IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. III (Jul.-Aug. 2017), PP 80-89
www.iosrjournals.org

To Improve Accuracy in Movies Reviews Using Sentiment Analysis

*Rasika Wankhede[1], Prof. A. N. Thakare[2]
[1](Dept of CSE, BDCE Sevagram, RTMNU, Maharashtra, India)
[2](Dept of CSE, BDCE Sevagram, RTMNU, Maharashtra, India)
Corresponding Author: Rasika Wankhede*

Abstract: *Opinion mining is one of the new concepts of data mining. As World Wide Web is growing at higher rate, this has resulted in enormous increase in online communications. The online communication data consist of feedback, comments and reviews on particular topic that are posted on internet by internet users. Sentiment analysis is a sub-domain of opinion mining where the analysis is focused on the extraction of emotions, specific view or judgment on certain topic. Sentiment analysis system classifies text data into their respective sentiments based on polarity. In this domain most of the previous researchers have focused on using one of the three classifiers like SVM, Naïve Bayes, and Maximum Entropy. There are some other robust classifiers which have ability to provide comparable or better results. In this work, we try to focus our task of sentimental analysis on four different sites such as bollymoviereviewz.com, rottentomatoes.com, timesofindia.com and filmiclub.com movie review databases. The proposed approach is based on SentiWordNet, which generates count of scores words into five categories like strong positive, weak positive, neutral, strong negative and weak negative for the opinion mining task and evaluated using Random Forest algorithm. By using Random Forest classification technique we have achieved the best accuracy of 94.14%.*

Keywords: *Classifier, Feature extraction, Movie reviews, Opinion mining, Polarity, Sentiment Analysis.*

Date of Submission: 15-07-2017 Date of acceptance: 27-07-2017

I. Introduction

Sentimental analysis is rapidly increasing research area in the field of text mining. Posting online reviews on different web sites has become an increasingly popular way for people to share their opinions about specific product or services with other users. Sentimental analysis or opinion mining is the computational study of people's judgment, attitudes and emotions towards an entity [1]. The entity can represent individuals, events or certain topics. Opinion mining extracts and analyses people's opinion about an entity while sentiment analysis finds the sentiment words expressed in a text document and then analyses it. Therefore, the main goal of sentiment analysis is to find opinions, identify the sentiments they express, and then classify their polarity. Sentimental analysis helps to find words that indicate sentiments and help to understand the relationship between textual reviews and the significance of those reviews. One such domain of reviews is the domain of movie reviews which affects everyone from audience, film directors to the production company [1]. The movie reviews present on various websites are not formal reviews but they are rather very informal reviews and are unstructured form of grammar.

Sentiment analysis of twitter messages has gained significance attention over the past few years. With the help of opinion mining, we can differentiate poor contents from high quality contents. It is possible that by using available technologies we can even know if a movie has good opinion then bad opinion and this helps users in their decision making. In this paper reviews from the four different movie websites are collected. We follow a lexical approach using the SentiWordNet for finding the overall polarity of the movie reviews [5]. We analyze the features that affect the sentiment score of the movie reviews.

On the basis of entered reviews by user it produces the result according to highest sentiments extracted based on polarity. We have also focused on finding the exact polarity result for phrases such as "NOT VERY BAD", "NOT SO INTERESTING" etc. We have used Natural Language Processing (NLP) for detecting the part-of-speech such as adjective, noun, verb etc. For example phrase such as "NOT VERY BAD", in this example "VERY BAD" is an adjective and NOT represents the negation. We find sense of each sentence using machine learning technique, with the help of neural network the system gets trained for detecting the sentiments from the reviews. According to the machine the phrase such as "VERY BAD" will be negative and it will score the sentence as negative. The "NOT" word also indicate the negation. So, this phrase is considered as negative phrase, but if we find sense of this phrase this is not negative phrase and it is incorrectly classified as negative phrase as machine is unable to handle the negation properly. Contextual understanding is difficult for a machine

to reach human level accuracy. To solve this problem we have used five feature labels such as Strong Negative- (-2), Weak Negative- (-1), Neutral- 0, Weak Positive- 1 and Strong Positive- 2 with the help of these feature labels we can find the exact polarity result for phrases which are neither completely positive nor completely negative phrase, such phrases are classified into weak positive or weak negative polarity based on their scores.

