest Neighbor approach. The paper provides the comparison with other existing technique, shows that the use of hybrid approach can improve the efficiency of sentiment analysis. The proposed hybrid approach gives better result as compare to the existing technique. The rest of the paper is described as follows: Section 2 introduces some related work done in this field. Section 3 discusses proposed work done along with explain the results and analysis obtained in Section 4. Section 5 presents the conclusion and future work for the proposed work.

2. RELATED WORK

The large numbers of researches have been conducted previously in the field of natural language processing of online review sentiments, upon which our research is based. The various researchers have different way to express the sentiments and the number of tools to classify these sentiments.

Tan et al [4] presented sentiment categorization based on the Chinese language documents with the size of 1024 documents. The feature selection methods (IG, MI, CHI, and DF) and machine learning methods (K-Nearest Neighbor, Winnow classifier, Centroid classifier, SVM and Naïve Bayes) are conducted to find out the Chinese language sentiment. The results produced suggest that IG performance best in sentiment phrases selection and SVM for sentiment classification and it contains the reviews from three different domains like movie, education and house, also found that sentiment classifier mainly dependent on the topics. The dataset consists of 507 documents related to education, 248 to house and 266 documents related to the movie. The Precision, Recall and F-Measure performance parameters are used.

Li et al [2] combined rule based classification, supervised learning and machine learning and movie reviews, product reviews and MySpace reviews considered for testing of methods. The paper shows that hybrid method is using different classifiers can improve the performance of sentiment analysis. The use of hybrid approach produces results with better efficiency in the case of macro- and micro-averaged F1 measure rather than individual classifier. F1 is a measure that takes both precision and recall parameters to measure the performance of any classifier. In addition, paper purposes a complementary and semi-automatic method to identify the high level of effectiveness and hybrid classification also used the induction algorithms and best results provided by the ID3 and SBC uses two rule sets (ID3 and SBC).

Nagamma et al [1] applied the sentiment analysis and machine learning methods to identify the online movie review and to find out the performance of movie box office revenue. The document level sentiment analysis is conducted along with term frequency and inverse document frequency is selected as features. The SVM and classification model is created for online movie reviews, and find the box office collection of movie according to these reviews. The data collected from the website, such as IMDb and apply TF-IDF with fuzzy clustering, which produces the positive and negative sentiments and accuracy obtained for prediction of box office revenue improved by clustering approach. With the clustering method the accuracy obtained by the SVM classifier from 62% to 89.65% and NB produced 72.4%.

3. PROPOSED WORK

This section provides the detailed description of the number of steps followed for the sentiment analysis of the technical article review and prediction of the SuperFetch review collection. The proposed work included the following steps:

- Dataset
- Preprocessing
- Sentiment classification
- Evaluate the results

These are discussed below:

3.1 Dataset

The SuperFetch dataset is used to search the raw input data. Www.osnews.com is a website provides the information of SuperFetch and other OS related articles. From the website consists of detailed information regarding the positive and negative reviews of SuperFetch given by various OS experts.

The reviews of SuperFetch are also collected from the different professionals of Universities/Colleges in proper format, including date, name of University/College, in the form of excel sheet. After the collection of reviews, data

is saved in sequential order and large information was drawn from this sheet like review, rating and efficiency along with the date.

