International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-3, Issue-9, Sept- 2016]*
https://dx.doi.org/10.22161/ijaers *ISSN: 2349-6495(P) | 2456-1908(O)*

Feature Based Opinion Mining on Movie Review

Tanya Shruti, Manish Choudhary

Student, Computer science & Engineering Department,YIT Jaipur, Rajasthan, India
Asst. professor, Computer science & Engineering Department,YIT Jaipur, Rajasthan, India

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 minning,which aims to determine what other people feel,think And Exprss.Sentiment or Opinions contain user generated comment about products,services,policies andPolitics.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 Minning 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 siteslike 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 minning, Sentiment Analysis, Machine learning, Feature Selection, Polarity, SentiWordNet.

I. INTRODUCTION

Opinion minning is the Automated technique of extraction of attitudes,feelings or appraisal of the people about particular topic,product or services.Opinion minning technology make it possible to aggregate the opinions of vast number of peoples.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 minning is a challenging text minning for automatic extraction.classification and summarization of sentiments and opinions expressed in online about product and services.There are many challenges in opinion minning[12].The first one is that opinion word is not always considerd 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,NDTVetc.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 minning of movie review is considered more challenging than opinion minning 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 containinf 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 positine or negative of feature.If there is more then

one compound sentences referring the asked feature then it provide 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 andfinally feature selection.Features in the context of opinion minning are the words,terms or pharases 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/pharase 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].

I. Document level
II. Sentence level
III. 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 sentimecnt 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 mining 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/neutral). 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 sentiment classification is a more pinpointed method to opinion minning.This type of classification mainly focused on feature of particular product or services.It give 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 Minning 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 minning 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 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 onopinion 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,directory 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 bu user.By the help of sentence delimiter the document is splitted into individual sentencesThe most of the reviews are available on mmovie forums or blog sites where userspost their opinion in informal language which dose not follow any grammatitical 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 same way.Those word are nor present in SentiWordNet are assigned zero for both positive and negative scores.

Feature based Scoring

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

$$SenPosScore(PS) = \frac{1}{n}\sum_{i=1}^{n} PosScore(i)$$

$$SenNegScore(NS) = \frac{1}{n}\sum_{i=1}^{n} NegScore(i)$$

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=Total No.of words inS.

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

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

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

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 pharases 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:

Polarity Based on the feature of The movie review	**Proposed Method**	**SentiWordNet Approach**
Accuracy	**80%**	**60%**

VI. CONCLUSION

- We find that some incomplete meaningless sentences or clauses are presented to the user as answer. It happens asour sentence segmentation based on rules is not proper. This is our future work to break sentence through machine learning methods.
- identification of feature is a tough task. Co reference resolution has also affected our method. In future,we will address this issue.
- Even different aspects of movie has different sub features hence segmentation based on sub feature is required for the opinion mining.
- We use SentiWordnet, general opinion lexicon dictionary for the purpose of opinion mining at movie domain.Hence domain specific dictionary could be more appropriate.

REFERENCES

[1] Philip,Beineke,TrevorHastie,Christopher- Manning and Shivakumar Vaithyanathan. An exploration of sentiment summarization. In 2003Proceedings of AAAI 2003,pp.12-15**.**

[2] Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In Proceedings of HICSS 2005, vol.4.

[3] Kushal Dave, Steve Lawrence and David M. Pennock.2003. Mining the peanut gallery: Opinion extraction and semantic classi¯cation of product reviews. In Proceedings of WWW 2003, pp.519-528.

[4] A Review of Feature Extraction in Sentiment Analysis.Muhammad Zubair Asghar1, Aurangzeb Khan2, Shakeel Ahmad1, Fazal Masud Kundi .ISSN 2090-4304Journal of Basic and Applied Scientific Research 2014.

[5] Information Technology and Quantitative Management (ITQM2013)The Role of Text Pre-processing in Sentiment AnalysisEmma Haddia, Xiaohui Liua, Yong Shib

[6] Bo Pang, Lillian Lee and Shivakumar Vai2002. Thumbs up? Sentiment classification using machine learning techniques.In Proceedings of EMNLP 2002,pp.79-86.

[7] ISSN 2090-4304,2014,Journal of Basic and Applied Scientific Research, Author: Muhammad Zubair Asghar,et al.A Review of Feature Extraction in Sentiment Analysis.

[8] S.Agrawal and T.J.Siddiqui, "Using syntactic and Contextual Information for Sentiment Polarity

Analysis" in Proceeding of ICIS,Seoul, Korea November 2009.

[9] B.Liu 2010. Opinion Mining and Sentiment Analysis: NLP Meets Social Sciences", STSC, Hawaii.

[10] G.Vinodhini and RM. Chandrasekaran 2012. Sentiment analysis and Opinion Mining:Asurvey International Journal of advanced Research in Computer Science and Software Engineering vol. 2Issue

[11] K. Denecke. "Using SentiWordNet for Multilingual Sentiment Analysis," in Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs,Social,Media,and,Web-2.0,Cancun,.2008

[12] M. Hu and B. Liu 2004. Mining and summarizing customer reviews,Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177.

[13] P.Turney 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceeding of Association for Computational Linguistics, pp. 417--424.

[14] E. Riloff, and J. Wiebe, 2003. Learning Extraction Patterns for Subjective Expressions, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan, Sapporo

[15] H. Yu, and V. Hatzivassiloglou, 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan, Sapporo

[16] T.Wilson,., J. Wiebe,., R. Hwa,. 2004. Just how mad are you? Finding strong and weak opinion clauses. In: the Association for the Advancement of Artificial Intelligence, pp. 761--769.

[17] A.M..Popescu,, O. Etzioni, 2005. Extracting Product Features and Opinions from Reviews, In Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, pp. 339–346.