Sentiment Analysis of Movie Reviews Using Lexicon Approach

Purtata Bhoir
Department of Computer Engineering
Saraswati College of Engineering
Kharghar, Navi Mumbai, India
bpurtata@gmail.com

Shilpa Kolte
Department of Information Technology
Saraswati College of Engineering
Kharghar, Navi Mumbai, India
shilpakolte17@gmail.com

***Abstract*— Sentiment analysis is a language processing task which is used to find out opinion expressed in online reviews to categorize it into different classes like positive, negative or neutral. The paper aims to summarize the movie reviews at aspect level so that user can easily find out which aspects of movie are liked and disliked by user. Before finding aspect and its respective opinion of movie, proposed system performs subjectivity analysis. Subjectivity analysis is one of the important and useful tasks in sentiment analysis .Online reviews may consist of both objective and subjective sentences .Among these, objective sentences consist of only factual information and no sentiments or opinion. Hence subjective sentences are considered for further processing i.e. to find feature- opinion pair and to find summery at aspect level. In this paper, two different methods are implemented for finding subjectivity of sentences and then rule based system is used to find feature-opinion pair and finally the orientation of extracted opinion is revealed using two different method. Initially the proposed system uses SentiWordNet approach to find out orientation of extracted opinion and then it uses the method which is based on lexicon consisting list of positive and negative words.**

Keywords— Sentiment analysis, Subjectivity, Objectivity, NLP, SentiWordNet.

I. INTRODUCTION

We always considers others opinion while doing anything like before going for shopping, before going to watch any movie or even at the time of giving vote to politician. In earlier days, we used to take this information from our friends, relatives or from consumer report. But now time has changed. Today in Internet era, more and more people can connect easily with each other. Internet has made it possible for us to find out the opinions and experiences of others that are neither our relatives nor well-known professional critics — that is, people we have never heard of. There are many ways (Forum, Blogs) on the Internet through which people can give their opinion. Many products as well as services are available online so it's now easier for manufacturer as well as service provider to establish direct interaction with customers through their online feedback in the form of reviews. This review helps both customer as well as manufacturer. For customers it helps to get the idea about product or service and for manufacturer it helps to improve quality of their product or services.

Because more and more people are willing to share their views or experiences on the Internet, huge volumes of data is produced .And because of this it becomes tedious task for any person or organization to make decision. Thus an automated system (Sentiment analysis system) is meant to automate the process of analysis, summarization and classification of data. The task of sentiment analysis can be carried out at different levels i.e. at document level, sentence level and aspect level. Sentiment analysis at document level assumes that reviews have opinion about single entity i.e. at this level system tries to find whether entire document expresses a positive or negative reviews. Analysis at sentence level determines opinion for individual sentence or review. The task at this level is related with subjectivity analysis. Both document and sentence level analysis fails to find out what exactly people liked and disliked. For example the review about a movie may consist of both positive and negative aspects about specific feature of movie. In this case it would be inappropriate to work at document or sentence level analysis. Hence aspect level is more appropriate solution for more complete and accurate system.

II. RELEATED WORK

In sentiment analysis, we analyze people's opinion towards any entity like any product or services. There is lot many work have been done in this area. In sentiment analysis, subjectivity is one of the main subtasks. In early research, subjectivity of sentences considered as a standalone problem .But, in recent research, many researchers considered it as one of the important task during sentiment classification. Identification of subjective sentences is much highly difficult compared with sentiment analysis [1] and subjective analysis system is responsible to increase overall performance of sentiment analysis.

Wiebe and Riloff et al, [2] used bootstrapping process for subjectivity classification. In this work, they have used two classifiers; one classifier searches given dataset for subjective sentences whereas other classifier searches for objective sentences .Classified sentences are then fed to an extraction pattern learner, which produces extraction pattern for subjective sentence.

The method proposed by authors in [3], is based on SentiWordNet. It used linguistic feature consisting of adjective, adverb and verb. They have performed sentiment analysis at document level as well as at aspect level.

978-1-4799-7849-6/15/$31.00 ©2015 IEEE

Authors in [4] propose subjectivity classifications using sentence similarity and a naïve Bayes classifier. The sentence similarity method is based on the assumption that subjective or opinion sentences are more similar to other opinion sentences than to factual sentences. They used the SIMFINDER system in to measure sentence similarity based on shared words, phrases, and WordNet synsets.

