Opinion Analysis Algorithm Using Mathematical Rule-Based Approach

Ms Roshni Pawar[1], Mr. Sachin Malviya[2]
[1]Dept of Computer Engineering
[2]Assistant Professor, Dept of Computer Engineering
[1, 2] SBITM, Betul, Madhya Pradesh, India

Abstract- *Opinion Analysis notion is to provides extensive contribution in Natural Language Processing (NLP) which promises with the computational measures of sentiment, subjectivity and objectivity in the given text. Opinion analysis is the process of extracting desired knowledge from the people's opinions, appraisals and emotions toward entities, events and their attributes. These opinions notably make impact on consumers to make their preference regarding, choosing products, watching movies and buying entities. As a result, it is desired to develop an efficient and effective openion analysis system for customer reviews and comments. We consider the sticky situation of determining the polarity of sentiments in reviews when negation words occur in the sentences.*

In this work we use SentiWordNet dictionary to assigns sentiment scores to each word found in user opinion comments. Sentiment words are assigned sentiment scores: Positivity, Negativity and Objectivity with a word which lies in between the range from 0 to 1. The final review prediction uses Rule-Based mathematical measures approach and gives the final output.

Keywords- SentiWordNet, Natural Language Processing, Sentiment Analysis System, Fuzzy measures, Web Opinion Mining ,Text Tokenization.

I. INTRODUCTION

Today the Internet holds an enormous amount of textual data, which is also growing every day. The text is in ubiquitous format on the web, since it is easy to generate and publish. People communicate through online resources, discussion forums, groups and blogs. What is hard now-a-days is not accessibility of valuable information but rather extracting it in the appropriate context from the vast ocean of text content. It is now beyond human supremacy and time to see through it manually and therefore, the research problem of automatic categorization and organizing text is perceptible. Textual information on web can be separated into two main fields: facts and opinions. While facts focus on objective data transmission, the opinions express the sentiment of their person behind. Currently Google searches for facts and facts can be expressed with keywords; but Google search does not discover opinions, because opinions are hard to articulate with keywords.

What Is Opinion Analysis?

Opinion analysis is a "Sentiment Mining", where the primary objective is to classify the opinions according to a variety and range. The boundaries on the range usually correspond to +ve or -ve feelings about a product or brand which in fact determines sentiment orientation of an individual or a group of people. Current ranking strategies are not appropriate for opinion mining. Primarily the research has mostly persistent on the classification of the text data. Sentiments are naturally subjective from individual to individual, and can be absolute illogical. It's critical to analyze relevant sample of data when attempting to measure sentiment. No particular data point is necessarily relevant. An individual's sentiment toward a brand or a product may be inclined by one or more mind and someone might have a terrible day and tweet a -ve remark about something they otherwise had a pretty neutral opinion.

With large enough samples, outliers are diluted in the aggregate. Also since sentiment are very likely to changes over the time according to a person's mood, world events, and so forth, it's usually important to look at data from the perspective of time. Like any other type of Natural Language Processing (NLP) analysis, context matters, it's an incredibly difficult issue, sarcasm and other types of ironic language are inherently problematic for machines to detect when looked at in isolation. It's imperative to have a sufficiently sophisticated and rigorous enough approach that relevant context can be taken into account. That would require knowing a particular person is ironic, exaggerated or sarcastic which offers evidence to conclude whether or not a phrase is ironic.

The focus of the system is to analyze the sentiments for the movie reviews. The input is to be taken from the movie review sites and the social networking sites on which the comments are posted for the particular movie. The SA is done

by three methods i.e. Rule-Based approach, Machine-Learning and the Hybrid Approach of the two. Rule-Based is the manually created Rules approach. Machine-Learning give the more efficient approach as compared to the Rule-Based approach. The Hybrid Approach of both above approaches will give the more superior and filtered outcome of the analysis for the movie reviews.

II. LITERATURE REVIEW

C. Hauff gives the way how to handle the negation words like not, no, neither, couldn't, etc words in the sentence. It may happen that even if the negative words are present in the sentence still its meaning is in a positive way[5].

