I.J. Education and Management Engineering, **2016, 6, 1-11**
Published Online November 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2016.06.01

Available online at http://www.mecs-press.net/ijeme

A Hybrid Approach to Sentiment Analysis of Technical Article Reviews

Babaljeet Kaur [a], Naveen Kumari [b]

[a] *Punjabi University Regional Centre, Mohali, India*
[b] *P unjabi University Regional Centre, Mohali, India*

Abstract

Sentiment analysis is similar to opinion mining, which is a popular research problem to search out in the field of NLP. Sentiment analysis determines the perspective of the author and identifies the positive, negative and neutral reviews. It provides the reviews or opinions of people's on text, article and product which can be positive, negative or neutral. Reviews on the different websites, social networking sites is an important source to collect the information regarding various brands of product and new features in technology (e.g. Windows, Mobiles). During the sentiment analysis various classification tools within the NLP are used to find out the positivity and negativity of reviews or comments. The paper presents a length aware hybrid approach to analyses the reviews either as positive or negative and present approach is tested on SuperFetch data set. The present approach is a combination of both supervised machine learning techniques that are Support Vector Machine and K-Nearest Neighbor in which SVM is working great for large size review and KNN is working best for small size review.

Index Terms: Sentiment analysis, SVM, KNN, SuperFetch review.

1. Introduction

With the rapid growth of online websites and social networking sites, everybody express their views regarding article, product on these sites [12]. The internet or web becomes a necessity in every person's life. On each day everybody is using online websites to express their reviews or comments, providing feedback and asking a question and other business or companies can take decision according to feedback. Sentiment analysis depicts the mood or feeling of individual about various entities and their attributes [11]. Sentiment analysis is a task to express people's opinions about an entity, article and product etc.

In natural language processing (NLP) different machine learning approaches are utilized to determine the

* Corresponding author:
E-mail address: babbaljeet001@gmail.com, naveencse2k4@gmail.com

sentiments of a huge amount of products, services, and text etc. Sentiment analysis provides many facilities to implement new applications which have a major role in business, medical, data mining and evaluate the feedback of different latest and old products.

The reviews or comments of individuals in sentiment analysis can be classified into various ways like positive and negative reviews. Positive reviews have high polarity than the negative reviews. According to the reviews by various experts, one can easily find out the quality of technique. The analysis of sentiment work up with few difficulties: Among different things, it must be resolved whether document or segment thereof is subjective or objective and whether or not the sentiment communicated is positive or negative.

The main steps involved in the research are Pre-processing of text, analysis of data and classification and evaluation of result. On numerous websites, people share their views in the form of 'comments'. These reviews and comments are the main elements that confirm the sentiment of people as these reviews as original. The sentiment analysis on the website is done in three steps: (a) Determine the sentiment expression. (b) Determining the polarity of reviews (positive, negative and neutral). (c) A classification approach is utilized to classify the sentiments on the website [2]. It has increased much consideration as of late and studies individual's emotions towards certain substances [1]. Sentiment analysis deal with several challenges. The description of element considered in this area. It is not a cup of tea. For example sentiments about any product can be determined by positive and negative opinions of people on the product.

In the present work find out the reviews of a specialized article (e.g. SuperFetch). Determine the positive and negative survey of individuals in this article. On the basis of the reviews of the expert define how much that article technically sound or not. The positive review concerning article gives high polarity. A combined approach of Support Vector Machine and K-Nearest Neighbor is used to classify the reviews. The rest of the paper is organized as follows: Section 2 describes the some related work done in this field. Section 3 explains the proposed approach in identifying the reviews. Section 4 provides the results obtained in the proposed work along with a brief discussion. Section 5 concludes work done in this research and work to be done in the future.

