International Journal For Technological Research In Engineering
Volume 4, Issue 8, April-2017 *ISSN (Online): 2347 - 4718*

NOVEL ALGORITHM FOR SPAM DETECTION AND SENTIMENT AND FEATURE LEVEL ANALYSIS OF REVIEWS

Rishabh Chand Sharma[1], Manish Mathuria[2]
[1]M.Tech. Scholar, [2]Head of CS Dept, Maharishi Arvind College of Engineering & Research Center

Abstract: The Review submission is the normal process when we talk about the social media like facebook , google+ etc… These reviews are very useful for companies for improving the products specifications and quality. But always the reviews are not of use as the internet is also in attack by spammers and people writing spam reviews. So the main problem is to filter these reviews. In our dissertation ,we have proposed the novel algorithm for detection of the spam reviews and together with the detection of the spam reviews , we have extended our work in classification of the reviews like positive reviews and negative reviews. In our work we have used spam and ham dataset for the identification of the reviews as spam or ham. And after the normal review is detected then using the sentiment analysis we have classified the reviews.

I. INTRODUCTION

It has become a common practice for people to read online opinions/reviews for different purposes. For example, if one wants to buy a product, one typically goes to a review site (e.g., amazon.com) to read some reviews of the product. If most reviews are negative, one will almost certainly not buy it. Positive opinions can result in significant financial gains and/or fames for businesses, organizations and individuals. This, unfortunately, gives strong incentives for opinion spamming.

1.1 Opinion Spamming

It refers to "illegal" activities (e.g., writing fake reviews, also called shilling) that try to mislead readers or automated opinion mining and sentiment analysis systems by giving undeserving positive opinions to some target entities in order to promote the entities and/or by giving false negative opinions to some other entities in order to damage their reputations. We believe that as opinions on the Web are increasingly used in practice by consumers, organizations, and businesses for their decision making, opinion spamming will get worse and also more sophisticated.

1.2. Feature Level Analysis:

Both the document level and the sentence level analyses don't discover what exactly people liked and did not like. Aspect level performs finer-grained investigation. Aspect level was earlier called feature level (feature-based conclusion mining and summarization).Instead of taking a gander at language builds (documents, sections, sentences, clauses or phrases), aspect level directly takes a gander at the assessment itself. It is based on the idea that an assessment consists of a sentiment (positive or negative) and a target (of conclusion).

1.3. Sentiment Classification Techniques

In general, Sentiment Classification should be possible with three techniques machine learning (ML) approach, lexicon based approach and hybrid approach.

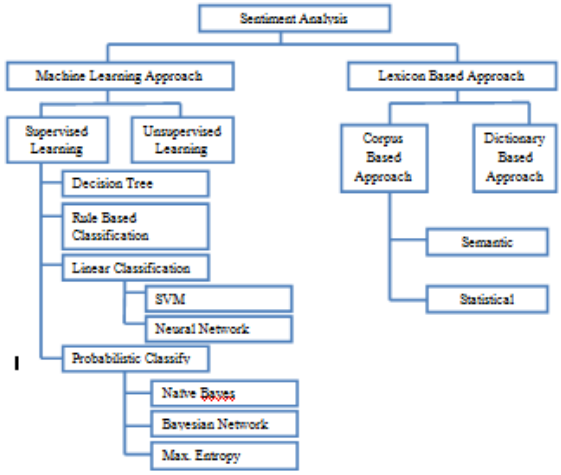

Fig. 1 Sentiment Classification Techniques

II. LITERATURE SURVEY

(Jindal & Liu, 2008). [1] The opinion spam problem was initially detailed by Jindal and Liu with regards to item reviews, (Jindal and Liu, 2008). By dissecting a few million reviews from the famous Amazon.com, they indicated how widespread the problem of fake reviews was. The current recognition methods can be part with regards to machine learning into managed and unsupervised approaches. Second, they can be part into three classifications by their features: behavioral, linguistic or those utilizing a blend of these two. They arranged spam reviews into three classifications: non-reviews, brand-just reviews and untruthful reviews. The creators ran a strategic relapse classifier on a model prepared on copy or close copy reviews as positive preparing data, i.e. fake reviews, and whatever is left of the reviews they utilized as honest reviews. They joined reviewer behavioral features with printed features and they meant to exhibit that the model could be generalized to identify non-copy review spam. This was the initially recorded research on the problem of opinion spam and consequently did not profit by existing preparing databases. The creators needed to assemble their own dataset, and the least difficult approach was to use close copy reviews as cases of misleading reviews. Despite the fact that this underlying model indicated great outcomes, it is as yet an early examination concerning this problem. Nidhi Mishra & C. K. Jha [14] "Opinion Mining from Text in Movie Domain" International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 4, Oct 2013. They computed opinion of the feature of movie such as story, star cast, direction etc. and present the

related text fragment to the user. The authors discussed about some existed research work as many search engine retrieved facts through keyword matching, popularity etc.

