Abstract— Rapid flow in internet users along with increasing power of online review sites and social media has given existence to Sentiment analysis or Opinion mining, which aims to determine what other people feel, think, and express. Sentiment or Opinions contain user generated comment about products, services, policies and politics. Opinion may be in the form of ‘positive’ or ‘negative’. Users can give various opinion about feature of the product or services. Therefore product feature or aspects have got significant role in sentimental analysis. This review paper analyse existing techniques and approaches for feature extraction in opinion mining and sentimental analysis. In this paper we proposed the technique to extract the feature from the Movie review dataset. There is a burst of movie domain opinion rich resources in the form of review sites like IMDB, Yahoo movies etc. In this paper we proposed the method to provide the review summarization based on the feature of the movie commented by the user.

Keywords – Opinion mining, Sentiment Analysis, Machine learning, Feature Selection, Polarity, SentiWordNet.

I. INTRODUCTION

Opinion mining is the automated technique of extraction of attitudes, feelings or appraisal of the people about particular topic, product or services. Opinion mining technology makes it possible to aggregate the opinions of vast number of people. The information gathering process is the major part of aggregate the opinion of people. Sentiment is a view, feeling, opinion [1] which is expressed in the form of positive or negative. Sentiment analysis or opinion mining is a challenging text mining for automatic extraction, classification and summarization of sentiments and opinions expressed in online about product and services. There are many challenges in opinion mining [12]. The first one is that opinion word is not always considered positive or negative. In one condition it may be positive and in another condition it may be negative. Second challenge is opinion could be in the form of simple sentence or compound sentence. To deal with the compound sentence is more challenging. There is not enough work done in opinion mining of compound sentences.

Our research focused on movie reviews. There are large amount of user-generated movies review are available on the internet like IMDB, Yahoo, NDTV etc. There are many challenges like one or more bad feature of the movie does not make it overall bad same as one or more good feature does not make it good overall. Therefore opinion mining of movie review is considered more challenging than opinion mining of other type of reviews.

Feature based sentiment analysis include feature extraction, sentiment prediction, sentiment classification and summarization. Feature extraction identifies the products feature [3] which are commented by the user. Sentiment prediction identifies the word in the sentence containing sentiment or opinion based on sentiment polarity [10] in the term of positive, negative or neutral and finally provide the summarization based on the feature. [4] Feature extraction process takes text as input and provide the extracted feature in any of the forms like Lexico-syntactic or Stylistic, Syntactic and Discourse based [7].

In this work we proposed the method to find the opinions about movie based on the feature which is extracted from the user review. The various user review may be collected from various movie review sites for eg: www.rediff.com/movies/reviews, www.hindustan.com/movies-reviews/, www.bollywoodhungama.com etc. The review data set consisting of compound sentences. It will extract the opinion from the compound sentences. The compound sentences which is the collection of more than one sentences or clauses. In compound sentences, a single sentence may express more than one opinion about product or thing. For example, the sentence, ”The storyline is great, the director made good sense but the performance of the cast is not good”, represent both positive and negative opinions. For “storyline” and “director”, the sentence is positive, but for “cast”, it is negative. It is also positive for the movie as a whole.

Our system will ask for the feature of the movie like story, music, cast, direction etc. It then it searches for the compound sentences, which gives opinion to the asked features, based on different feature it divide compound sentence into single sentence and provide the opinion in the form of positive or negative of feature. If there is more then
one compound sentences referring the asked feature then it provide the opinion of all compound sentences individually and opinioted as positive or negative.

II. RELATED WORK
This section review the related work performed on opinion minning,feature extraction,feature classification in sentiment analysis. The opinion minning is associated with the information retrieval. The opinion minning is works on subjective data whether it is positive or negative. The concept of opinion minning is given by Hu and Liu[14].They provide basic component of an opinion are:

- **Opinion Holder:** It is the person which give opinion about thing.
- **Object:** It is the thing on which opinion is given by the user.
- **Opinion:** It is a view, sentiment, emotion what the user feel about the thing.

There are major feature extraction and manipulation techniques available which are summarized in below sections.