A. Neviarouskaya performs fine grained categorization of sentences using ten categories: nine emotions ('Anger', 'Disgust', 'Fear', 'Guilt', 'Interest',' Joy', 'Sadness' ('distress'), 'Shame' and 'Surprise') and neutral. The proposed rule-based approach processes each sentence in stages, including symbolic cue processing, detection and transformation of abbreviations, sentence parsing and word/phrase/sentence-level analyses. Each analyzed sentence is automatically annotated with emotion or neutral label and numerical intensity; the strength of emotion. It also mentioned the links where the datasets would be available[7].

B. Tierney presents the results of applying the SentiWordNet lexical resource to the problem of automatic sentiment classification of film reviews. It comprises counting +ve and -ve term scores to determine sentiment orientation, and an improvement is presented by building a data set of relevant features using SentiWordNet as source, and applied to a machine learning classifier[8].

M. Thelwall gives the hybrid knowledge for involving the Rule-Based and Support Vector Machines method. The hybrid approach gives the maximum efficiency for analysis of sentiments[9].

S. Kawathekar studies the role of negation in an opinion-oriented information-seeking system. They investigate the problem of determining the polarity of sentiments in movie reviews when negation words, such as not and hardly occur in the sentences[1].

A. Shukla presented a tool which tells the quality of document or its usefulness based on annotations. Annotation may include comments, notes, observation, highlights, underline, explanation, question or help etc. Collective sentiments of annotators are classified as positive, negative, objectivity[3].

Recent research has tried to automatically determine the "PN-polarity" of subjective terms F. Sebastiani in order to aid the extraction of opinions from text, i.e. identify whether a term that is a marker of opinionated content has a positive or a negative connotation[10].

P. Bhattacharyya proposed a technique for the effective SA of movie reviews. It also describe a novel approach to process the predictions for individual documents of the test dataset to improve the accuracy over the entire set. It presented a WorldNet based method for the effective incorporation of linguistic information in the system without any kind of experts' intervention. It also presents a generic method that can be used to improve the accuracy of classification over a test dataset in any kind of classification task. It shows how the application of this technique to SA helps to attain the best accuracy so far in this field[11]

III. IMPLEMENTATION FRAMEWORK

The development of a complete review or opinion mining application might involve attacking each of the following problems. If the application is integrated into a general-purpose search engine, then one would need to determine whether the user is in fact looking for subjective material. This may or may not be a difficult problem in and of it: perhaps queries of this type will tend to contain indicator terms like "review", "reviews", or "opinions", or perhaps the application would provide a "checkbox" to the user so that he or she could indicate directly that reviews are what are desired. Besides the still-open problem of determining which documents are topically relevant to an opinion-oriented query, an additional challenge we face in our new setting is simultaneously or subsequently determining which documents or portions of documents contain review-like or opinionated material. Sometimes this is relatively easy, as in texts fetched from review-aggregation sites in which review-oriented information is presented in relatively stereotyped format: examples include Epinions.com and Amazon.com. However, blogs also notoriously contain quite a bit of subjective content and thus are another obvious place to look and are more relevant than shopping sites for queries that concern politics, people, or other non-products, but the desired material within blogs can vary quite widely in content, style, presentation, and even level of grammaticality.

Once one has target documents in hand, one is still faced with the problem of identifying the overall sentiment expressed by these documents and/or the accurate opinions regarding particular features or aspects of the items or topics in question. Again, while some sites make this kind of extraction easier for e.g., user reviews posted to Yahoo!

Movies must specify grades for predefined sets of distinctiveness of films more free-form text can be much harder for computers to analyze, and indeed can pose additional challenges; for e.g., if quotations are included in a newspaper article, care must be taken to attribute the views expressed in each quotation to the correct entity.

Rule Based Mathematical Approach Algorithm:

Algorithm 1 Find the total count of review sentences in the document.

Input: Review document comprises of different opinions

Output: Total Count of review sentences present in the document

1. **for** each review document **do**

$$\text{Count} = \sum_{I=0}^{N} \text{Openion}$$

2. **End for.**
3. Output Opinion Count.