III. PROBLEM DESCRIPTION

This chapter incorporates the issue definition and the destinations alongside the subtle elements of the procedures utilized as a part of the proposed work.

Objectives

- Spam Filter of the reviews
- To gather the reviews for motion picture space from various social locales.
- To perform information pre-preparing as
  - Tokenization,
  - Stop word evacuating
  - Stemming and grammatical feature tagging on gathered reviews for information readiness.
- To concentrate every one of the components from the reviews and store in the database.
- To decide the polarity of the basic sentence and compound sentences at highlight
- level utilizing proposed calculation.

To look at our proposed approach utilizing existing sentiment examination device (Opinion Finder, SentiWordNet, and WordNet spread).

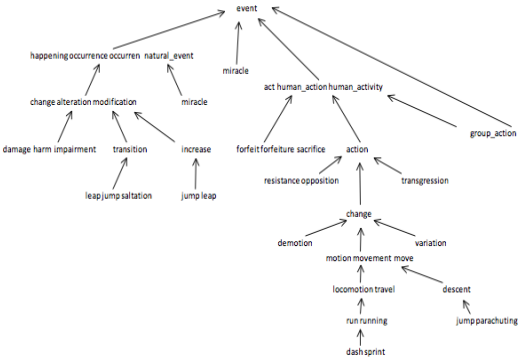

Fig 2 WordNet

IV. PROPOSED METHODOLOGY

4.1 Algorithm Adopted For Spam Detection

Step1 : Read the document file containing the Review

Step 2: Read the dataset containing the HAM keywords.

Step 3: Read the dataset containing the SPAM keywords.

Step 4: Analyze the document for the SPAM and HAM keywords on the basis of the occurrences.

4.2 Algorithm Adopted For Classification

Proposed model for Movie Review Analysis using the Sentiments scoring based on SentiWordNet and feature level analysis of the documents as shown in figure 4.1

Step 1: Input

Input Review document for generating review result that whether the review is positive or Negative (.txt file).

Step 2: Segmentation

Divide the document into sentences using segmentation.

Step 3: Feature Level Filtering

We will read the file is the feature is all then all the files will consider for the review analysis otherwise the filtration is done. And the algorithm is shown in Algorithm-1 [22].

ISFEATURE (line, feature)

[Perform feature level extraction]
1. Check for the Feature and its Synonyms.
2. If the line contains words which are feature itself or its synonyms
3. Return the line granted for review analysis.
[End of for loop]

Algorithm-1: Feature Level Filtering Algorithm

Step 4: Tokenization

Each sentence is divided into tokens.

An example: Friends, has been colouring and roman lend me, your field; Hence after tokenization we get: Friends has been colouring and roman lend me your field. Basically we need to omit the commas, punctuations, (carefully apostrophes), question marks etc [23].

Step 5: POS tagging

The pos tagging is the tagger which specify the token as nouns, verbs, adverbs, adjectives [24] show in Algorithm-2.

STEMTAGG (line)

[Perform stemming and then tagging]
1. Using the WORDNET API, the base for of words is obtained
2. Tagged the line using the MAXENTTAGGER class.
3. Return the finally tagged line.

Algorithm-2: Stemming & Tagging Algorithm

Step 5: Splitting the Line and processing for line score.

In this we will split the line into the array and we will find the score of the each word in order to get the score for the complete sentence. And this score will further led to the scoring for the entire document. The scoring is done using the SentiWordNet and here we will use the SentiWordNet library for getting the scoring as described in the Algorithm-3.

SENTISCORE (Word, POS_tagg)

[Perform scoring based on SentiWordNet]
1. Call the sdata.txt library file for SentiWordNet scoring.
2. Call extract (word, tag) for getting the score for the word.
3. Return the score.

Algorithm-3: Scoring Word using SentiWordNet

Step 6: Intensifier Handling

In this we will examine that the word which we are reading from the line is intensifier or not, if it is intensifier then the score is to be handled accordingly, as the intensifier will further enhance the score show in Algorithm-4.