Phrases such as "Worst Movie" will be classified as strong negative phrase and "Very inspirational story really a good movie" will be classified as strong positive sentences. The sentence which does not contain any positive or negative sentiments is neutral. For this the SentiWordNet is used which produces the count of feature scores and Random Forest classifier, classifies them into five features such as strong positive, strong negative, neutral, weak positive, weak negative. These score counts are used to perform the task of sentiment analysis [5]. All these five features are represented in graphical form and the average score of these features are represented using emoticons for better visualization of users. Organization of this paper provides the following details: Section II discusses the related work done in this domain. Section III describes the issues in existing system. Section IV explains the proposed work done in this paper in depth. Section V discusses the experiment and result analysis. Section VI gives the conclusion for the proposed work.

II. Related Work

A large number of works have been carried out previously on opinion mining and sentiment analysis. Nagamma P et al. [1] proposed different data mining techniques for classification of movie reviews and it also predicts the box office collection for the movie. Classification accuracy for pretending was improved substantially by clustering method. The online movie review data collected from IMDB dataset, the box office collection and the success or failure of the movie is predicted based on the reviews. Pang et al. [2] applied the machine learning technique for classification of reviews present on IMDB movie reviews database according to sentiment. They have used Naïve Bayes, Maximum Entropy and SVM classifiers for sentiment classification task. However, SVM classifier achieves higher accuracy than Naïve Bayes and Maximum Entropy. They noticed that using unigrams as features with term presence on SVM always gain highly accurate results, but using bigrams as features, the accuracy was lower as compared to that of the unigrams. They also found that the machine learning techniques outperformed human produced baselines. J. Erman et al. [3] studied three types of clustering algorithms namely K-Means, DBSCAN and AutoClass algorithm for the classification of network traffic problem. This study is based on the ability of each algorithm for forming clusters having higher predictive power of a single traffic class and for determining the ability of each algorithm to generate small number of clusters that has many connections. The AutoClass algorithm is compared with DBSCAN and K-Means algorithm and the result indicates that both K-Means and DBSCAN work faster than AutoClass algorithm. Turney et al. [4] presented an unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or as not recommended (thumbs down).This algorithm calculated the point wise mutual information (PMI) of the candidate word for semantic orientation with two given key words such as, ''poor'' and ''excellent''. The algorithm depended on the patterns of two consecutive words where one word is an adverb or adjective used for orientation and the other word is used to represent the context. Adverbs and adjectives with different patterns of term categories were used for the semantic orientation, and a review was classified as recommended if the average semantic orientation of its phrases were positive and as not recommended if the average semantic orientation of its phrases were negative.

Stefano, Andrea and Fabrizio [5] present SentiWordNet 3.0, it is a lexical resource used for sentiment classification. SentiWordNet 3.0 is an open resource platform for all researchers all over the world, for different types of research projects it has supported more than 300 research groups worldwide. SentiWordNet assigns to each word of three sentiment score such as positivity, negativity and objectivity. Rudy Prabowo et al. [6] studied the hybrid SVM classification method for sentiment classification. They used Sentiment Analysis Tool for achieving good level of effectiveness. Ion Smeureanu et al. [7] presents a method of sentiment analysis based on movie reviews. They have used Naïve Bayes technique for classifying reviews into positive and negative class. Kennedy et al. [8] studied two methods for determining the sentiments present in movie reviews. First method classifies reviews into positive and negative polarity and second method uses positive and negative terms from different sources such as dictionary of synonym and web corpus.

One of the latest works on feature level analysis of opinions was reported by Zhai et al. [9] studied a semi-supervised technique for feature grouping as this technique plays an important role in the summarization of opinions. As the same features can be expressed by different synonyms, words or phrases, for producing a useful summary, these words and phrases were grouped. With respect to feature grouping, the process generated an initial list to bootstrap the process using lexical characteristics of terms. This method has empirically achieved good results.

