Design Approach for Accuracy in Movies Reviews Using Sentiment Analysis

Rasika Wankhede
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Sewagram (Wardha), India
wankhede.rasu@gmail.com

Prof. A.N.Thakare
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Sewagram (Wardha), India
amu_thak@rediffmail.com

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, a specific view or judgment on certain topic. Sentiment analysis system classifies text data into their respective sentiments of positive polarity, negative polarity or neutral. 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 paper, we try to focus our task of sentimental analysis on "times of india" movie review database. We examine the sentiments present in the text document for classification of movie reviews based on polarity (positive/ negative/ neutral). Also we have used the Random Forest classifier for the evaluation of performance and for finding the accuracy. By using Random Forest classification technique we have achieved the best accuracy of 90%.**

Keywords- Opinion Mining, Sentiment Analysis, Polarity, Sentiments, Movie reviews, Feature extraction, Classifier.

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 "times of india" website 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 from class of positive, negative and neutral. 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 negative score. Contextual understanding is difficult for a system to reach human level accuracy. We have used five feature labels such as Strong Negative- (-2), Weak Negative- (-1), Neutral- 0, Weak Positive- 1, Strong Positive- 2 with the help of these feature labels we can find the exact polarity result for phrases such as "NOT VERY BAD" sense that it is neither completely negative nor completely positive phrase but this phrase is weak positive phrase as it is not having completely negative polarity. Other phrases such as "NOT SO INTERESTING" will be weak negative phrase as it is not

having completely positive polarity and phrases such as "SO BORING" will be strong negative phrase and "VERY INTERESTING" will be the strong positive phrase. The sentence which does not contain any positive or negative sentiments is neutral. Finally the weak positive and strong positive features get converted into positive polarity and weak negative and strong negative features are converted into negative polarity. The result is graphically displayed in the form of positive and negative polarity.

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, by forming the list of 14 keywords which are useful in finding the baseline for classification accuracy. The machine learning techniques like Naïve bayes, SVM, achieves higher accuracy over the baseline. 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] studied the unsupervised learning algorithm for sentiment classification process. They determined the similarity of words with help of NEAR operation and developed a classifier for finding polarity result.

Stefano, Andrea and Fabrizio [5] present SentiWordNet 3.0, it is a lexical resource developed 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. 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. Rui Yao et al. [7] proposed a simplified version of sentiment aware autoregressive model this model can be used for producing the good accuracy for prediction of box office sale revenue using online movie review data. NB classifier is used for the sentiment classification. Mullen and Collier et al. [8] proposed an approach for classification of text data into positive or negative polarity using SVM. Their work involved extraction of value phrases (two word phrases conforming to a particular part-of-speech) and assigning them sentiment orientation values by using point wise mutual information.

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) Classifiers such as SVM and Naïve Bayes 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 feature scores are not present, so it becomes difficult to decide which phrase is either completely positive or completely negative for example phrase such as "NOT VERY BAD". This problem is solved by providing labels to each features, 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 Ranking and second using machine learning techniques. We use "times of india" 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 Figure.2 explains the overall methodology.

A. Input Data

The input data is in the form of reviews from the "times of india" movie review dataset. 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.

B. Pre-Processing

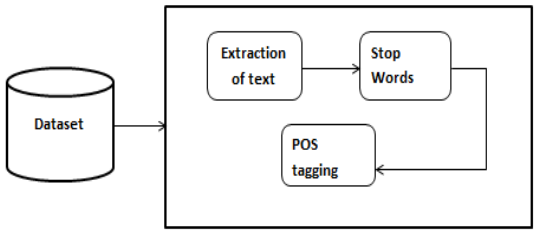

Fig.1: Pre-processing of dataset.

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

  “This movie is amazing”

  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, amazing is tagged with tag JJ which indicates ‘amazing’ is an adjective where as a ‘movie’ is tagged as NN which indicates noun.

C. Text Transformation

In the process of text transformation the score of each sentence in the source document is calculated by sum of weight of each term in the corresponding sentences. The weight of each term is calculated by multiplication of that words based on adjective word extracted from part-of-speech. The output of pre-processing process is given as input to text transformation process.

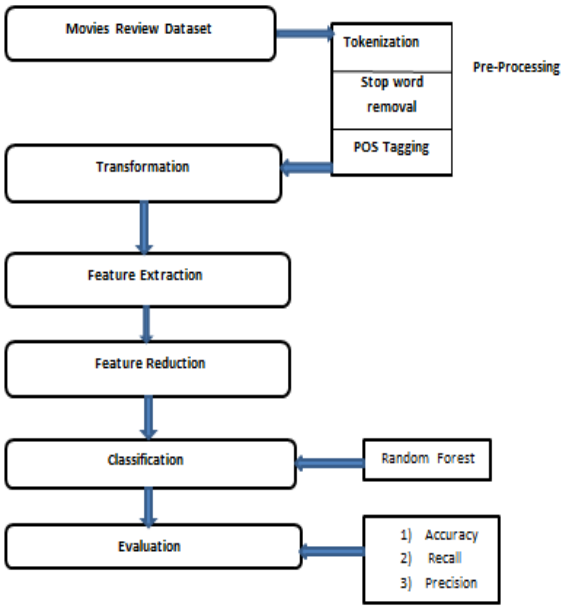

Fig.2: Work Flow diagram.