Proposed method in [5] is based on latent semantic analysis to identify features .This method also let the people to choose the features of entity in which they are interested. This reduces the size of summery.

Proposed scheme in [6] studies aspect based opinion mining. In this method, a multi-aspect bootstrapping method is proposed for aspect identification and then aspect-based segmentation is used segment a multi-aspect sentences.

Authors in[7] , used min-cut method for subjective sentence extraction , and further classification algorithm like Naïve Bayes ,Support Vector Machine are applied for further processing.

Wiebe and Riloff [8] have worked on discovered patterns to generate a rule-based method to produce training data for subjectivity classification. The rule based subjective classifier classifies a sentence as subjective if it contains two or more strong subjective clues (otherwise, it does not label the sentence). In contrast, the rule-based objective classifier looks for the absence of clues: it classifies a sentence as objective if there are no strong subjective clues in the sentence, and several other conditions. The system also learns new patterns about objective sentences using the information extraction system Autos log-TS (Riloff, 1996) [9], which finds patterns based on some fixed syntactic templates.

Proposed scheme in [10] applies different rules to review document to extract feature and opinion pair .For subjectivity classification, they have used decision tree classifier of Weka.

B.Pang et al. [11] used machine techniques as Bayesian, Maximum Entropy and SVM algorithm to classify the review of movie.

III. SYSTEM DESIGN AND IMPLEMENTATION

A. System Structure

The proposed system aims to find out orientation of movie reviews. The task is conducted in following main steps. Shown as Fig.1.

Step 1: Initially the reviews are given as input to document processor where preprocessing is done on collected reviews. In this phase, all slang as well as informer words are replaced with proper words .For example if any review consist of word *congrat*s then with preprocessing it will get converted with word *congratulations.*

Step 2: Pre-Processed data is then given to next module i.e. document parser which applies Part-Of-Speech tag to each of the word which is then used to find out category of the word.POS tagger try to find out POS tags for given words, because morphological analyzer can't make decision about any specific word. For example, duck whether it is verb or noun. A POS tagger can make decision by looking the adjacent or neighboring words. For example "Duck is delicious for dinner"; here duck will be tagged with noun tag.

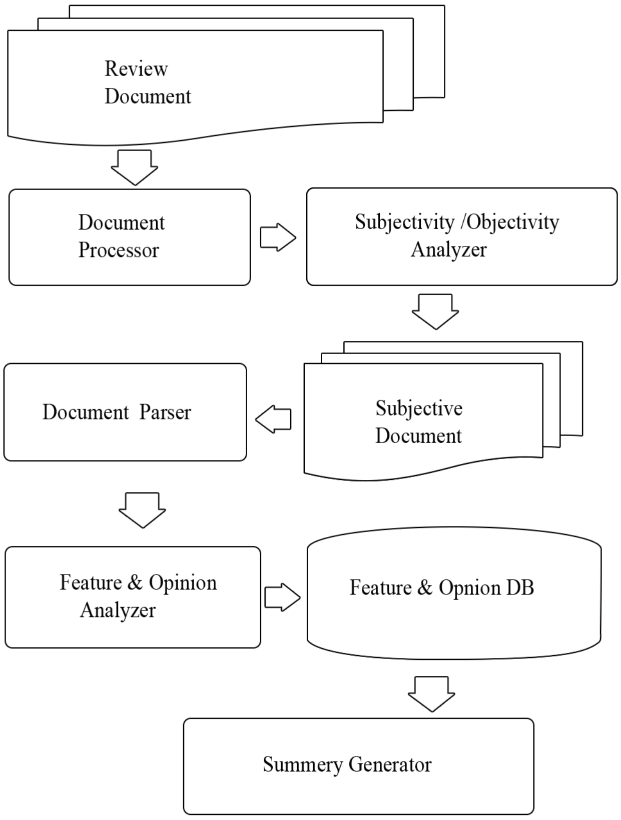

Fig.1. Architecture of Proposed System

Step 3: Once tagging is done reviews are taken for further processing of subjectivity analysis where system finds sentences which are more subjective in nature and filter out those that are more objective. We have used two different methods –machine learning and SentiWordNet to find subjectivity of sentences.