1.1. Applications

- Aid in decision making- Decision making is an integral part of our life. It ranges from "which technology's feature to use", "which product to buy", "which bank insurance policies to go for", "which restaurant to go". Sentiment Analysis (SA) can be utilized to decide and select from the available options on the basis of general opinions expressed by other users.
- Business Strategies- Many of business strategies are being guided with regard to response from the public. The various companies' aims to satisfy the requirements and demands of the users, so strategic moves of the companies are driven through public views and opinions. With the world connected via technology events have a global impact; issue/failure on one part of the globe has an impact on opposite corners of the world. Therefore it becomes quite necessary to drive products/services per the general public view point.
- Application as sub-component technology- The sentiment analysis has a major role in enabling technology for different systems. In some websites whenever ads are displayed in the sidebars, it's helpful to show ads when relevant positive sentiments are detected and nix the ads once negative sentiments are detected. The question-answering is another area where sentiment analysis may be helpful as opinion oriented question might need completely different treatment [13].
- Application to review related websites- The most common application in within the reviews of costumer products, services, and articles. There are several websites that give reviews related to the article. For example os.news.com is the website that provides numerous reviews related to mobile phones, Windows, DOS, and Memory management. Summarizing opinions or reviews of the public are an important task today, so before using the latest features comes in the technology, people can aware to make decision regarding the technology means good or not for future [10].

2. Related Work

This section presents a comprehensive literature survey of research related to various sentiment analysis classification approaches. Various sentiment analysis related research papers till date have been studied and their brief is presented in this section.

Mudinas *et al*. The paper presented a combination of both lexicon and learning based approaches for sentiment analysis. The paper mainly focused on the anatomy of pSenti- a concept level sentiment analysis and combined approach are tested on two types of data set including CNET software reviews and IMDB movie reviews. In the concept level sentiment analysis, pSenti is originated by combining both lexicon and learning based techniques. The supervised machine learning part is not simply responsible for smaller tasks such as adjusting sentiment values or finding additional sentiment words, however is really responsible for evaluating all the ingredients of sentiment process as well as semantic rules utilized to get the final output. The main advantage of hybrid approach employing a lexicon/machine learning techniques is to achieve more effective of each word- stability also as readability from a correctly designed lexicon, and also the high accuracy of the popular supervised machine learning approach. The of performance of sentiment analysis mainly depends on the sentiment of comments or reviews; if there's a clear separation between positive and negative value distribution, the lexicon based technique would work best, otherwise machine learning would well raise the performance. The result shows that even without subjective detection, the combined approach pSenti can achieve 82.50% accuracy that is just slightly below the bag-of-words Support Vector Machine [8].

Amira Shoukry *et al*. The paper presented a sentiment analysis of Egyptian dialect tweets and applied sentence level analysis utilizing a hybrid approach. This approach combines each machine learning (ML) approach using Support Vector Machine (SVM) and semantic orientation (SO) techniques. This approach includes building a classifier utilizing the unigrams, bigrams and trigrams as well as new feature for the semantic orientation score that sums the weights of total sentiment words and smiley faces (e.g. ) also available in the tweets. Just in case of machine learning accuracy by utilizing thresholds for N-gram (unigrams, bigrams and trigrams), Unigrams produces the best results at threshold 0. It is good to combine the all unigrams, bigrams and trigrams as features to increase the performance of sentiment analysis. The accuracy achieved by Support Vector Machine (SVM) classifier is 82.9% in case of unigrams. When comparing the results of both obtained in ML experiment and semantic orientation, the accuracy (0.806) obtained in SVM learning algorithm greater than achieved using semantic orientation technique's accuracy (0.719). After combining both ML and SO accuracy (A) is 0.844 and precision (P) is 0.842, recall (R) is 0.844 and f-measure is 0.842 [3].