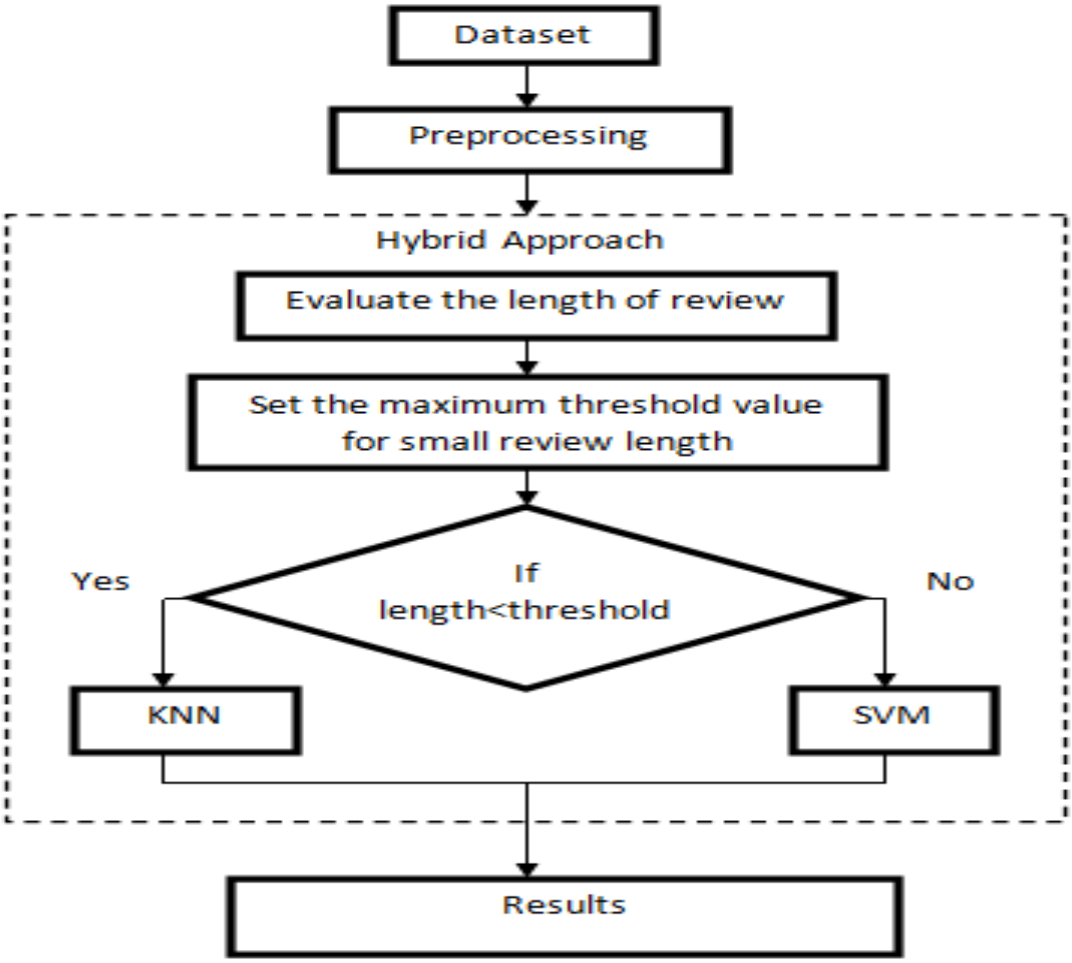

Fig. 3.1 Proposed Work

3.2 Preprocessing

The online data consist of lots of vague information, it needs to clean. The preprocessing is the method used to prepare the data before using it for further classification. The text preprocessing tasks are contained the following categories.

- Removal of Stopwords: The stop words are the list of words which are removed before the processing of natural language data. In natural language processing system and a search engine may involve a variety of stop words, one per language, or it may involve a single stop-list that can be multilingual [1]. Some of most often used stop words in the English language include a, and, the, he, she, it, these are called as functional words which don't contribute any information of sentiments. When using the data and contents of natural language, meaning can be expressed more clearly by discarding the function words.
- WordNet: The WordNet is a lexical database of English language based on psycholinguistic studies and all English words is linked with each other by semantic relationships. It was formed as a text preprocessing resource that covers lexico-semantic categories called as synsets. The synsets are the sets of synonyms that collect lexical items having equal significances [5]. Synsets refers to the actual fact that a lot of word forms will have multiple meaning, a characteristic referred to as "Polysemy". That is, one word will have totally different meaning in several contexts and its even probability for one word to be used as different elements of speech (noun, verb etc.). For example, "fly" will consult with associate inspect (noun) or the act of moving through the air (a verb).

3.3 Sentiment Classification

The proposed hybrid approach is used to classify the reviews of SuperFetch in which K-Nearest Neighbor gives the best performance in case of small reviews if the review length is less than the threshold value and Support Vector Machine more efficient classifier for large type of review like if the length of a review is greater than the given threshold value. In the proposed work for sentiment classification, first compute the length of reviews and set the maximum threshold value for the effectiveness of hybrid approach in term of sentiment analysis of SuperFetch review.