Step 4: In this step, system takes input from document parser and outputs the feature and its opinion pair by analyzing noun phrases and the associated adjectives by using rule-based system.

Step 5: In this step the final feature-based review summary is get generated. Here each discovered feature, related opinion is put into positive and negative categories according to the orientation of opinion.

B. Subjectivity / Objectivity Analyzer

For subjectivity analysis, two different methods are explored i.e. Naïve Bayes and SentiWordNet. The Naïve Bayes classifier is a probabilistic model based on the Bayes theorem. We have used this theorem to calculate probability to determine the class of individual sentence. For classification we have considered publically available dataset from http://www.cs.cornell.edu/people/pabo/movie-review-data. This dataset consists of 5000 subjective and 5000 objective sentences. The task of classification is implemented in two main phase's i. e. training phase and testing phase. The steps of training phase are shown in Fig. 2

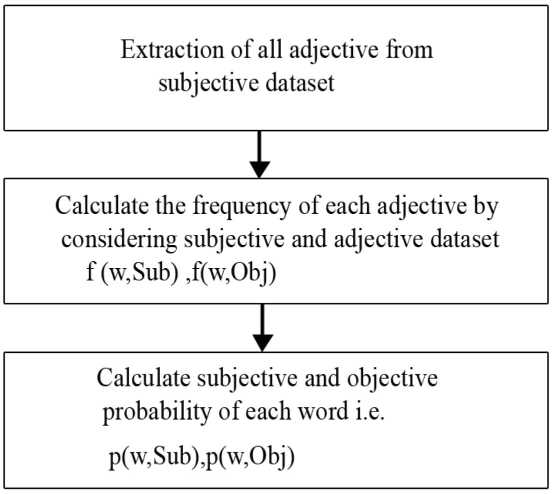
Extraction of all adjective from subjective dataset
Calculate the frequency of each adjective by considering subjective and adjective dataset f (w,Sub) ,f(w,Obj)
Calculate subjective and objective probability of each word i.e. p(w,Sub),p(w,Obj)

Fig.2. Training Phase

The training phase consists of following steps.

1. In first step we apply POS tagging on dataset and extracts adjectives from it.
2. In second step, we retrieve the subjective and objective score for all words belonging to sentence.
3. In third step, we calculate probability of word being classified as subjective using equation (1).

$$p(w,Sub)=\frac{\mathrm{f(w,Sub)}}{\mathrm{f(w,Sub)+f(w,Obj)}} \quad (1)$$

Similarly we calculate probability of word being classified as objective using equation (2).

$$p(w,Obj)=\frac{\mathrm{f(w,Obj)}}{\mathrm{f(w,Obj)+f(w,Sub)}} \quad (2)$$

Thus, in this step we create table consisting 3 entries first is word which we have extracted from subjective dataset, second is subjective probability p(w,sub) and third is objective probability p(w,obj).

In testing phase we consider testing data i.e. Movie reviews as input. The testing phase consists of following step

1. In first step we apply POS tagging on reviews and extracts adjectives from it.
2. In second step, we retrieve the subjective and objective score for all words belonging to sentence.
3. Finally, average subjective and objective scores of all adjectives of a sentence are computed using following equation.

$$Sub(s)=\frac{\sum_i p(wi,Sub)}{n} \quad (3)$$

$$Obj(s)=\frac{\sum_i p(wi,Obj)}{n} \quad (4)$$

Where, s is sentence & n is no. of words.

4. In last step, we discard objective sentences from reviews and only subjective sentences are processed further. To select only subjective sentences, we simply use constraint i.e. if sub(s)>obj(s) then sentence is marked as subjective .The flow of testing phase is shown in fig.3.