Changliang Li *et al***.** The paper introduced a Chinese feeling Treebank over social data because there are not very many resources introduced in sentiment analysis for Chinese and progress is kept down because of expansive, marked corpus and capable models. It identified 13550 marked sentences from movie reviews. The paper introduced a Recursive Neural Deep model (RNDM) figuring out how to sentiment label based on recursive deep learning. The paper considers the sentiment about the sentence and identifying review is positive and negative. For sentence level based sentiment, this model used baselines such as Naïve Bayes, Maximum Entropy and Support Vector Machine. The total accuracy obtained by RNDM is maximum (90.8%) compared to Naïve Bayes (78.65%), Maximum Entropy (87.46%) and Support Vector Machine (84.9%) and RNDM gets the highest performance [7].

Grandi *et al*. In this paper 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 complete profile [6].

Tan *et al*. The paper developed sentiment categorization on the basis of 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 achieved suggest that IG performance best in sentiment phrases selection and Support Vector Machine (SVM) for sentiment classification and it contains the reviews from three different areas like movie, education and house, also found that sentiment classifier mainly dependent on the topics. The dataset contains 507 documents related to education, 248 to house and 266 documents related to the movie. The Precision, Recall and F-Measure parameters are used to evaluate the performance of all the classifiers [4].

Guerrero *et al.* The paper analyzed people's emotions, reviews, feelings about the product, services and organizations. Different tools and techniques are implemented for sentiment analysis. Compared some free access web services, analyze their capabilities and score to different pieces of text according to their sentiments. Machine learning and lexicon based approaches are utilized. 15 web services computed for the purpose of sentiment analysis. Some of these services belong to personal companies, however, still they allow to restricted free access to their functionalities and others are completely free services. This fact is interesting to the users who want to incorporate sentiment analysis capability among their own platforms while not having to develop their own algorithms; thus, these tools are particularly interesting for researching purposes and fast prototyping. Besides, because of the actual fact that the chosen services can work as web services, inclusion of them into any platform is basically simple. From the results produced, the services like Alchemy and Semantria might be taken into consideration for any kind of text. Sentiment analysis could also be extremely interesting to the user if the analyzed text is quite high and you would like to classify them [5].

3. Methodology

This section provides the information about the present hybrid approach and various parameters that are used to evaluate the results and also the various steps that are taken into consideration to complete the proposed work.

3.1. Data set

The data set is a collection of technical article reviews (e.g. SuperFetch), which is a memory management feature find out in various Windows. The reviews of the SuperFetch have collected from the websites such as anandtech.com, os.news.com and also from the various professionals of Universities or Colleges. These websites provide detailed information related to positive and negative reviews of different articles given by various professionals. The collected reviews are stored in an excel file, then directly fetch to the MS SQL server for further processing. All the collected reviews are divided into three parts like 50, 100 and 120 to evaluate the result.

3.2. Preprocessing

The data preprocessing or cleaning step is more important in sentiment analysis. It is the process of preparing the text before the classification. The preprocessing includes the removal of extra words which are not helpful for identification of sentiments. In case of word level, different words in sentence don't have any impact if treated as one dimensional in the classification process [9]. Including those words in the text make the classification more tough. So, there is a process to overcome the noise level in text should help increase the performance of the classifier is called preprocessing. The main step in this process is the removal of stop words is described below. The preprocessing step mainly includes the removal of stop words. In Information Retrieval, stop words removal is a common process to get rid of words that are extremely common which don't add substantial value to the classification process. These common words like is, he, it, a, an, and they're

collectively known as stop words. Because the inclusion of these words in a review does not give any useful information, they are removed.