CHECKINTENSIFIER (Word)

[This algorithm will check whether the word is the intensifier or not.]
1) Read the File "intensifier.txt" into fstream
2) Repeat till EOF (End of File)

3) Read INTENSIFIER
4) If Word is INTENSIFIER then :
Return true
Else
Return false
[End of If structure]
[End of inner for loop]

Algorithm-4: Check for Intensifier

Step 7: Negation Handling

In this we will examine that the word is the negative word, as the negative word will negate the score of the coming word show in Algorithm-5.

CHECKNEGATION (Word)

[This algorithm will check whether the word is the negation or not.]
1) Read the File "negation.txt" into fstream
2) Repeat till EOF (End of File]
3) Read NEGATION
4) If Word is NEGATION then :
Return true
Else
Return false
[End of If structure]
[End of inner for loop]

Algorithm-5: Check for Negation

Step 8: Conjunction Handling

In this we will examine that the word is the conjunction word, as the conjunction word will split the sentence into two different sentences and the working of conjunction handling algorithm is shown in Algorithm-6.

CHECKCONJUNCTION (Word)

[This algorithm will check whether the word is the conjunction or not.]
1) Read the File " conjunction.txt" into fstream
2) Repeat till EOF (End of File]
3) Read CONJUNCTION
4) If Word is CONJUNCTION then :
Return true
Else
Return false
[End of If structure]
[End of inner for loop]

Algorithm-6: Check for Conjunction

Step 9: Review Analysis

This algorithm describes how the overall process of Review Analysis generated.

Generate (fname,feature)
[The function Generate will which whether the file contains the positive review for the movie or not. Fname is the name of the file containing the movie review and feature is the feature selected by the user for analyzing the review document file.]
Set docscore := 0,nlines :=0, br :=openfile(fname)
while ((strLine = br.readLine()) =! NULL) {
Set slines1[nlines1]=strLine, Set nlines1 := nlines1 +1 ; [Line will be taken according to sentence] [each line is a sentence]
}
[End of while loop]

Repeat for nl:=0 to nlines1-1 by 1 do:
If ISFEATURE(slines1[nlines1]) and CHECKCONJUNCTION(slines1[nlines1] then :
(a) Set slines[nlines]=slines1[nlines1], Set nlines := nlines +1 ;
[End of If structure]
} [End of while loop]

Repeat for nl :=0 to nlines-1 by 1 do:
call STEMTAGG(nlines[nl])
Split nlines[nl]-→Words array
Set LScore:=0.
Repeat for I := 0 to Word.Length-1 By 1 do:
Set sc:=SENTISCORE(Word[I],Pos_Tag).
If CHECKINTENSIFIER(Word[I-1]) then :
(a) Set LScore :=LScore –Pscore.
(b) Set sc :=sc * Pscore.
Else If CHECKNEGATION(Word[I-1])then :
(a) Set LScore :=LScore –Pscore.
(b) Set sc :=sc*-1.
End if
Set LScore : = LScore +sc.
[End of for loop]
Set docscore:=docscore+LScore.
[End of for loop]
If docscore>=0 then
Write "Movie Review is Positive"
Else :
Write "Move Review is Negative"
[End of If structure]

Algorithm-7: Review Analysis Algorithm.

V. IMPLEMENTATION

The proposed algorithm is implemented in Eclipse Java Enterprise Edition (J2EE) Integrated Development Environment (IDE) for Web Developers Version: Kepler Service Release 1 Build id: 20130919-0819 with SentiWordNet 3.0 and Stanford Tagger tools.

To run the above software the required hardware are core i3 processor 2.30 GHz of 4 GB of RAM.

Now the chapter follows with explanation of implementation of algorithm with the help of screenshots of my work I have taken during my practical work.