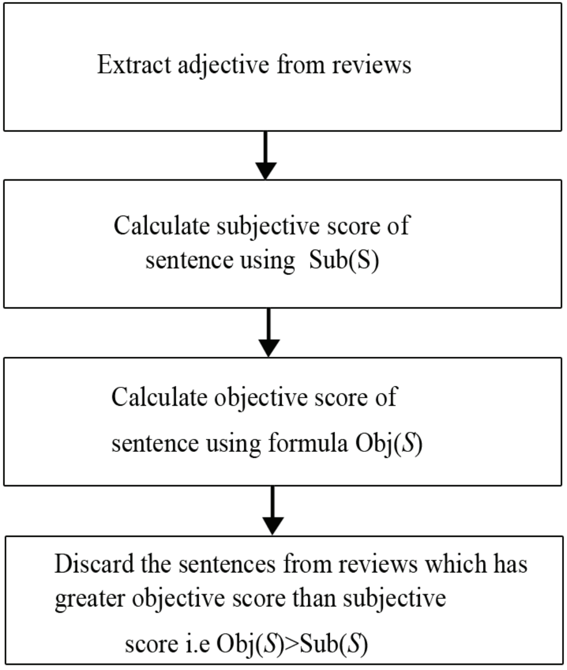
Extract adjective from reviews
Calculate subjective score of sentence using Sub(S)
Calculate objective score of sentence using formula Obj(S)
Discard the sentences from reviews which has greater objective score than subjective score i.e Obj(S)>Sub(S)

Fig.3. Testing Phase

In SentiWordNet approach, we have used two linguistic feature schemes. In first scheme we consider only adjectives from individual reviews .In other we extracts both adjective, verb along with adverb.

SentiWordNet is a lexical resource for opinion mining and it assigns positive and negative score to each synsets of WordNet. Since adjectives play vital role in expressing opinion; it has more weight than other part-of-speech word. For Example, "I like the story of that movie" .In this sentence, the word "like" is an adjective and expresses positive sentiment about movie.

The subjective score of an adjective can be calculated using equation no (5).

$$Sub(wi)=\alpha*(|Pos(wi)|+\ |Neg\ (wi)|) \quad (5)$$

Similarly, the objective score of an adjective is calculated using following equation no. (6).

$$\mathrm{Obj(wi)}=\alpha*\left(1-\mathrm{Sub(wi)}\right) \quad (6)$$

Further, if a word is an adverb, verb then subjective scores is calculated using equation (7)

$$\mathrm{Sub(wi)}=\beta*(|\mathrm{Pos(wi)}|+\ |\mathrm{Neg\ (wi)}|) \quad (7)$$

Similarly, the objective score is calculated using following equation (8).

$$\mathrm{Obj(wi)} = \beta * \left(1 - \mathrm{Sub(wi)}\right) \tag{8}$$

Here, Pos (wi) is a positive score of word and Neg (wi) is a negative score of word from SentiWordNet. The symbol α and β are constant where α > β.

This way, we find subjective and objective score of all words from individual sentence from online review .Further, the final subjective as well as objective score of sentence is calculated using equation no. (9) and (10) respectively.

$$\mathrm{Sub}_{\mathrm{score}} = \frac{\sum_{i=1}^{n}(\mathrm{Sub(Wi)})}{n} \tag{9}$$

$$\mathrm{Obj}_{\mathrm{score}} = \frac{\sum_{i=1}^{n}(\mathrm{Obj(Wi)})}{n} \tag{10}$$

Once the total score of an individual sentence is calculated, we discard the sentences for which the condition ObjScore >SubScore is true. This way we just consider subjective sentences from given document.

C. *Feature and Opinion Analyzer*

This module extracts the opinion and feature pair from subjective document. To find out feature and opinion pair, it takes typed dependency relation as input generated by Stanford Parser. This module is implemented as a rule based system. Following are rules of rule based system,

Rule1: In a typed dependency relation R, if we find nn(w_1, w_2) relation between w_1 and w_2 word and nsubj(w_3, w_1) relation between w3 and w1 word , where POS tag of w_1 word is NN, POS tag of w_2 word is NN, POS tag of word w_3 is JJ and w_1, w_2 words are not stop words or if we get nsubj(w_3,w_4) relationship between w_3 and w_4 word, such a way that POS tag of w_3 is JJ, POS tag of w_4 is NN and w_3, w_4 are not stop words then from this relationship we can conclude that either word (w_1, w_2) or w_4 is extracted as a feature and w_3 as opinion.