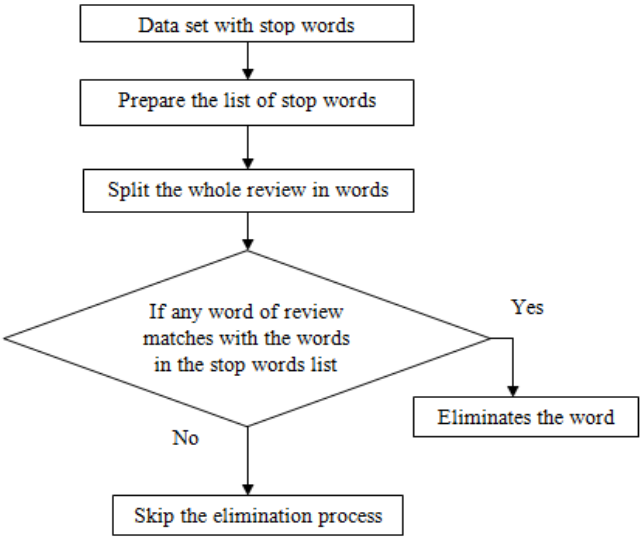

Fig.1. Flow Chart of Removal of Stop Words

3.3. Proposed Approach

The proposed hybrid approach is a combination of supervised machine learning approaches: SVM and KNN. The present hybrid approach is utilized to classify the reviews of SuperFetch in which Support Vector Machine classifier is working well to evaluate the large size reviews and KNN is working best to evaluate the small size review. Both approaches are combined in such a way that to increase the effectiveness of the present hybrid approach.

- Support Vector Machine

Support Vector Machine is well known for their good generalization performance and has been applied to many sentiment analysis problems. Training and testing data are involved in classification task, which consists of data instances. One class label and various features are contained in each instance in the training set. In basic form, a SVM learns to find a hyper plane that separates both positive and negative examples maximum margins. The Support Vector Machine classifier is most effective classifier to increase the performance using large size sentence during the classification.

- K-Nearest Neighbor

It is the classifier that relies on the labels of category and finds a k-nearest neighbor during the classification. Assign the weight to the neighbor based on their distance from the query point. KNN is an effective classifier to classify the small size sentence. The Euclidean distance equation is a main alternative to the similarity measure.

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

3.4. Performance Evaluation Metrics

Finally the result is evaluated using the Accuracy, F-Measure, Precision and Recall. These four parameters are used to evaluate the performance of the classifiers.

4. Results and Discussion

The SuperFetch Reviews is considered as data set to test the performance of the present hybrid approach. The reviews are collected from the websites as well as from the various professionals. The present hybrid approach is tested on a SuperFetch data set in which three cases are considered to evaluate the results. All cases were carried out in Visual Studio 2010 are discussed in this section.

Reviews are classified as positive and negative by the hybrid method. The effectiveness of the hybrid method is determined by the following parameters.

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (2)

Where;

TP is true positive
TN is true negative
FP is false positive
FN is false negative

- Precision = $\frac{TP}{TP+FP}$ (3)
- Recall = $\frac{TP}{TP+FN}$ (4)
- F-Measure = $\frac{2\times Precision\times Recall}{Precision\pm Recall}$ (5)

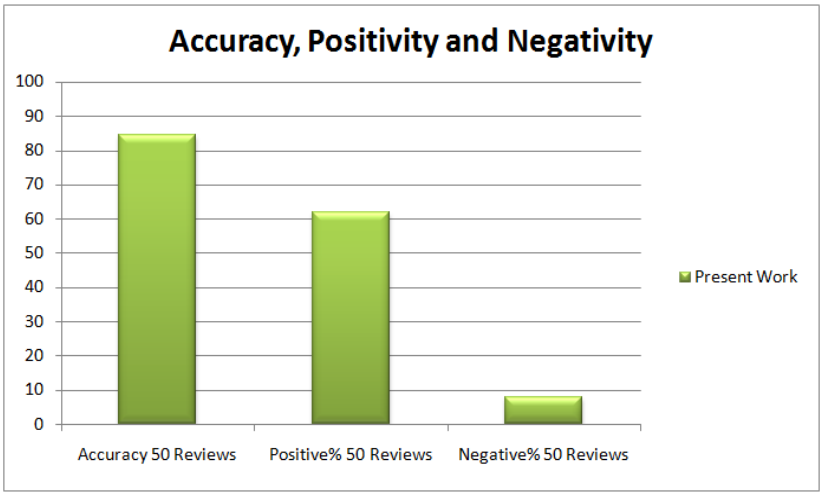

Fig.2. Accuracy, Positivity and Negativity in case of 50 Reviews

The Figure 2 shows the graphs which represent the accuracy, positivity and negativity in case of using 50 reviews. The positivity comes in the 50 reviews is much greater than the negativity.