Support Vector Machine: SVM learns to assign a label to the text and give a polarity score like +1 or -1, it should be classified a text as positive and negative. Vocabulary is generating to write every word in alphabetical order. Eventually for each sentence, need to generate a document-term matrix. Before train SVM classifier or model to review classification, there will be need to prepare the data. In the creation of the document-term matrix contains the following steps:

- Initially start the line with class of sentence like +1 or -1.
- Write the index of the word, based on the vocabulary.
- Finally, add a colon, and then the number of time word comes in the sentence.

K-Nearest Neighbor: It is the simplest classifier mainly depends on the category labels. KNN uses the parameter "k", in classifying the object. The main concept behind the KNN is first trained the system for existing data and find a predefined number of training sample nearest in distance to the new point and estimate the labels from these. The amount of samples can be user defined constant and the common alternative for similarity measure is the Euclidean distance equation.

$$d_{Euclidean}(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad 3.1$$

3.4 Evaluate the Result

The different performance measures parameters are used to evaluate the results are Accuracy, Precision, Recall, F-Measure.

4. RESULTS AND ANALYSIS

The article reviews for the article "SuperFetch" was considered as dataset for the sentiment analysis. These reviews were obtained from the OS news online site and from College/ University's professtionals and in this dataset consists of 120 reviews stored in the form of excel file. All the experiments were carried out in VISUAL STUDIO 2010.

4.1 Evaluation Metrics

The classification metrics considered for the sentiment analysis are Accuracy, Precision, Recall and F-Measure and these parameters are evaluated based on the calculated positivity and negativity of reviews by the proposed hybrid approach. These parameters are also determined the effectiveness of the proposed hybrid approach and the common way of computing these metrics based on the contingency table as shown below:

Table-4.1 Contingency table

	Correct	**Not Correct**
Selected	True Positive (TP)	False Positive (FP)
Not Selected	False Negative (FN)	True Negative (TN)

- Accuracy: Accuracy is a common measure for the classification performance and it's proportional of correctly classified instances to the total number of instances, whereas the error rate uses incorrectly classified rather than correctly. Eq: 2 show the mathematical formula for accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad 4.1$$

- Precision: Precision is the percentage of selected items that are correct. Eq. 4.2 show the mathematical formula for precision.

$$\text{Precision} = \frac{TP}{TP+FP} \quad 4.2$$

- Recall: Recall is the percentage of correct items that are selected. Eq. 4.3 indicate the mathematical formula for recall.

$$\text{Recall} = \frac{TP}{TP+FN} \quad 4.3$$

- F-Measure: A combined measure that evaluates the $\frac{P}{R}$ tradeoff is the F- Measure and also called as weighted harmonic mean. Eq. 4.4 shows the mathematical formula for F-Measure.

$$\text{F-Measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad 4.4$$

4.2 Experimental Setup

The LIBSVM is integrated software for SVM in Visual Studio was used for the classifying the text sample and presents the experiment analysis on article review available in MS SQL server. The proposed work is divided into three experiments [8].

Experiment 1: In this experiment evaluates the described parameters using 50 reviews and as compare the positive and negative polarity it indicates that reviews have more positivity than negativity so the article is excellent. The article got 62% positivity and 8% negativity.

Experiment 2: In this experiment evaluates the described parameters using 100 reviews and positivity is greater than negativity. Small amount of review has negative polarity and article got 70% positivity and 5% negativity in this case.

Experiment 3: Evaluate the described parameters using 120 reviews of article. The article got 68.33% positivity and 5% negativity.