Rule 2:
In a typed dependency relation R, if we find nn(w_1, w_2) relation between w_1 and w_2 word and nsubj(w_3, w_1) relation between w_3 and w_1 word such that POS tag for w_1 is NN, POS tag for w_2 is NN, POS tag for w_3 is VB and w_1, w_2 are not a stop-words, or if, there exist a relationship nsubj(w_3,w_4) where POS tag for w_3 is VB, POS tag for w_4 is NN and w_4 is not a stop-word, then another relation i.e. acomp(w_3,w_5) relation is searched. And if acomp relationship exists such that POS tag of word w_5 is JJ and w_5 is not a stop-word then either (w_1, w_2) or w_4 is extracted as the feature and w_5 as opinion.

Rule 3:
In a typed dependency relation R, if we find relation nn(w_1, w_2) between word w_1 and w_2 and nsubj(w_3,w_1) between word w_3 and w_1 such that POS tag of word w_1 is NN, POS tag for word w_2 is NN, POS tag for word w_3 is VB and w_1, w_2 are not stop-words, or if we find relationship nsubj(w3,w4) between word w_3 and w_4 such that POS tag for w_3 is VB, POS tag for w_4 is NN and w_4 is not a stop-word, then we search for another relation called as dobj(w_3, w_5) relation between word w_3 and w_5. And if there exists a dobj relationship where POS tag for w_5 is NN and w_5 is not a stop-word then either (w_1,w_2) or w_4 is extracted as the feature and w_5 as opinion.

Rule 4:
In a typed dependency relation R, if we find relation amod (w_1, w_2) relation between word w_1 and w_2 where POS tag for word w_1 is NN, POS tag for word w_2 is JJ, and w_1 and w_2 are not stop-words then word w_2 is extracted as an opinion and w_1 as a feature.

D. *Summery Generator*

For each extracted opinion, we need to find its polarity suggesting whether opinion is positive or negative. Summery generator module exploits each feature-opinion pair to find polarity and generates summarized view. Generated feature based summery helps the customer to know positive and negative aspects of movie without going through large no. of reviews.

To get more accurate result, we used two different methods. First we extract orientation with SentiWordNet (SWN). SWN is a publicly available lexical resource which assigns positive and negative e numerical score to each extracted term. For each word, SWN gives part-of-speech (POS) tag, unique ID, its positive and negative score. For example, for word ‘awesome’, SWN gives details as (a, 01282510, 0.875, and 0.125) where ‘a’ indicates adjective which is POS tag for word ‘awesome’, 01282510 is its unique ID, 0.875 is positive score and 0.125 is its negative score. After finding orientation with SWN, we find it with our own lexicon of two text files consisting list of positive and negative words. Finally to find resultant orientation of word we compare the result of two methods and consider best result among these two methods. The algorithm to find orientation with lexicon consisting positive and negative list of words is mentioned in fig 4.The Algorithm to find resultant orientation is presented in fig, 5.

```
String s1= “Extracted Opinion Word for which we need to find orientation ”
Integer POS, NEG
Integer FLAG=0
String Orientation
    1.  If PosWd.contains (s1) Then
        FLAG=1; POS++; Orientation is POSITIVE
        //Where PosWd is list of positive words.

    2.  If NegWd.contains (s1) Then
        FLAG=1; NEG++; Orientation is NEGATIVE
        //Where NegWd is list of negative words.

    3.  If (FLAG! =1)
        Orientation is NEUTRAL
```

Fig.4. Orientation Extraction with own Lexicon

```
String s1= "Opinion Word"
String Orientation = ""
String lab1=Orientation of s1 w.r.t. own lexicon consisting positive and negative list of words.
String lab2=Orientation of s1 w.r.t. SentiWordNet (SWN).
    1. If lab1==lab2 , Orientation =lab1/lab2
    2. Else if lab1 is Neutral, Orientation =lab2
    3. Else if lab2 is Neutral , Orientation =lab1
    4. Else if lab1 is positive and lab2 is Negative
       Then
            a. If POS (s1) >
               Negative_score_SentiWordNet (s1)
               THEN Orientation =lab1
               Else
                 Orientation =lab2
// where POS (s1) is positive score (s1) from Own lexicon from Algorithm 1.

    5. Else if lab1 is Negative && lab1 is Positive
       Then
            a. If POS (s1) >
               Negative_score_SentiWordNet (s1)
               THEN Orientation =lab1
               Else
               Orientation =lab2
```

Fig.5. Orientation Extraction by Comparing Result of Two Methods

IV. EXPERIMENTAL RESULT AND DISCUSSION

In this section, we present the experimental details of proposed system. For subjectivity analysis, we used the Naïve Bayes, SWN schemes. For analysis, we obtained reviews from popular movie website i.e. www.imdb.com. To evaluate the performance we have used standard IR performance measure in which we consider TP (true positive) FP (false positive), TN (true negative) and FN (false negative).