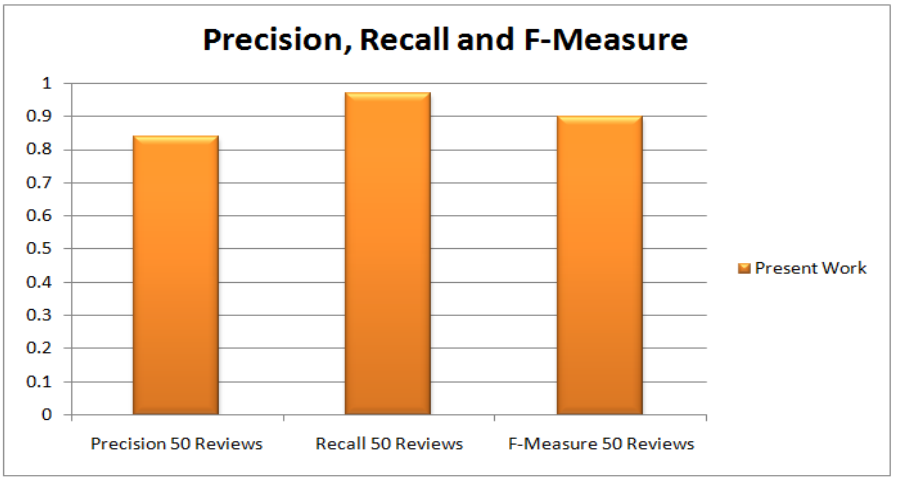

Fig.3. Precision, Recall and F-Measure in Case of 50 Reviews

The figure 3 indicates the Precision, Recall and F-Measure in case of 50 reviews. The value of Recall is greater than the Precision and F-Measure.