D. Feature Extraction

In the process of feature extraction, movie features are extracted from every sentence. For finding the polarity of text document, it is necessary to understand the sentiment score with its usage as well as their relationship with all the nearby words. Following are some features that affect the polarity of the document.

1) Positive Sentiment Words: These are the words which are having a positive sentiment score according to SentiWordNet. For example: Nice, Good, Fantastic, Pretty, Outstanding etc.

2) Negative Sentiment Words: These are the words which are having negative sentiment score according to SentiWordNet. For example: Bad, Awful, Disgusting, Pathetic etc.

3) Unigram Model: In unigram model the whole sentence is divided into number of words, and those words are used as features.

Unigram example “it is an interesting movie”.
Output feature set {it, is, an, interesting, movie}.

4) Bigram Model: In bigram model the combination of two words is used for creating the feature vector.

Bigram example: “it is not good movie”
Output feature set {it is, is not, not good, good movie}.

5) N-Gram Model: This model refers to combination of more words together for forming feature vector and using that feature vector for data classification.

6) *Sentiment words combined with adjectives*: These are the positive or negative sentiment words which are preceded by an adjective. For example: "An excellent action movie" or "A boring movie is just a waste of money".

7) *Sentiment words having repeated letters:* These are the positive or negative sentiment words having repetition of letters. For example: Awwwesome, excellenttttt, Awwwfull, borrrring etc.

E. Feature Reduction Approach

One of the biggest problems of sentimental analysis is dealing with text data which are available in very high dimensions which may affect the performance of classifier. So, there is a need for such technique which will eliminate those features that are not relevant and keeping only those features which are much important and the techniques which will help to differentiate the sentences into class labels such as positive and negative. The Information Gain and Gain Ratio are the most popular techniques among number of feature reduction techniques.

- *Information Gain:* Information Gain technique is mainly used for finding importance of a feature in decreasing overall entropy. It works on the basis of information required for a text document to be classified in a respective class, depending on presence and absence of words in that text document. It is a statistical property that identifies how features separate the training samples corresponding to the target classification. Information Gain process is mainly based on the measure entropy. The entropy measure indicates the impurity of collected samples. The entropy is defined as:

$$Entropy\ (D) = -\sum_{i} P\ (i/D)\ log_2\ P\ (i/D)$$

Let D be the dataset and P be the probability of random text in data set is positive, negative or neutral and P(i/D) be the fraction of D belonging to class i. The class is represented by positive, negative or neutral. The information gain of feature y related to a collection of dataset D, is defined as:

$$IG(D, y) Entropy\,(D) - \sum_{x \in value(y)} \frac{|Dx|}{|Dy|} Entropy\,(Dx)$$

Where, value(y) is the set of all possible values of feature y. these values can be positive, negative or neutral value. Dx be the subset of D with the class x related to feature y. Sy is the subset of feature y it denotes the cardinal number indicating the number of features present in the set.

- *Gain Ratio*: It is modification of information gain technique. In gain ratio the contribution of all features will be normalized before the classification of the document. The gain ratio works better as compared to information gain technique.

*F. **Algorithm for finding Sentiment Score and Sentiment Label**.*

Algorithm 1: Algorithm for calculating SentiScore (Sentiment Score).

All features are initialized to 0.
(Check every sentence for finding the features present in it) for loop.
(Check every feature extracted from the sentences) for loop.
If negative sentiment word occurs && adjective, then
SentiScore= Score(W)+ (Score(W-1)*1.5))
else if, (W-1) is negative sentiment word, then
SentiScore= Score(W)+(Score(W-1)*1.6)
If positive sentiment word && adjective, then
SentiScore= Score(W)+(Score(W-1)*1.7)
else if, (W-1) is positive sentiment word, then
SentiScore= Score(W)+(Score(W-1)*2.3)
SentiScore= Score(W).

Where, W is the sentiment word and SS (Sentiment Score) it is calculated according to five features we have defined. The weight of each score features depends upon the information gain ratio and feature ranking. Once the SS (Sentiment Score) is obtained we find the result using following algorithm.

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

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

G. 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, decision tree and K-Nearest Neighbour technique. The classification is done with the Random Forest classifier to determine the sentiment labels for a machine and to predict the class of a movie reviews whenever it arrives in the form of positive or negative polarity. We have performed the feature impact analysis by computing 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 number of positive and negative sentiment. The sentiments in the sentence are classified according to polarity and the result of classification process is appears as shown in Figure.4 there are total 49 reviews among which 37 reviews are positive reviews and 12 reviews are negative reviews.