Dave et al. [10] presents a model for review classification based on features for various machine learning techniques and classification. Their approach depends on a manually annotated corpus whereby each of the annotated corpuses is described by features related to positivity and negativity polarity. The test document is

classified through an annotated corpus by using similarity scores. The classifier depends on information retrieval techniques for feature extraction and scoring. As such, this paper proposed that a group of sentences such as documents or a full review can provide a more reliable analysis than an individual sentence as a sentence based performance analysis is limited due to noise and ambiguity.

III. Issues

There are three main issues in existing systems which are overcome in the present system. The three main issues are as follows

1) Many classifiers used in previous system do not give much accuracy. In this paper we have used Random Forest classifier which provides better accuracy than other machine learning algorithms.
2) Inadequate reviews that leads to wastage of time and money. This issue is overcome in present system by the process of pre-processing. The exact reviews are provided to the users in the form of graphs based on polarity result which is easily understood by the user and it saves time as well as money of users.
3) In existing system, there are certain situations where it becomes difficult to decide for a machine that which phrase is either completely positive or completely negative for example phrase such as "NOT VERY BAD". This problem is solved by evaluating the sentiment scores; we have used five feature labels such as strong positive, strong negative, weak positive, weak negative and neutral features by which the exact result based on polarity is obtained.

IV. Proposed Work

This section gives the description of the steps followed for the movie dataset mining for sentiment analysis. In this work we have focused on two areas like first Feature Selection and second using machine learning techniques. We use "timesofindia", "bollymoviereviewz", "rotten tomatoes", "filmiclub" movie review dataset and we provide label to the polarity as follow Strong Negative- (-2), Weak Negative- (-1), Neutral- 0, Weak Positive- 1, Strong Positive- 2. The flowchart as shown in Fig.2 explains the overall methodology.

4.1 Input Data

The input data is in the form of reviews from the "times of india.com", "bollymovie reviewz.com", "rottentomatoes.com","filmiclub.com" movie review datasets. Particular movie is selected from the dataset and reviews regarding that movie are displayed on web page. After releasing of any new movie the reviews of that movie are added to the dataset.

4.2 Pre-Processing

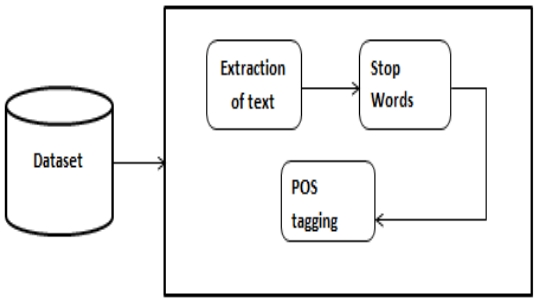
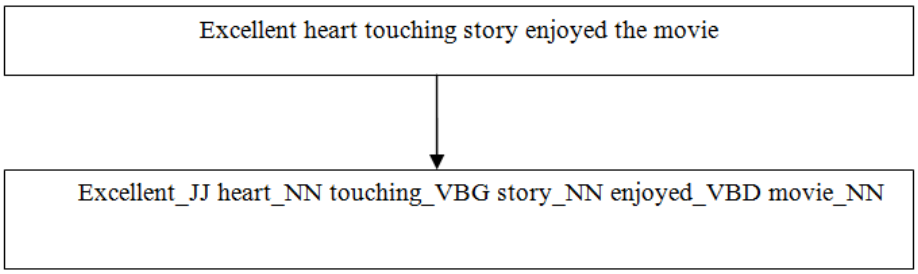


Fig.1: Pre-processing of dataset.

The text pre-processing techniques are divided into three subcategories:

- Tokenization: The data present in the text document contains block of characters called tokens. These text documents are separated as tokens and used for further processing of data.
- Removal of Stop Words: A web search tool or other natural language processing system may contain collection of stop-records, or it may contain a solitary stop-list. Most of the more frequently used stop words in English are "an", "a", "of", "the", "you", "and" these are some words which do not carry any meaning. Hence, those words which appear too often that support no information for the task are removed.
- Part of Speech Tagging: POS tagger parses a sentence or document and tags each term with its part of speech. For part-of-speech tagging we used the Stanford part-of-speech tagger. This tagger used by splitting text data into sentences and to produce the POS tag for each word (whether the word is a noun, verb, adjective). Consider following example



In part-of-speech (POS tagging), each word in review is tagged with POS (such as noun NN, adjective JJ, verb RB). In tagged sentence, Excellent is tagged with tag JJ which indicates 'Excellent' is an adjective where as a 'heart', 'movie', 'story' are tagged as NN which indicates noun and 'touching', 'enjoyed' are tagged as verbs.