Precision (π): It is the ratio of true positive records considering all sentences .The precision is calculated using equation no. (11)

$$\pi = \frac{TP}{TP+FP} \tag{11}$$

Recall (ρ): It is ratio of true positive considering only positive sentences. It is calculated using equation no. (12).

$$\rho = \frac{TP}{TP+FN} \tag{12}$$

F1- Measure (F1): It is harmonic mean of recall and precision .It is calculated using equation no. (13).

$$F1 = \frac{2\rho\pi}{\rho+\pi} \tag{13}$$

Accuracy (τ): It gives the ratio of sum of TP and TN over total positive and total negative instances. Accuracy is calculated using equation no. (14)

$$\tau = \frac{TP+TN}{TP+FP+FN+TN} \tag{14}$$

Table I shows the performance measure values for subjectivity and Fig. 6 and fig 7 shows the pictorial view of results shown in Table I.

TABLE I. PERFORMANCE EVALUATION OF SUBJECTIVITY

Method Name	**Precision (%)**	**Recall (%)**	**F1-Measure (%)**	**Accuracy (%)**
Naïve Bayes	75	75	75	71.42
SWN	60	37.5	46.15	53.33

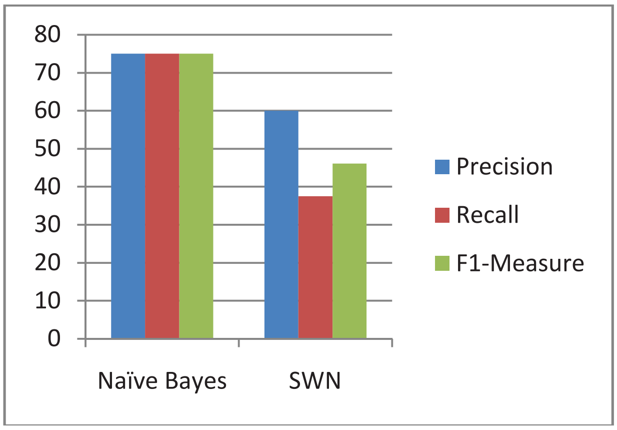

Fig.6. Precision, Recall, F1 Measure of Subjectivity

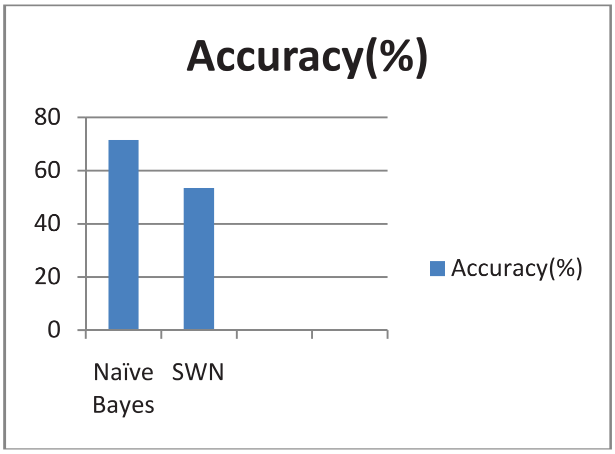

Fig.7. Accuracy Result

The experimental result for subjectivity shows that the Naïve Bayes method performs better than SWN approach.

Further, proposed system is analyzed for feature and opinion extraction. Here, we have collected 20 reviews of 5 different Hindi movies from website www.imdb.com. We have manually extracted feature and opinion from these reviews and compare result with proposed system. Table II shows the performance measure values for feature and opinion extraction process. Fig. 8 shows the pictorial view of results shown in Table II.