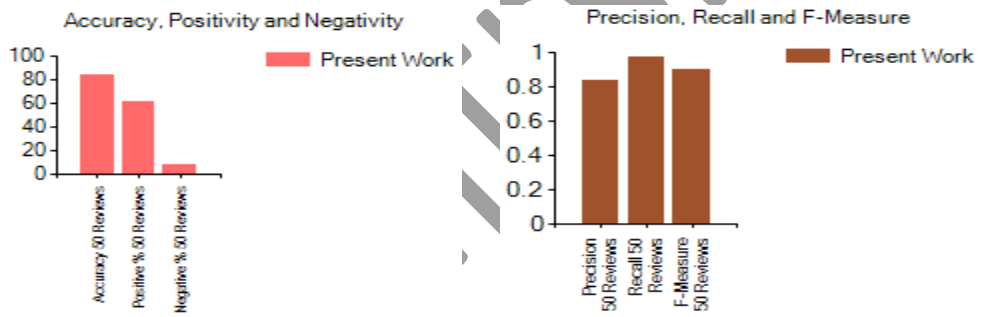

Fig. 4.1 Plot of 50 article reviews

The Fig. 4.1 shows the two graphs in which first graph represents the accuracy, positivity and negativity in case of using 50 reviews. The positivity comes in the 50 reviews is much greater than the negativity. The second graph represents the Precision, Recall and F-Measure in case of 50 reviews and value of Recall is greater than Precision and F-Measure. So positivity comes in case of 50 reviews also prove that the hybrid approach is effective to classify the reviews of article.

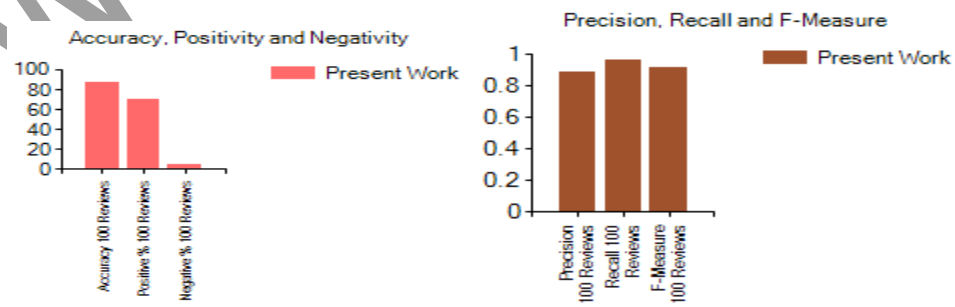

Fig. 4.2 Plot of 100 Article Reviews

The above figure indicates that the result evaluation using 100 reviews. The result produces using 100 reviews greater than using 50 reviews in terms of Precision, Accuracy, F-Measure and Positivity. So result evaluation using 100 reviews quite effective than the result evaluation using 50 reviews.

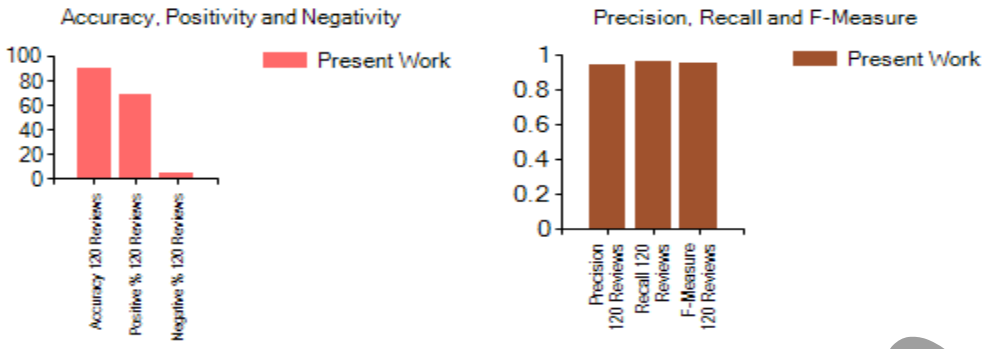

Fig. 4.3 Plot of 120 article reviews.

The above figure represents result evaluations using 120 reviews. The Precision, Recall and F-Measure in 120 reviews are at almost same level. The accuracy achieved is better than the previous two cases. So, the hybrid approach is more effective in this case mainly in terms of Accuracy.