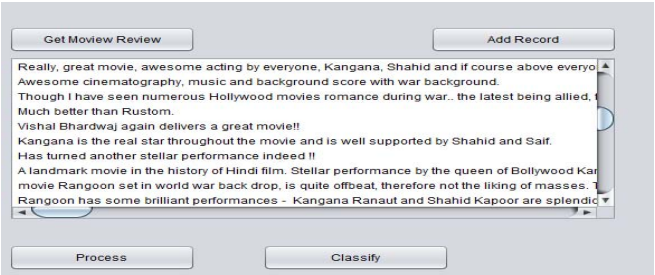

Fig.3: List of reviews for the selected movie.

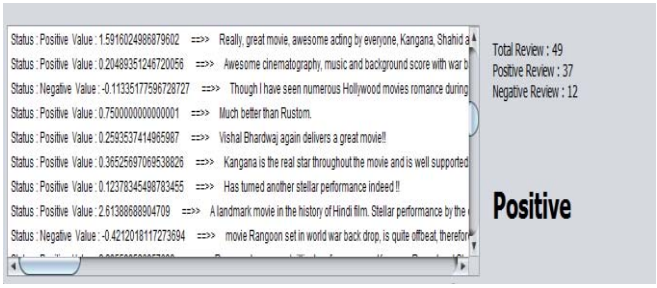

Fig.4: Result of classification process.

V. EXPERIMENT AND RESULT ANALYSIS

A. Dataset Description

We have collected reviews from times of india (http://timesofindia.Indiatimes.com/entertainment/hindi/movie_reviews) movie review dataset. The dataset contains all the positive and negative reviews. We have considered approximately 1,000 reviews from times of india dataset 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 1.1.

B. Evaluation Measures

The easy way to calculate the accuracy is to validate the performance of the system by using the known sentiment words in dataset (times of india dataset is used). 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 class like positive, negative or neutral. 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 class such as positive, negative or neutral.

- *Performance Measures*

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

TABLE I. TABLE FOR CONFUSION MATRIX

Human/ Machine	Machine indicates yes	Machine indicates no
Human express yes	True Positive	False Negative
Human express no	False Positive	True Negative

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

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

$$\text{Precision} = \frac{\text{True positive elements}}{\text{True positive elements} + \text{False positive element}}$$

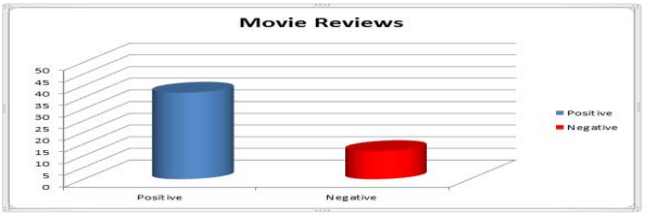

Fig.5: Movie review analysis result based on polarity.

C. *Comparison with previous model*

Many researchers have work on the domain of movie reviews, Pang et al. [2] have used three classification techniques NB, ME, SVM with the help of these technique they have achieve accuracy of 82.90%. Prabowo et al. [6] used hybrid SVM method for classification method it achieves accuracy of 87.3%. Rui Yao et al. [7] work on NB classification technique and achieve accuracy of 78.75%. Mullen and Collier et al. [8] performs sentiment classification using SVM and achieves accuracy of 86%. On comparing our result with the previous models we see that our approach achieves highest accuracy level than previous model used for classification of movie reviews.

TABLE II . COMPASISON WITH PREVIOUS MODELS

Authors	Classification Methods	Feature Selection	Accuracy
Pang et al. [2]	NB, ME, SVM	No	82.9%
Prabowo et al. [6]	Hybrid SVM	Yes	87.3 %
Rui Yao et al. [7]	NB	Yes	78.75%
Mullen and Collier et al. [8]	SVM	Yes	86%
Proposed Method	RF	Yes	90%

VI. CONCLUSION

Opinion mining has become popular research area due to the increasing number of internet users, social media etc. In this work, we extracted new features that have a strong impact on finding the polarity of the movie reviews. We then perform the feature impact analysis by estimating the information gain for each feature in the feature set and used it to derive a reduced feature set. The main goal of this work is to classify the sentences according to its sentiment by using Random Forest classification technique. This process of extracting the text having sentiment deals with finding the sentiment feature set from the sentences. As final output is displayed graphically it becomes easier for user to understand the exact polarity result.

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 combined approach." Journal of Informetrics 3.2 (2009): 143-157.

[7] Rui Yao and J. Chen "Predicting movie sales revenue using online reviews". In GeC, 2013, pp. 396-401.

[8] Mullen, Tony, and Nigel Collier. "Sentiment Analysis using SVM with Diverse Information Sources." *EMNLP* Vol. 4. 2004.

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

[10] 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.

[11] Godbole, Namrata, ManjaSrinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7 (2007): 21.

Gang Li, Fei Liu, "A Cluster-based Approach on Sentiment Analysis",IEEE 2010.