TABLE II. PERFORMANCE EVALUATION OF FEATURE EXTRACTION

Movie Name	Precision (%)	Recall (%)	F1-Measure (%)
Bajrangi Bhaijaan (BB)	86.12	69.42	75.38
ABCD 2	78.20	53.43	62.23
Roy	79.24	56.50	64.55
Badlapur	82.08	60.62	69.34
Gabbar is Back	82.45	61.23	68.11

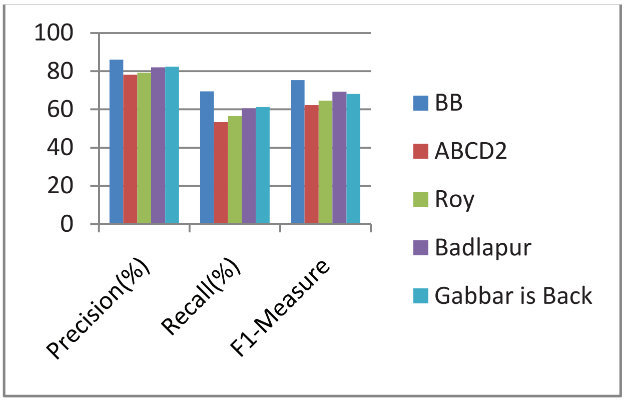

Fig.8. Precision .Recall and F1-Measure of feature Extraction

The experimental result shows that the values for precision are high. It indicates that the extracted opinion and feature pair are correct.

V.CONCLUSION AND FUTURE WORK

Subjectivity deals with extraction of subjective sentence and it is one of the important tasks in sentiment analysis which increases the system performance both in terms of efficiency and accuracy. In this paper, we have presented system which implements two different methods to find subjectivity of sentences. Among these two methods, Naïve Bayes classifier gives more accurate result than SentiWordNet. As there is need to find different aspects of movie and its respective opinion, we implemented rule based system which allows user to easily check different aspect of movie liked or disliked by other user. In our future work we will implement system that would analyze reviews which are in language other than English.

REFERENCES

[1] Wei Jiang, “Study On Identification of Subjective Sentences in Product Reviews Based on Weekly Supervised Topic Model”, Journal of Software, Vol.9, No.7July 2014,pp. 1952-1959.

[2] Riloff Ellen and Janycee Wiebe, “Learning Extraction Patterns for Subjective Expressions”, In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03),pp.105-112.

[3] V.K. Singh, R. Piryani, A. Uddin, P. Waila, “Sentiment Analysis of Movie Reviews .A new Feature-based Heuristic for Aspect-level Sentiment Classification”, IEEE 2013,pp.712-717.

[4] Hong Yu and Vasileios Hatzivassiloglou, “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences”, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03),pp.129-136.

[5] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, “Movie Rating and Review Summarization in Mobile Environment” , IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 3, May 2012, pp.397-407.

[6] Jingbo Zhu, Member, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, and Matthew Ma, “Aspect-Based Opinion Polling From Customer Reviews” , IEEE Transactions On Affective Computing, Vol. 2, No. 1, January-March 2011, 2011 IEEE Published By The IEEE Computer Society, pp.37-49.

[7] B.Pang, L.Lee, “Sentiment Analysis Using Subjectivity Summarization Based on minimum Cuts”, Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, 2004,pp.271-278.

[8] Janycee Wiebe and Ellen Riloff, “Creating subjective and objective sentence classifiers from Unannotated Texts”, Proceeding of CICLing-05, Volume 3406 of the series Lecture Notes in Computer Science,2005,pp.486-497.

[9] Ellen Riloff, “Automatically generating Extraction Pattern for untagged Text” In the procedding of Thirteenth National Conference on Artificial Intelligence (AAA I-96),volume 2,pp.1044-1049.

[10] Tanvir Ahmad,Mohammad Najmud Doja, “Rule Based System for Enhancing Recall for Feature Mining from Short sentences in Customer Review Documents”,International journal on Computer Science and Engineering(IJCSE), ISSN : 0975-3397 Vol. 4 No. 06 June 2012,pp.1211-1219.

[11] Pang, Bo,Lillian Lee, and Shivkumar Vaithyanathan, “Thumps Up ?sentiment classification using machine learning techniques” .In procedding of the confrence on Empirical Methods in Natural Language Processing (EMNLP),2002,pp 79-86.