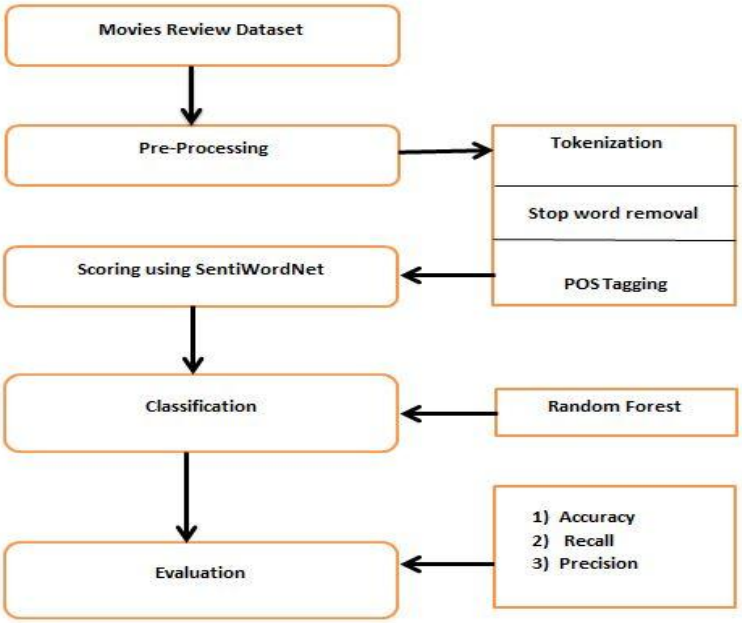


Fig.2: Work Flow Diagram.

4.3 Sentiment scoring and classification

Many approaches are mainly classified into two categories namely lexicon based approach and machine learning based approach. We have used lexicon based approach using SentiWordNet for finding the overall polarity of movie reviews. We use well known classifier namely Random Forest classifier, for sentiment classification. The classification is done with the Random Forest classifier to determine the class labels for a machine and to predict the class of a movie reviews whenever it arrives in the form of five polarities. We have performed the feature impact analysis by computing the information gain for each feature in the feature set and used it to derive a reduced feature set. The reduced features are provided as input to classification process and the classification is based on the five features such as strong positive, strong negative, neutral, weak positive and weak negative.

- Scoring using SentiWordNet

SentiWordNet is a publicaly available lexical resource, where each synset is associated to two numerical scores such as positive synset and negative synset, describing how positive and negative terms are present in the synset. If the analyzer finds pre-defined keywords in a reviews or comments for a specific movie, then it looks for the modifiers associated for that keyword, it obtains its score from SentiWordNet for the further process. For sentiment analysis, a novel approach using SentiWordNet is proposed that produces the count of scored words by classifying them into the five possible categories such as strong positive, strong negative, neutral, weak positive and weak negative features. These score counts are used to perform the task of sentiment analysis and finally the average of all score features is calculated and the average score is represented using emoticons. For example, the sample output generated for the online movie data review by proposed approach using SentiWordNet is shown in Fig.3. Each line represents a word's score value and its associates score category.

- Random Forest

The random forest classifier is used for classifying the sentiments based on five features such as strong positive, strong negative, neutral, weak positive and weak negative. The sentiment labels are assigned to the average score values in SentiWordNet by classifying them into five features using random forest classifier.

4.4 Algorithm for finding Sentiment Score and Sentiment Label.

Algorithm 1: Algorithm for calculating Sentiment Score.

Step 1: To find the count of reviews present in each document.

Input: Unstructured form of movie reviews

Output: Total count of polarity.

1. For each review present in document do
   Positive Score= ∑strong positive + weak positive
   Negative Score=∑strong negative+ weak negative
2. End for
3. Count = Total Positive Score + Total Negative Score.

Step 2: Find the average score of each movie reviews present in the form of text.

Input: List of sentiment words extracted from each movie reviews.