Table-4.2Result Table for all Experiments

	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

4.3 Comparative Analysis

In comparative analysis, compare the output achieved using the proposed hybrid approach with the output obtained by existing approach. To compare the result, both proposed hybrid approach and lexicon approach used the similar polarity dataset which consists of 120 reviews. The following Table 4.3 indicates the comparison of the proposed approach with the lexicon approach [7].

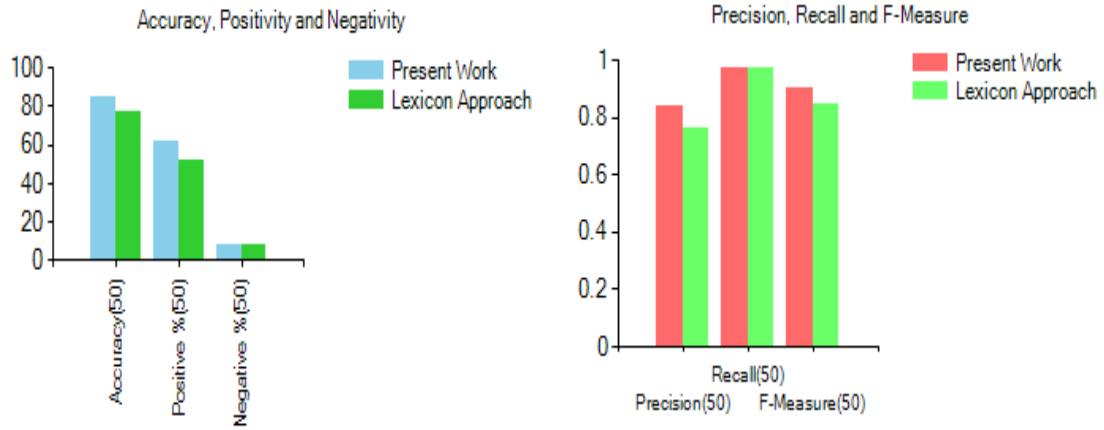

Fig. 4.4 Comparison of the results of two approaches using 50 reviews

The above figure shows the comparison of proposed approach with lexicon approach in evaluation of all the parameters. . Positivity comes in proposed approach is greater than lexicon approach. The proposed method is working well in terms of Accuracy, Precision, F-Measure and Positivity.

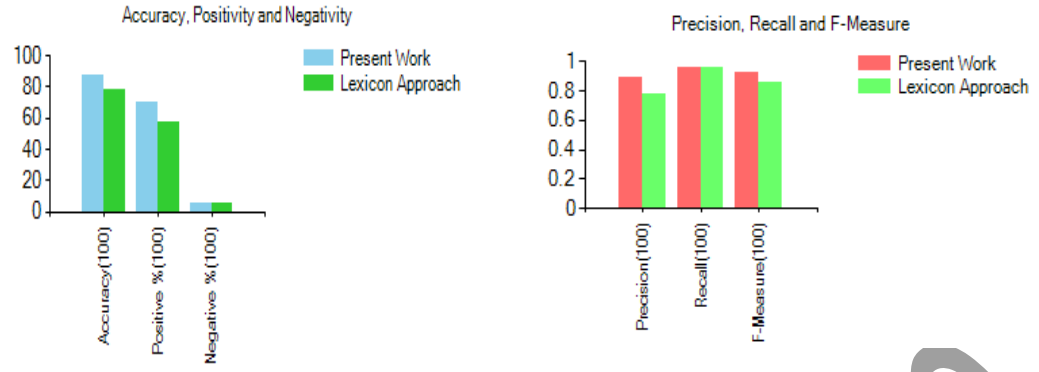

Fig. 4.5 Comparison of the results of two approaches using 100 reviews

The fig. 4.5 represents the comparison of the results of proposed approach and lexicon approach for all parameters in case of 100 reviews. In first graph accuracy achieved is 87.13% that is higher than the lexicon approach. The second graph also shows that proposed hybrid approach is best than the lexicon approach.