Algorithm 2 Find the average score of each review found in document

Input: List of sentiment words extracted from comments of input review document

Output: Sentiment score and Sentiment Review

1. **for** each sentiment word from List **do**

Get polarity as well as sentiment scores using *SentiWordNet*.dictionary

2. Compute the polarity(+/-) intensity of each word using Rule based mathematical algorithm.
3. Find ultimate Sentiment Score

$$\text{Sentiment Score} = \sum_{I=0}^{N} \frac{\text{Max(Polarity)}}{\text{Count}}$$

4. Output Sentiment Score.
5. Output Overall Sentiment Review.

RULE BASED ANALYISIS WITH MATHEMTICAL ALGORITHM

We calculate the weight of the extracted opinionated phrases as the weights of individual words in the phrases.

Case 1: There are few adverbs like very, really, extremely, simply, always, never, not, absolutely, highly, overall, truly, too, etc. (we imply 15 such adverbs) which may be used positively or negatively like {very good, very bad}, {extremely acceptable, extremely unacceptable}, {too good, too bad}, {simply outstanding, simply disgusting} etc.

Case 2: Again never, not etc. will change the orientation of the opinion like {not good will be bad} {never accepted will be always rejected} etc.

Case 3: Case 1 and Case 2 may also come together in opinion phrases like {not very good}, {not absolutely recommended}, {not truly reliable one} etc. We consider them accordingly in the calculation of fuzzy weights of opinion phrases.

Case 4: It is also possible that in some cases like {It is not only good but also it's an awesome} etc. appears in sentiments.

In Case 1, we consider the weight (W) of the opinion like (very /extremely/highly etc) (Adj)

$$= \sqrt{\text{Value Of (Adj)}} \quad \text{if ValueOf(Adj)} \geq 0.5$$
$$= (\text{Value Of (Adj)})^2 \quad \text{if ValueOf(Adj)} < 0.5$$

Table 1: Fuzzy Measure of (Adverb, Adjective) Phrases

Good	0.625	Very good	0.7906
Bad	0.25	Very bad	0.0625
Awesome	0.875	Simply Awesome	0.9354
Pathetic	0.375	Highly Pathetic	0.1406

In Case 2, we consider the weight (W) of the opinion like (not/never)(Adj /Verb)

= 1 - ValueOf(Adj/Verb)

Table 2: Fuzzy Measure of (Not/Never, Adjective) Phrases

Good	0.625	Not good	0.375
Bad	0.25	Not bad	0.75
Conclusive	1.0	Never Conclusive	0.0
Stunning	0.75	Never Stunning	0.25

In case 3, we consider the weight (W) of the opinion like (not/never) (very /extremely/highly etc)(Adj)

$$= \sqrt{(A * B)}$$

Where A = (very/extremely/highly etc) (Adj)

$$= \sqrt{\text{Value Of (Adj)}} \quad \text{if ValueOf(Adj)} \geq 0.5$$

$= (\text{Value Of } (\text{Adj}))^2$ if ValueOf(Adj) < 0.5

And B = (not/never) (Adj)

Table 3: Fuzzy Measure of (Not, Adverb, Adjective) Phrases

Good	0.625	Not Good	0.375	Not Very Good	0.5445
Bad	0.25	Not Bad	0.75	Not Very Bad	0.2165

In case 4, we consider the weight (W) of the opinion like (but/also/nor) words. Firstly we divide the sentiment into two opinions group. Then these two opinions separately calculated as follows. Finally the average of both cases will be taken as a Fuzzy measure for the given sentence.

F= $\sqrt{(\sqrt{A}+\sqrt{B})/2}$ if ValueOf(Any Adj) ≥ 0.5

$\mathbf{F = ((A^2 + B^2)^2)/2}$ if ValueOf(Both Adj) < 0.5

Table 4: Fuzzy Measure of (But, Also) Phrases

Good	0.625	Awesome	0.875
$\sqrt{\text{Good}}$	0.7906	$\sqrt{\text{Awesome}}$	0.9354
This is a not only good but also awesome			0.9289
Bad	0.25	Pathetic	0.375
Bad^2	0.0625	Pathetic^2	0.1406
This is not only bad but it's pathetic			0.0206