**Pre processing**
Pre-processing is the process of cleaning the data and preparing the text for classification. This process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. Features in the context of opinion minning are the words, terms or phrases that strongly express the opinion as positive or negative. In the preprocessing step,[5] first the sentence boundary is identified and then the text is tokenized. Extra white spaces, html tags, new lines and unrelated extra characters and special symbols are removed. Stop words are also removed as they do not belong to any of the four parts of speech (Noun, Adjective, Verb, and Adverb) present in the SentiWordNet[13] and they do not affect the opinion expressed in the document. The list of stop words used in this work excludes adverbs like very, more etc. and conjunctions such as and, but, etc. which can affect the subjective information of text. We parse the sentence through Stanford parser to determine part of speech of each word in sentence [10].

III. DIFFERENT LEVEL OF SENTIMENT ANALYSIS
Sentiment anlaysis are mainly divided into document level, sentence level and feature level/attribute level/aspect level/phrase level to find whether the given text is providing positive opinion, negative opinion or neutral. This is also known as ‘sentiment polarity prediction’. Hence sentiment analysis is carried out into three levels[11].

- **Document level**
- **Sentence level**
- **Feature level**

**Document level sentiment classification:**
It is classifying the opinionated text given by the user in whole document as positive, negative or neutral about a certain subject or object. Hence subjective or objective classification is necessary in document level classification. The problem arise in this classification when the informative text is to extract for deducing sentiment of the entire document. Pang et al. [6] present a work based on classic topic classification techniques. The proposed approach aims to test whether a selected group of machine learning algorithms can produce good result when opinion minning is perceived as document level, associated with two topics: positive and negative. He present the results using nave bayes, maximum entropy and support vector machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets. Turney [15] present a work based on distance measure of adjectives found in whole document with known polarity i.e. excellent or poor. The author presents a three step algorithm i.e. in the first step; the adjectives are extracted along with a word that provides appropriate information. Second step, the semantic orientation is captured by measuring the distance from words of known polarity. Third step, the algorithm counts the average semantic orientation for all word pairs and classifies a review as recommended or not.

**Sentence level sentiment classification:**
This type of classification refer to calculate the polarity of each sentence. The sentence level classification mainly focused on two things. First one is, to identify that the opinionated sentence is objective or subjective. The second one is, to identify the opinionated sentence is positive, negative or neutral[1]. Riloff and Wiebe [16] use a method called bootstrap approach to identify the subjective sentences and achieve the result around 90% accuracy during their tests. In contrast, Yu and Hatzivassiloglou [17] talk about sentence classification (subjective/objective) and orientation (positive/negative/neural). For the sentence classification, author’s present three different algorithms: (1) sentence similarity detection, (2) naïve Bayens classification and (3) multiple naïve Bayens classification. For opinion orientation authors use a technique similar to the one used by Turney [15] for document level. Wilson et al. [18] pointed out that not only a single sentence may...
contain multiple opinions, but they also have both subjective and factual clauses.

**Feature level sentence classification:**
The feature level sentence classification is a more pinpointed method to opinion mining. This type of classification mainly focused on feature of particular product or services. It gives the opinion based on the feature of the object. Analysis of the object based on their feature called as feature based sentiment analysis. It extract the feature of the object and conclude the opinion in the form of positive, negative or neutral [2][3]. Liu [11] used supervised pattern learning method to extract the object features for identification of opinion orientation. To identify the orientation of opinion he used lexicon based approach. This approach basically uses opinion words and phrase in a sentence to determine the opinion. Hu and Liu do customer review analysis [14] through opinion mining based on feature frequency, in which the most frequent features is accepted by processing many reviews that are taken during summary generation. Popescu and Etzioni [19], improved the frequency based approach by introducing the part-of relationship and remove the frequent occurring of noun phrases that may not be features.

**Opinion Mining On Movie Domain:**
The earliest work at document level [6] the authors used several machine learning approaches with common text features to classify movie reviews from IMDB. Dave et.al 2003 [3] designed a classifier based on information retrieval techniques for feature extraction and scoring. K-Denecke [13] performs opinion mining on movie review at document level. The author used SentiWordnet for word scoring. The score of words of whole documents are are accumulated to give final score. The rules are followed to calculate the score of all synsets and averaged to give final score. S. Agarawal [11] presents the summarization of the movie based on the feature of the movie. The author present the method to generate the ratings based on the individual feature of the movie. Liu [14] used supervised pattern learning method to extract the object features for identification of opinion orientation. To identify the orientation of opinion he used lexicon based approach. This approach basically uses opinion words and phrase in a sentence to determine the opinion.