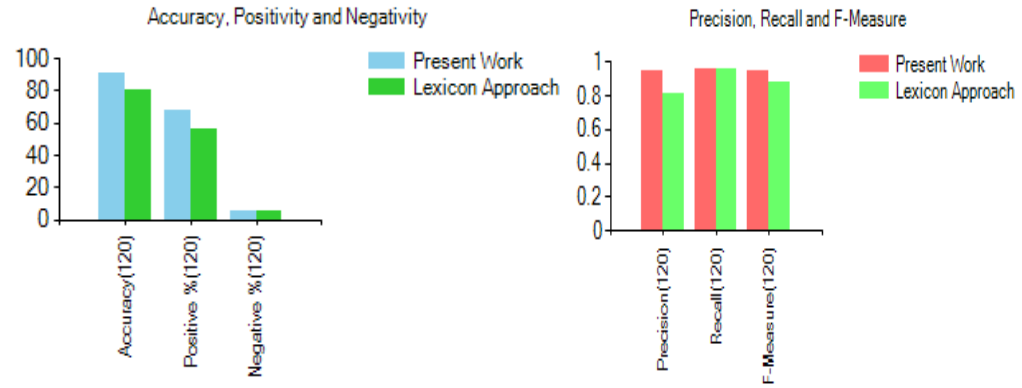

Fig. 4.6 Comparison of the results of two approaches using 120 reviews.

The fig, 4.6 shows that result evaluation of both approaches in case of 120 reviews. In first graph the accuracy obtained by proposed hybrid method is 90.74% that is higher than all the cases. So from the results it is prove that the proposed approach is more efficient than the lexicon approach.

Table -4.3 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

From the Table 4.3, it is represented that the accuracy computed in the case of the present method is better as compared to lexicon approach. While using the same dataset size, the positive polarity in proposed approach is better than the lexicon approach.

The comparison shows that lexicon approach worse than the proposed hybrid approach. In lexicon based approach, performance is relying upon the sentiment or opinion words that are enclosed within the dictionary. If the dictionary contains fewer words, it results in a decrease in performance. The hybrid approach combines the advantage of both the Support Vector Machine and K- Nearest Neighbor techniques. It is inheriting more accuracy using supervised machine learning approaches and providing good stability against the lexicon approach [6].

CONCLUSION AND FUTURE WORK

The paper considered the combination of two supervised machine learning techniques to technical article review data and also predicted the positive and negative reviews by people on the SuperFetch. The hybrid approach which contains the combination of SVM and KNN produced better results on the basis of Accuracy, Precision, Recall and F- Measure. K- Nearest Neighbor approach improved the performance in the case of small reviews and Support Vector Machine improved the performance just in case of large reviews are working as a single hybrid approach.

There are two more parameters positivity and negativity are computed that shows most of the reviewers have positive thoughts concerning SuperFetch and the negativity percentage is very less, the remaining reviews are considered as neutral. So the results show that SuperFetch is a good feature in the memory management system. The future work includes analyzing and improves the performance of present work 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] Nagamma P*, Pruthvi H.R, Nisha K.K^{+}, and Shwetha N H, "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.

[2] R. Prabowo, and M. Thelwall, "Sentiment analysis: A combined approach," Journal of Informetrics, pp. 143-157.

[3] V. S. Jagtap, and K. Pawar, "Analysis of different approaches to Sentence-Level Sentiment Classification," International Journal of Scientific Engineering and Technology, vol. 2, pp. 164-170, April 2013.

[4] S. Tan, and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," Expert Systems with Applications 34 (2008), pp. 2622-2629.

[5] Z. Elberrichi, A. Rahmoun and M. A. Bentaalah, "Using WordNet for Text Categorization," The International Arab Journal of Information Technology, vol. 5, no. 1, pp.16-24, Jan. 2008.

[6] H. Rahmath, and T. Ahmad, "Sentiment Analysis Techniques- A comparative study," International Journal of Computational Engineering and Management, vol. 17, issue 4, pp. 25-29, July 2014.

[7] A. Tripathy, A. Agrawal, and S. K. Rath, Classification of Sentimental Reviews using Machine Learning Techniques, in Proceeding of 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), pp. 821-829.

[8] R. S. Rahate, E. M, "Feature Selection for Sentiment Analysis by using SVM," International Journal of Computer Applications, vol. 84, no. 5, pp. 24-32, Dec. 2013.