IV. RESULT

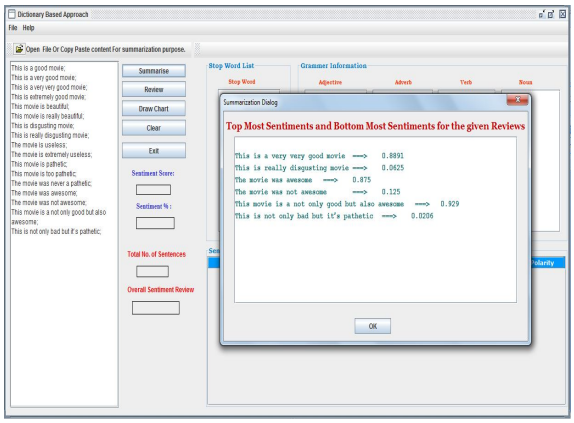

Figure 3: Summary for Topmost and Bottom Most Sentiments

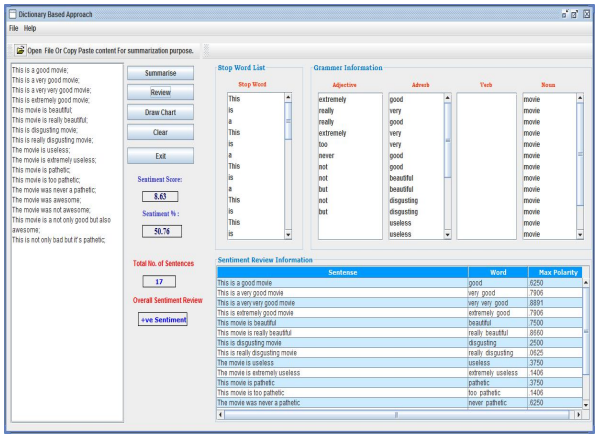

Figure 4: Preprocessing, Classification and POS tagging

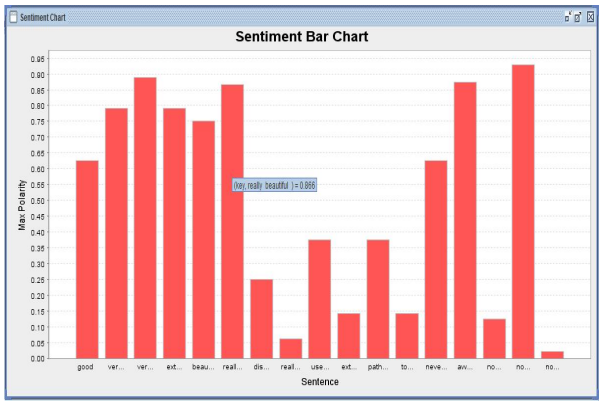

Figure 5: Fuzzy measure graph for SA

So, if the above value is positive (i.e. >= 50%) then, sentiment of document is positive, otherwise negative.

Here, sentiment of collective annotator over document is positive.

V. ADVANTAGES

- Here we take direct, unlimited, unbiased as well as real-time opinions of legitimate users.
- It provides cost effective approach of capturing user feedback.
- Uninterrupted with wide geographic reach, real time feedback.
- This approach provides much better reaction time for service and quality improvement for market.

VI. CONCLUSION

Implementation of Opinion based system provides major benefits to improve current market strategies so that maximum benefits can be achieved by applying proper and market suitable strategies. The major opinions sources regarding film or any product are Web, Social Networking

web-sites for instance, Facebook , Orkut and Twitter and other many web services from where subject related information can be gathered.

There are various challenge, more companies and researchers are working in this area until one day it would be easy for users and companies to minimally obtain complete and wealthy summarized fact about the opinions from the web in order to uphold them in the decision making process in their daily life.

This research work gives idea regarding how document based sentiment analysis implementation can be done by applying rule based and fuzzy system to create decision making system.

REFERENCES

[1] Swati A. Kawathekar and Dr. Manali M. Kshirsagar, *Movie Review analysis using Rule-Based & Support Vector Machines methods*, IOSR Journal of Engineering, Vol. 2(3) pp: 389-391, March 2012.