Output: Final Sentiment Score and Sentiment Summary of all reviews.

1. For each sentiment word present in sentences from List do
2. Obtain the polarity as well as sentiment scores by using SentiWordNet.
3. Find the intensity of each word using SentiWordNet.
4. Final Sentiment Score is evaluated as

$$Sentiment\ Score = \sum_{t=0}^{N} \frac{Maximum\ (Polarity)}{Count}$$

5. Output overall sentiment summary.

Algorithm2: Algorithm for finding the result in the form of Sentiment label.

Initialize C=0
C=Number of positive sentiments (strong positive and weak positive) + Number of negative sentiments (Strong negative and weak negative).
Average of Feature Score= SS/C
If Average of Feature Score >= 1, then:
SL=2
else if Average of Feature Score> 0 && < 1, then:
SL= 1
else if Average of Feature Score < 0 && >-1, then:
SL= -1
else if Average of Feature Score <= -1, then:
SL= -2
else SL=0

Where, C is the counter having total number of positive and number of negative sentiments and SL is the Sentiment Label these labels are given according to sentiment words.

Status: weak positive value: 0.3587747==> simile face entire time

Status: strong positive value: 1.0053347==> true story brave girl great direction

Status: weak negative value: -0.3740736==> not be very clear insane like

Status: strong negative value: -1.0008625==> such boring movie

Status: neutral value: 0.0

Status: strong negative: -3.1037393==> worst movie

Fig.3: Sample output of proposed approach.

4.5 System Execution Details

The one site among the four sites is selected for the particular movie and the list of reviews appears on the screen for the selected movie. Then all the three pre-processing techniques are applied on the dataset and the scores are calculated using SentiWordNet and finally the average of all sentiment summaries are calculated and it is represented in the form of emoticon for better visualization of user. The total number of review are shown in the form of graph based on five categories such as strong positive, strong negative, neutral, weak positive and weak negative.

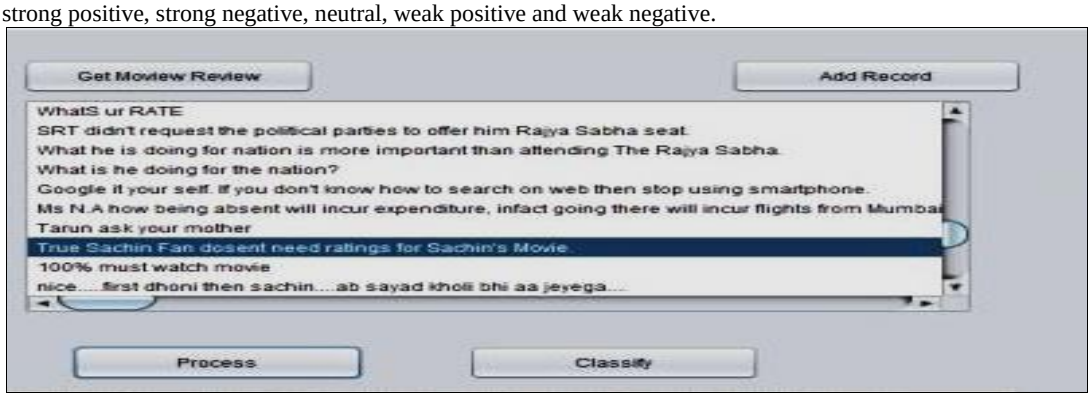


Fig.4: List of reviews for the selected movie.