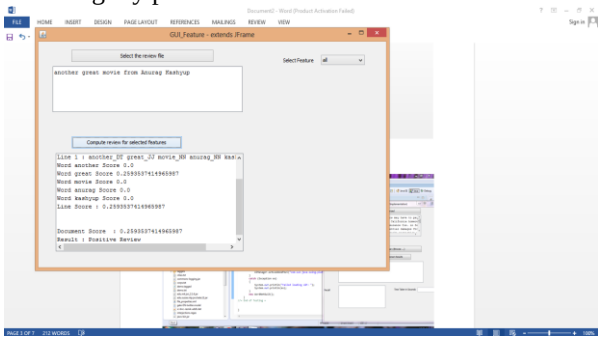

Fig:3 Score calculation for movie review

VI. TEST RESULT

6.1 Comparison Test Results for Negation Handling

Review Text	Proposed Work	Base paper
There was no good story in that movie.	-0.659	0.608
There was good story in that movie.	0.608	0.608
The story of the movie was not good.	-0.684	-0.042
The story of the movie was good.	0.582	0.582
The music of the songs was not bad.	0.953	-0.812
The music of the songs was bad.	-0.187	-0.187
The hero acting was not bad in Bajirao Mastani movie.	0.712	-1.053
The hero acting was bad in Bajirao Mastani movie.	-0.428	-0.428

Table 6.1 Comparison Test Results for Negation Handling

6.2. Comparison Test Results for Intensifier Handling

Review Text	Proposed Work	Base paper
The movie was very good.	0.717	0.612
The movie was good.	0.6337	0.6337
The songs were too bad.	-0.487	-0.653
The songs were bad.	-0.570	-0.570
The movie was directed very badly.	-0.349	-0.234
The movie was directed badly.	-0.432	-0.432

Table 6.2 Comparison Test Results for Intensifier Handling

Result Accuracy Base Paper	Result Accuracy Proposed Work
85%	97%

The percentage is calculated on the average of the number of reviews compared in both the implementation and some samples are shown in the tables presented above.

REFERENCES

[1] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining", In Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, Geneva, Italy, 2006.

[2] N. D. Valakunde, Dr. M. S. Patwardhan, "Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013.

[3] Na Fan, Wandong Cai, Yu Zhao "Research on the Model of Multiple Levels for Determining Sentiment of Text" IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2008.

[4] Si Li, Hao Zhang, Weiran Xu, Guang Chen and Jun Guo," Exploiting Combined Multi-level Model for Document Sentiment Analysis" 2010 International Conference on Pattern Recognition 1051-4651/10 $26.00 © 2010 IEEE, 2008

[5] Raymond Y.K. Lau and Wenping Zhang, Peter D. Bruza "Learning Domain-specific Sentiment Lexicons for Predicting Product Sales" Eighth IEEE International Conference on e-Business Engineering, 2011.

[6] NIDHI MISHRA & C. K. JHA ,"OPINION MINING FROM TEXT IN MOVIE DOMAIN", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 4, Oct 2013, 121-128

[7] Aniket Dalal, Kumar Nagaraj, Uma Sawant" Hindi Part-of-Speech Tagging and Chunking : A Maximum Entropy Approach"

[8] Ankit Ramteke , Akshat Malu , Pushpak Bhattacharyya " Detecting Turnarounds in Sentiment Analysis: Thwarting" 2012

[9] Aurangzeb khan Baharum Baharudin,"Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs",2011 IEEE.

[10] Buddhika H. Kasthuriarachchy*t, Kasun De Zoysa*+ and H.L. Premaratne*§ "Enhanced Bag-of-Words Model for Phrase-Level Sentiment Analysis" 2014 International Conference on Advances in ICT for Emerging Regions (ICTer): 210 – 214

[11] Chinsha T C, Shibily Joseph," A Syntactic Approach for Aspect Based Opinion Mining;" Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)

[12] Dr. Muhammad Shahbaz1, Dr. Aziz Guergachi2, Rana Tanzeel ur Rehman3," Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text" 978-1-4799-3010-9/14/$31.00 ©2014 IEEE

[13] Farah Benamara, Carmine Cesarano, and Diego Reforgiato Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone ICWSM Boulder, CO USA, '2007.

[14] J. Ashok Kumar, S. Abirami, and AN

EXPERIMENTAL STUDY OF FEATUR EXTRACTION TECHNIQUES IN OPINION MINING" International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.4, No.1, February 2015 DOI: 10.5121/ijscai.2015.4102

[15] Jasmine Bhaskar, Sruthi K, Prema Nedungadi, "Enhanced Sentiment Analysis of Informal Textual Communication in Social Media by Considering Objective Words and Intensifiers" 978-1-4799-4040-0/14/$31.00 ©2014 lEEE Kerstin Denecke "Using SentiWordNet for Multilingual Sentiment Analysis" 2008.