[2] Alaa Hamouda, Mahmoud Marei and Mohamed Rohaim, *Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis*, Journal of Advances In Information Technology, Vol. 2, No. 4, pp 199-203, November 2011.

[3] Archana Shukla, *Sentiment Analysis Of Document Based On Annotation*, International Journal of Web & Semantic Technology (IJWesT) Vol.2, No.4, pp 91-103, October 2011

[4] Aurangzeb Khan, Baharum Baharudin and Khairullah Khan, 2011. Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews. Trends in Applied Sciences Research, Vol. 6, pp. 1141-1157, July, 2011.

[5] C. Hauff, Dadvar, Maral and Jong de, Franciska, *Scope of negation detection in sentiment analysis*, Dutch-Belgian Information Retrieval Workshop, Netherlands, February 2011.

[6] Animesh Kar, Deba Prasad Mandal, *Finding Opinion Strength Using Fuzzy Logic on Web Reviews*, International Journal of Engineering and Industries, volume 2, Number 1, pp 37-43, March, 2011

[7] Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka, *Affect Analysis Model: novel rule-based approach to affect sensing from text*, Natural Language Engineering, Cambridge University, Vol. 17, pp. 95- 135, September 2010.

[8] Brendan Tierney and Bruno Ohana, *Sentiment Classification of Reviews using SentiWordNet*, 9th IT&T Conference, Dublin Institute of Technology, Ireland, October, 2009.

[9] Mike Thelwall and Rudy Prabowo, *Sentiment Analysis: A Combined Approach*, Journal of Informatics**,** School of Computing and Information Technology, University of Wolverhampton, UK, Volume 3, Issue 2, pp. 143-157, April 2009.

[10] Fabrizio Sebastiani and Andrea Esuli, *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining*, Proceedings of the 5th Conference on Language Resources and Evaluation, Genoa – Italy, pp. 417-422, May 2006.

[11] Pushpak Bhattacharyya, Akhel Agrawal, *Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified*, International Conference on Natural Language Processing, IIT Kanpur, India, December, 2005.

[12] Bo Pang, Lillian Lee, and S. K. Vaithyanathan, *Thumbs up? Sentiment Classification using Machine Learning Techniques***,** Proceedings of the Conference on Empirical Methods in Natural Language Processing, Volume 10, pp. 79-86, 2002.

[13] Gang Li and Fei Liu, *A Clustering-based Approach on Sentiment Analysis*, International Conference on Intelligent Systems and Knowledge Engineering, Hangzhou, pp. 331-337, November 2010.

[14] Jianxiong Wang and Andy Dong, *A Comparison of Two Text Representations for Sentiment Analysis*, International Conference on Computer Application and System Modeling, Taiyuan, China, pp. 35- 39, October 2010.

[15] Aditya Joshi, Balamurali AR, Pushpak Bhattacharyya and Rajat Mohanty, *C-Feel-It: A Sentiment Analyzer for Micro-blogs*, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 127-132. June, 2011.

[16] Siva RamaKrishna Reddy V., DVLN. Somayajulu and Ajay R. Dani, *Classification of Movie Reviews Using Complemented Naive Bayesian Classifier*, International Journal of Intelligent Computing Research, UK, Volume 1, Issue 4, pp. 162-167, December 2010.

[17] Adnan Duric and Fei Song, *Feature Selection for Sentiment Analysis Based on Content and Syntax Models*, Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon, USA, pp. 96-103, June, 2011.

[18] V. Rentoumi, S. Petrakis, M. Klenner, G. A. Vouros and V. Karkaletsis, *United we stand: improving sentiment analysis by joining machine learning and rule based methods*, 7th International Conference on Language Resources and Evaluation, Malta, pp. 1089- 1094, May 2010.

[19] Julia Maria Schulz, Christa Womser Hacker and Thomas Mandl, *Multilingual Corpus Development for Opinion Mining*, Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 3409-3412, May 2010.

[20] Xiaoxu Fei, Huizhen Wang and Jingbo Zhu, *Sentiment word identification using the maximum entropy model*, International Conference on Natural Language Processing and Knowledge Engineering, Beijing, pp. 1-4, September 2010.