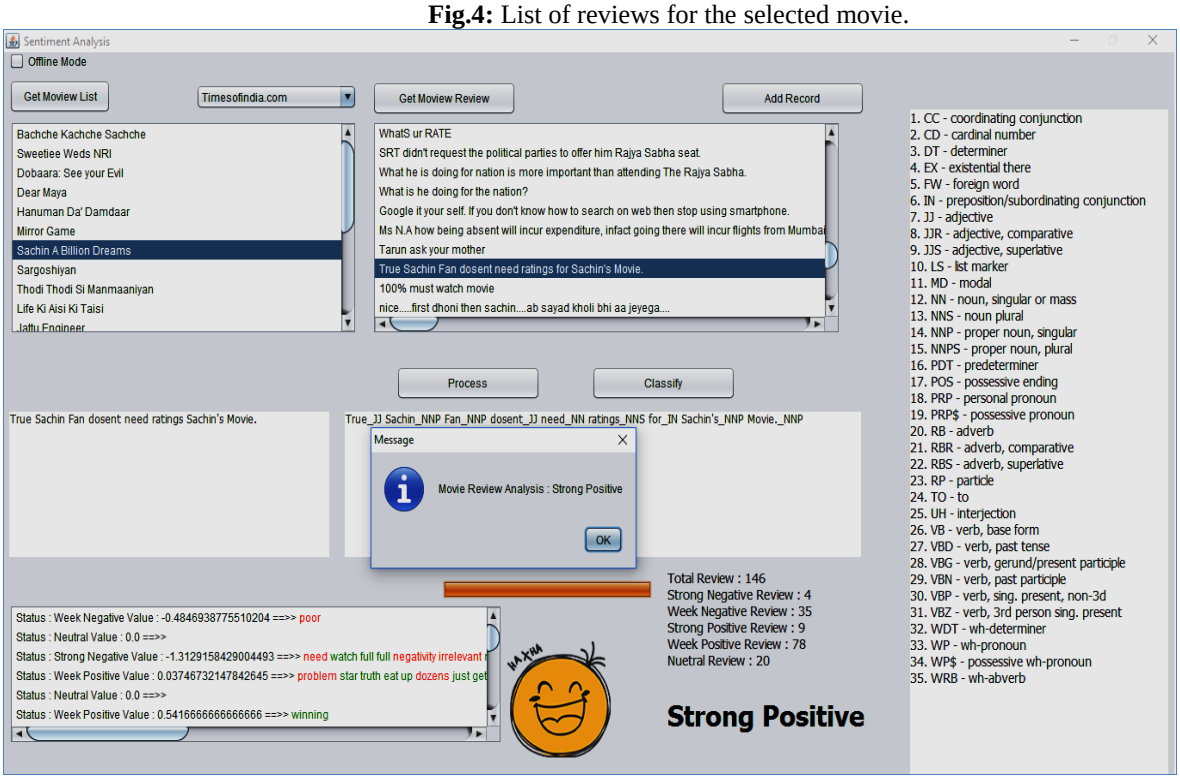


Fig.5: Result of classification process.

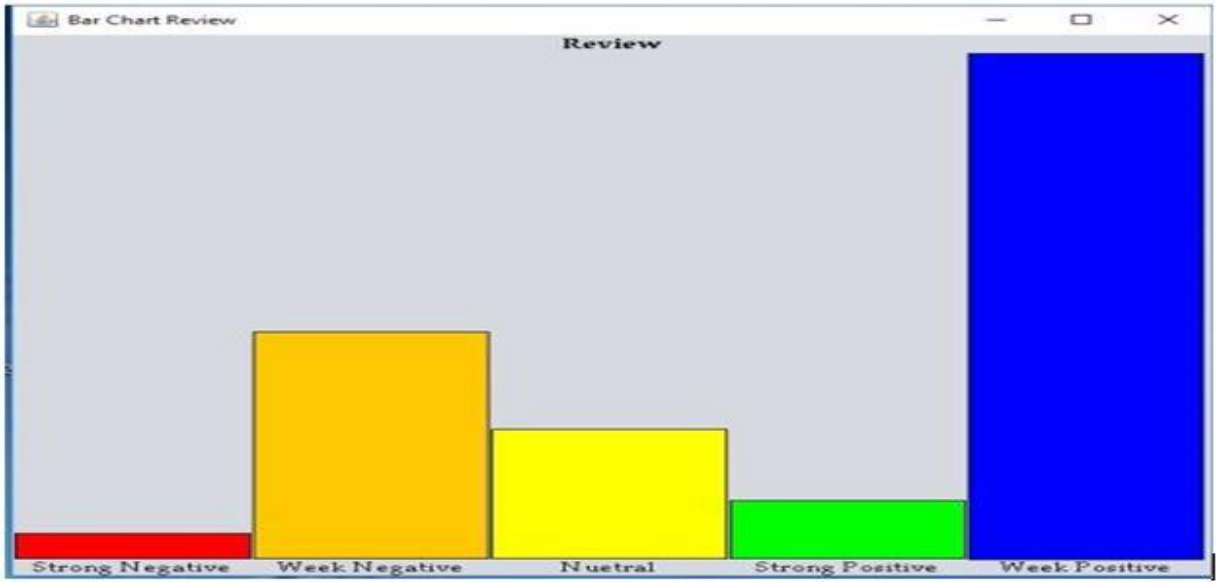


Fig.6: Movie review analysis result based on polarity.