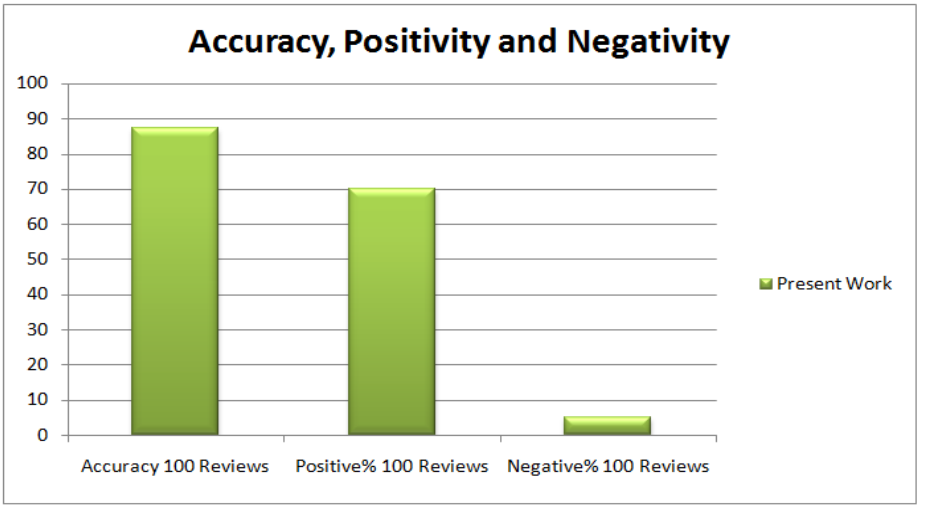

Fig.4. Accuracy, Positivity and Negativity in Case of 100 Reviews

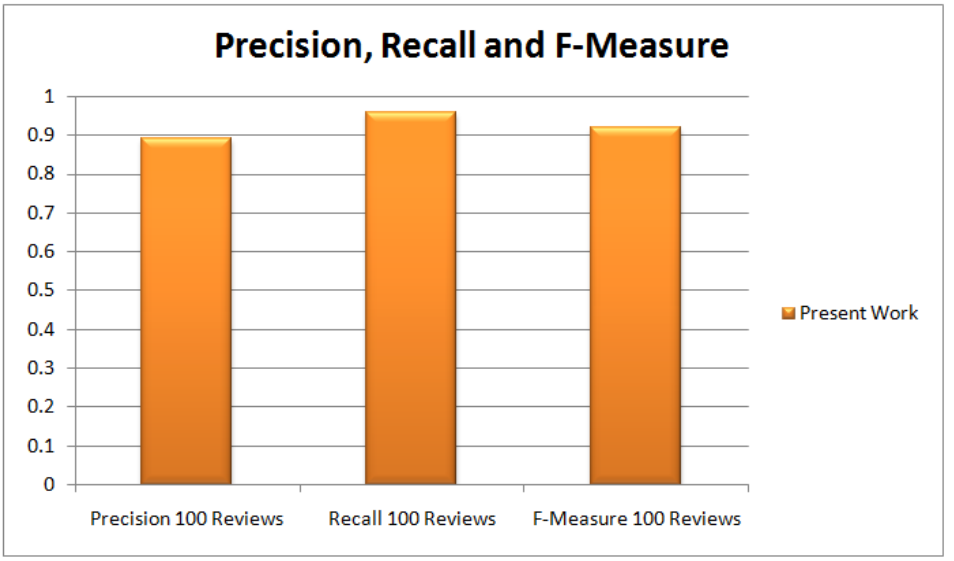

Fig.5. Precision, Recall and F-Measure in Case of 100 Reviews

The figure 5 and 6 show that the result evaluation using 100 reviews. The results achieved using 100 reviews greater than using 50 reviews in Precision, Accuracy, F-Measure and Positivity. So result evaluation using 100 reviews quite effective than the result evaluation using 50 reviews.