[16] Lizhen Liu, Xinhui Nie, Hanshi Wang," Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis" 2012 5th International Congress on Image and Signal Processing (CISP 2012) 978-1-4673-0964-6/12/$31.00 ©2012 IEEE

[17] Mohsen Farhadloo , Erik Rolland," Multi-Class Sentiment Analysis with Clustering and Score Representation" 2013 IEEE 13th International Conference on Data Mining Workshops DOI 10.1109/ICDMW.2013.63

[18] Monalisa Ghosh, Animesh Kar ," Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiWordNet 3.0" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 9, September – 2013

[19] Ms.K.Mouthami, Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran "Sentiment Analysis and Classification Based On Textual Reviews"

[20] Mostafa Al Masum Shaikh, Helmut Prendinger and Mitsuru Ishizuka ,"An Analytical Approach to Assess Sentiment of Text" ,1-4244-1551-9/07/$25.00 ©2007 IEEE

[21] Nabeela Altrabsheh, Mihaela Cocea, Sanaz Fallahkhair" Sentiment analysis: towards a tool for analysing real-time students feedback" IEEE 26th International Conference on Tools with Artificial Intelligence, 2014.

[22] Na Fan, Wandong Cai, Yu Zhao "Research on the Model of Multiple Levels for Determining Sentiment of Text" IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application,2008.

[23] NIDHI MISHRA & C. K. JHA "OPINION MINING FROM TEXT IN MOVIE DOMAIN" International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 4, Oct 2013, 121-128

[24] Nidhi Mishra1, Dr. C.K. Jha ." An Insight into Task of Opinion Mining" Passent Elkafrawy (Ed.): SPIT 2012, LNICST pp. 182–187, 2012. © Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2012

[25] N. D. Valakunde, Dr. M. S. Patwardhan "Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process" International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2013.

[26] Ohana, B. & Tierney, B. (2009)" Sentiment classification of reviews using SentiWordNet" 9th. IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, 22-23 October.

[27] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79--86 (2002)

[28] Raisa Varghese, Jayasree M," Aspect Based Sentiment Analysis using Support Vector

[29] Jindal, N., & Liu, B., Opinion Spam and Analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 219–230). New York, NY, USA: ACM. doi:10.1145/1341531.1341560,2008

[30] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. , Detecting Product Review Spammers Using Rating Behaviors. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (pp. 939–948). New York, NY, USA: ACM. doi:10.1145/1871437.1871557,2010

[31] Wang, G., Xie, S., Liu, B., & Yu, P. S., Identify Online Store Review Spammers via Social Review Graph. ACM Trans. Intell. Syst. Technol., 3(4), 61:1–61:21. doi:10.1145/2337542.2337546,2012.

[32] Xie, S., Wang, G., Lin, S., & Yu, P. S. ,Review Spam Detection via Time Series Pattern Discovery. In Proceedings of the 21st International Conference Companion on World Wide Web (pp. 635–636). New York, NY, USA: ACM. doi:10.1145/2187980.2188164,2012.

[33] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana,"COMPARISON AND ANALYSIS OF SPAM DETECTION ALGORITHMS",International Journal of Application or Innovation in Engineering & Management (IJAIEM),2013.

[34] R.Malarvizhi, K.Saraswathi,"Content-Based Spam Filtering and Detection Algorithms- An Efficient Analysis & Comparison", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 9- Sep 2013.

[35] Rekha, Sandeep Negi ,"A Review on Different Spam Detection Approaches",International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 6 - May 2014.

[36] Muhammad Iqbal1,Malik Muneeb Abid2, Mushtaq Ahmad3 and Faisal Khurshid4,"Study on the Effectiveness of Spam Detection Technologies",I.J. Information Technology and Computer Science, 2016.

[37] Megha Rathi,Vikas Pareek,"Spam Mail Detection through Data Mining – A Comparative Performance

Analysis",I.J. Modern Education and Computer Science, 2013.

[38] Rohit Giyanani, Mukti Desai,"Spam Detection using Natural Language Process ing",IOSR Journal of Computer Engineering (IOSR-JCE),Oct. 2014.

[39] Marco Túlio Ribeiro, Pedro H. Calais Guerra, Leonardo Vilela, Adriano Veloso, Dorgival Guedes, Wagner Meira Jr,"Spam Detection Using Web Page Content: a New Battleground".

[40] Reena Sharma, Gurjot Kaur,"Spam Detection Techniques: A Review",International Journal of Science and Research,2013.

[41] Vandana Jaswal, Astt Professor. Nidhi Sood,"Spam Detection System Using Hidden Markov Model",International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