V. Experiment And Result Analysis

5.1 Dataset Description

In this proposed system we have considered the domain of movie reviews, where we have collected reviews from four web sites such as timesofindia.com, bollymoviereviewz.com, rotten tomatoes.com and filmiclub.com movie review datasets. The dataset contains all the unstructured form of reviews. We have considered approximately 2,000 reviews from all four datasets as new movies are released the reviews of that movies are added into the datasets. We split the dataset equally into training and testing sets. Random Forest Classifier is used for achieving better accuracy. We have implemented this using java and tools used for development are Netbeans IDE 8.0.2, jdk 8 and My SQL 5.0.

5.2 Evaluation Measures

The easy way to calculate the accuracy is to validate the performance of the system by using the known sentiment words in datasets. In the proposed system each sentence in the document is represented as a sentiment features and then opinion orientation algorithm is used to capture these features. It classifies the movie reviews according to the classes based on polarity. For classification, system categorized the sentence according to noun/verb/adjective with the help of part-of-speech tagger and calculates score of each sentence with the help of SentiWordNet and finally score is compared with the classes such as strong positive, strong negative, weak positive, weak negative or neutral.

- Performance Measures

The classification performance can be evaluated in three terms: accuracy, recall and precision as defined below.

Table 1: Table of Confusion Matrix

Actual/ Predicted	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

$$Accuracy = \frac{True\ positive\ samples + True\ negative\ samples}{Total\ number\ of\ samples}$$

$$Recall = \frac{True\ positive\ samples}{True\ positive\ samples + False\ negative\ samples}$$

$$Precision = \frac{True\ positive\ samples}{True\ positive\ samples + False\ positive\ samples}$$

Table 2: Table for Performance Parameters

Sr. no	Name of movies web sites	Accuracy	Recall	Precision
1.	Bollymoviereviewz.com	90.90%	95.07%	94.73%
2.	Timesofindia.com	94.19%	94.16%	94.49%
3.	Rotten tomatoes.com	92.28%	90.27%	89.00%
4.	Filmiclub.com	91.74%	93.50%	92.30%