**IV. PROPOSED METHODOLOGY**
In this section we focus on opinion of a movie review that gives the opinion based on the individual feature of the movie and also determine the sentiment score based on various feature of a movie, such as cast, director and music. Sentiment scores are used to classify the sentiment polarity (i.e. positive, negative or neutral) of clauses or sentences. We use SentiWordnet [5] scores to each sentence according to individual feature of the movie. We use Stanford NLP to POS tagging and text preprocessing. The following steps in our approach are discussed below:

**Document Preprocessing using StanfordNLP**
In the document preprocessing, first the sentence area is delimited and then the text is tokenized. Then some unnecessary things have to remove such as extra white spaces, HTML tags, new lines, extra characters and special symbols. The words that does not belong to any of the four parts of speech (Noun, Adjective, Verb and Adverb) and does not affect the opinion expressed in the document, those stop words are also removed. The stopwords excludes adverb like very, more etc. and conjunction such as and, but, etc. which can affect the subjective information of text. We use Stanford parser to parse the sentence to determine the part of speech of each word in sentence [11].

**Based on the feature splitting the document into sentences and clauses**
Select the document which is movie review generated by user. By the help of sentence delimiter the document is split into individual sentences. The most of the reviews are available on movie forums or blog sites where users post their opinion in informal language which dose not follow any grammatical rules. We use rule based pattern matching to identify sentence boundary.

**Classification of sentences**
We identify the sentences in review. Review may be in simple sentences or compound sentences [2]. A compound sentences contains two or more sentences or clauses that are related. These two or more sentences or clauses are usually connected by a conjunction. The conjunctions are used such as ‘and’, ‘but’, ‘for’, ‘or’, ‘nor’, ‘yet’, ‘so’. We use plain pattern matching to find the presence of coordinating conjunction. We split the compound sentences into sentences and identify the feature of the movie through pattern matching. The boundary of sentences identified by the punctuation marks such as comma, semicolon, full stop or coordinating conjunction. Hence we get the individual feature of each sentence and calculate the score of each sentence based on the feature of the sentence. Finally aggregate the score of each sentence to give final score. We follow average scoring method to compute the score of individual feature.

**Word Scoring**
Each word in the document that is present in the SentiWordnet is assigned a positive, negative and objective...
score. The positive score is calculated as the average of the positive score of all the synsets in movie review document in SentiWordnet. The negative score is calculated in the same way. Those words are not present in SentiWordNet are assigned zero for both positive and negative scores.

**Feature Based Scoring**

Feature based scoring is computed by taking the average of the scores of sentences or clauses related to feature.

\[
\text{SenPosScore}(PS) = \frac{1}{n} \sum_{i=1}^{n} \text{PosScore}(i)
\]

\[
\text{SenNegScore}(NS) = \frac{1}{n} \sum_{i=1}^{n} \text{NegScore}(i)
\]

SenPosScore(PS), SenNegScore(NS) are the positive and negative respectively of sentence S or clause S. PosScore(i), NegScore(i) are the positive, negative score respectively of ith word in sentence S or clause S. 

\(N=\text{Total No. of words in S.}\)

The score of sth sentences or clauses SenScore(S) of feature F calculated as:

\[
\text{SenScore}(S) = \frac{1}{n} \sum_{i=1}^{n} \text{PosScore}(i) + \frac{1}{n} (- \sum_{i=1}^{n} \text{NegScore}(i))
\]

\[
\text{FeatureScore}(F) = \frac{1}{n} \sum_{s=1}^{n} \text{SenScore}(S)
\]

Where FeatureScore(F) is score of Feature(F) and ‘n’ is number of sentences (s) or Clauses (s) which expresses opinion on feature(F).

If feature score(F) is positive, then it is positive opinion.

If feature score(F) is negative, then it is negative opinion.

**V. RESULT**

We evaluate proposed method and analyse the result. The movie dataset which contain more than two hundred words and contain phrases or clauses, which is evaluated using proposed method and also evaluated by SentiWordNet approach and analyse the result for both. Proposed methodology provide better result and 20% more accuracy than SentiWordNet approach. That is shown below:

<table>
<thead>
<tr>
<th>Polarity Based on the Feature of The Movie Review</th>
<th>Proposed Method</th>
<th>SentiWordNet Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>80%</td>
<td>60%</td>
</tr>
</tbody>
</table>