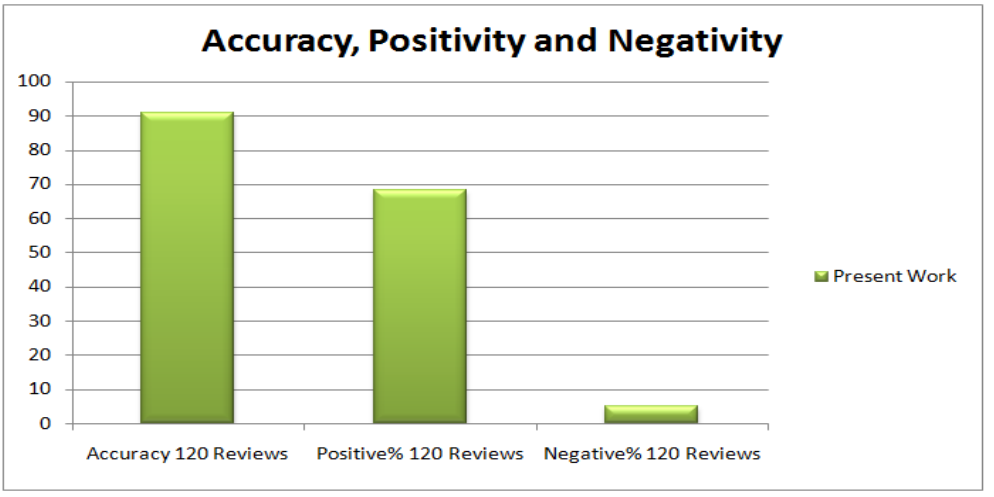

Fig.6. Accuracy, Positivity and Negativity in Case of 120 Reviews

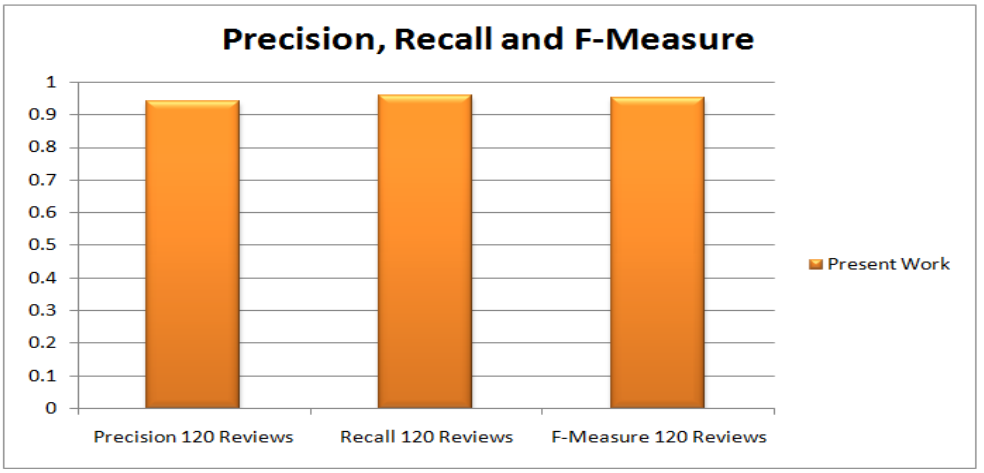

Fig.7. Precision, Recall and F-Measure in Case of 120 Reviews

The figures 7 and 8 represent 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 1. Result Table for All Cases

Parameters	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

5. Conclusion and Future Work

Present work concluded that the combination of K-Nearest Neighbor and Support Vector Machine produced better results on the basis of Accuracy, Precision, Recall and F-Measure. K-Nearest Neighbor improved the performance in the case of small reviews and Support Vector Machine improved the performance in case of large reviews are working as a single hybrid approach. There are two more parameters positivity and negativity are evaluated that shows most of the reviewers have positive thoughts regarding SuperFetch and the negativity percentage is very less, the remaining reviews are considered as neutral. So the results indicate that SuperFetch is a good feature in the memory management system. Future work includes the comparison of the present technique with existing techniques.

Acknowledgements

I am grateful to my guide Assistant Professor Mrs. Naveen Kumari for all help and valuable suggestion provided by her during the study and special thanks to Associate Professor Mr. Michael Swift who help me in collection of reviews.

References

[1] X. Fang, and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, pp.1-14, 2015.

[2] Y. Sharma, V. Mangat, and M. Kaur, "Sentiment analysis and Opinion mining*," In Proceedings of 21st IRF International conference*, March-8, 2015, Pune, India, pp. 35-38.

[3] 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.

[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] J. S. Guerrero, J. A. Olivas, F. P. Romero, and E. H. Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Information Sciences,* pp. 18-38, 2015.

[6] 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.

[7] C. Li, B. Xu, G. Wu, S. He, G. Tian, and H. Hao, "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.

[8] A. Mudinas, D.Zhang, and M. Levene, "Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis," *In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM*, August-12, 2012, pp. 1-8.

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

[10] Y. Sharma, V. Mangat, and M. Kaur, "A Practical Approach to Sentiment Analysis of Hindi Tweets*," In Proceedings of 1st International Conference on Next Generation Computing Technologies (NGCT-2015*), Dehradun, India, 4-5 September 2015, IEEE, DOI: 10.1109/NGCT.2015.7375207, pp. 677-680.

[11] J. 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.

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

[13] B. Pang, and L. Lee, "Opinion mining and Sentiment analysis," *Foundation and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

Authors' Profiles

Babaljeet Kaur has completed her M.Tech in Computer Science and 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,"A Hybrid Approach to Sentiment Analysis of Technical Article Reviews", International Journal of Education and Management Engineering(IJEME), Vol.6, No.6, pp.1-11, 2016.DOI: 10.5815/ijeme.2016.06.01