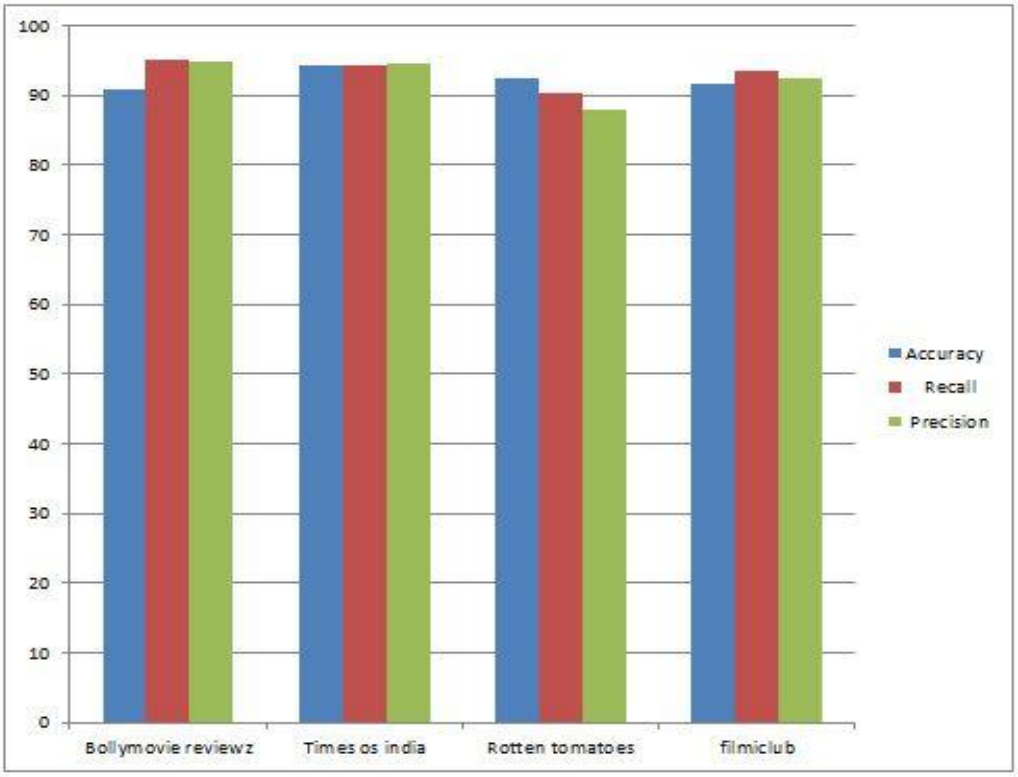


Fig.7: Performance Parameters Graph.

VI. Conclusion

Opinion mining has become popular research area due to the increasing number of internet users on conventional media and social media etc. In this work, we extracted new features that have a strong impact on finding the polarity of the movie reviews. A novel approach using SentiWordNet is proposed that produces the count of scored features by classifying them into the five possible categories such as strong positive, strong negative, neutral, weak positive and weak negative polarity. These score counts are used to perform the task of sentiment analysis. The main goal of this work is to classify the sentences according to its sentiment by using random forest classification technique. The proposed approach is experimented on online movie reviews from four different web sites such as timesofindia.com, bollymovie reviewz.com, rotten tomatoes.com and filmiclub.com and the experimental result reveals the efficiency of the proposed approach along with the accuracy. In future work we would like to apply the concept of NLP in more detail for the better prediction of the polarity results. We would like to use the best classification technique for achieving the highest accuracy. This technique can also be implemented on other domains of opinion mining such as product reviews, political discussion forums, hotels, tourism etc.

Acknowledgment

This work is a part of the postgraduate level project work and I represent my sincere gratitude to all my teachers for their constant guidance throughout the work and providing excellent atmosphere for Dissertation work.

References

[1] P.Nagamma, Pruthvi H.R, Nisha K.K, Carlos Soares," An ImprovedSentiment Analysis of Online Movie Reviews", IEEE 2015, International conference on Computer and Inforamation Technology.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp. 79–86.

[3] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in Proceedings of the 2006 SIGCOMM workshop on Mining network data. ACM, 2006, pp. 281–286 A. Baloglu, Mehmat A. Aktas, "An Automated Framework for Mining Reviews from Blogosphere," International Journal on Advances in Internet Technology, vol. 3, 2010.

[4] Turney, Peter, and Michael L. Littman. "Unsupervised learning of semantic orientation from a hundred-billion-word corpus." (2002).

[5] Baccianella, Stefano, Andrea Esuli and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC. Vol. 10. 2010.

[6] Prabowo, Rudy, and Mike Thelwall. "Sentiment analysis: A combine approach." Journal of Informetrics 3.2 (2009): 143-157.

[7] Ion Smeureanu, Cristian Bucur, "Applying Supervised Opinion Mining Techniques On Online User Reviews", Informatica Economică, 2012.

[8] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125,2006.

[9] Zhai Zhongwu, Liu Bing, Xu Hua, Jia Peifa, 2011. Clustering product features for opinion mining. Paper presented at the fourth ACM international conference on Web search and data mining, Hong Kong, China.

[10] Dave Kushal, Lawrence Steve, Pennock, M. David, 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Paper presented at the www2003, Budapest,Hungary.

[11] Singh, V. K., et al. "Sentiment analysis of movie reviews: A new featurebased heuristic for aspect-level sentiment classification". Automation, and the Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on. IEEE, 2013.

[12] X. Yu, Y. Liu, J. X. Huang, A. An, Mining online reviews for predicting sales performance: A case study in the movie domain, IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 4, APRIL 2012.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

Rasika Wankhede. "To Improve Accuracy in Movies Reviews Using Sentiment Analysis." IOSR Journal of Computer Engineering (IOSR-JCE) 19.4 (2017): 80-88.