RESEARCH ARTICLE OPEN ACCESS

Opinion Mining and Improvised Algorithm for Feature Reduction in Sentiment Analysis

Siddharth. J[1],Ashish .S[2], Shreyash.A[3], Rishi.A[4],Prashant.U[5,]
[1,2,3,4] *Research scholar, Department of Computer Engineering,*
[5]*Assistant Professor, Department of Computer Engineering*
NMIMS University, MPSTME, Shirpur, Maharashtra, India.

ABSTRACT
Nowadays organisations use the power of the web to analyse the review of the product by customer. The organisation cannot trust star based reviews because it can be faked by robots. That is why textual review is preferable. Opinion mining is used to find the approximate sentiment of the review. Sentiment analysis is a part of opinion mining which helps an organisation to get valuable feedback of the product by extracting the polarity of reviews. The review of a product may be used to improve productivity of the organisation as it could improve its product's features based on reviews. It provides us ways to analyse a given review. In our review paper, we have emphasised on the content based analysis of the review rather than deciding the contextual polarity by its topic. We have reached to our proposed algorithm by referring to BoPang and LillianLee's [2] paper and Tirath Prasad Sahu and Sanjeev Ahuja's [3] paper.

I. INTRODUCTION

With internet shopping becoming a daily trend, ensuring products liability is a very necessary task. Sentiment analysis, an integral part of opinion mining, focuses on mining of reviews and feedbacks by customers and summarizing this information such that it helps future customers to make decision with ease. This can be done by mining the reviews on products from particular sites and looking for sentiment words that tend to express the nature of the review being negative or positive[5].

Opinion mining on a brief context is reaching out for needed data from a forest of information. Since the use of social media sites, blog sites and online discussion forums has become more and more common, people tend to seek out to these sites to share their impressions on a particular product. This kind of informal reviews is accompanied by unstructured form of grammar but on the upside, it impacts everyone, from the audience, to the production company and also to the retailers and hence opinion mining is gaining more and more popularity in this digital world.

World web has brought everyone and everything closer and closer. People are a touch away from getting any kind of information. To get the most out of it many sites use the technique called sentiment analysis. Sentiment analysis is the technique which distinguishes a bad review from a good one. This is done by analyzing the contextual polarity of the whole review and dividing the language into positive, negative, and neutral categories. To do so different approaches are used which mainly focuses on subjective classification which reduces the data set by discarding the objective sentence and in-turn reducing the computational power. But given the inherent uncertainty of natural language of people, this process can sometimes provide us with over-confident decisions thus failing to provide best results.

II. OPINION MINING - A PART OF DATA MINING

Data mining is the extraction of data using relevant tools and using this unstructured data to find a meaningful pattern. Opinion mining can be referred to as finding polarity of opinion of customers from their textual review. The opinion of a person is their thoughts of the relevant product after using it. After a review is mined for opinion we come to know whether the review is positive or negative. In our view opinion, can be classified in two ways, 1. Analyzing contextual polarity of topics 2. Analyzing contextual polarity of whole review The web based reviews are extracted and collected. These reviews are filtered and classified on the basis of contextual polarity and then are summarized [1]. The detailed procedure is depicted in Fig 1.

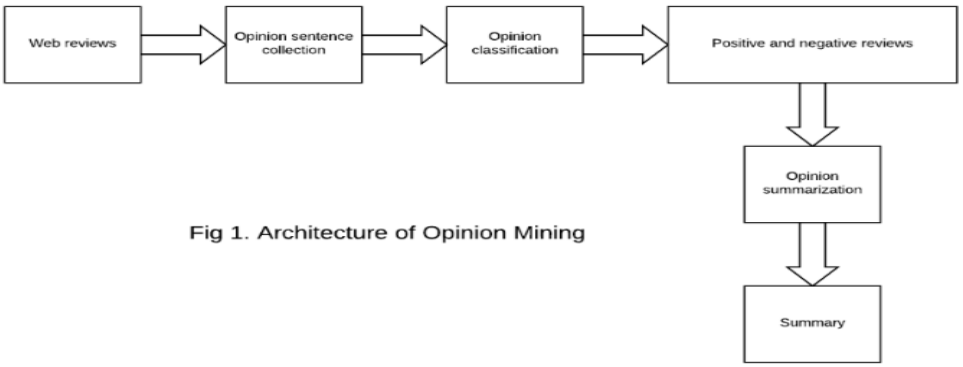

Fig 1. Architecture of Opinion Mining

III. CLASSIFICATION OF SENTIMENT LEVEL

Sentiment level is the scale at which a document is analyzed for sentiment classification [6][8].

- 3.1 Document Level – In this type of sentiment the overall contextual polarity is calculated by analyzing the whole document. The positivity or negativity is calculated by summarizing the whole document. The topic of the document is not taken into account for finding its polarity.
- 3.2 Sentence Level – As the name suggest the contextual polarity is calculated by analyzing the sentence. The overall polarity of the review is calculated by the number of positive or negative reviews. If more number of positive reviews the review is classified as positive. If the polarity of reviews is equivalent, then it will be considered neutral else negative.
- 3.3 Phrase Level – The opinionated words are found and classified on a phrase level. It is a more precise approach. It is rarely used as it may contradict the actual review of the customer ultimately compromising the real context

IV. EXISTING METHODS OF CLASSIFICATION

According to Bo Pang and Lillian Lee the document can be classified further based on its subjectivity [2]. That is if the review document is subjective or opinionated then only the review is sent to classifier else it is discarded. The detailed illustration is stated in fig 2 below.

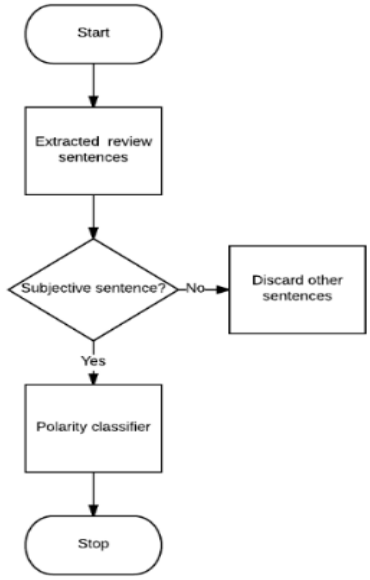

Fig 2. Polarity classification via subjectivity detection proposed by Bo Pang & Lillian Lee

In this method, objective sentences are discarded which in turn helps us to classify the review better. Because the text input is reduced in this, in our understanding computational power required will be comparatively less. But this methodology has a drawback i.e. it could not precisely calculate the polarity, because polarity is also further classifiable. The polarity classification was proposed by

Tirath Prasad Sahu and Sanjeev Ahuja [3]. They proposed a methodology which precisely and efficiently processes the review. The flowchart in fig 3. depicts their proposed method.

What they have proposed is that first the dataset is pre-processed using the following techniques:

1. Stemming: It is a process of converting all morphological words to the root word. For example, post, posted and posts are rooted to a common word post.

2. Stopping: A stop-word list is made which includes common group of words like a, the, an etc. These are removed from the text to decrease the content size.
3. POS tagging: POS is Part of Speech tagging. Each word of the sentence is tagged with their respective POS [4] such as noun, adjective, verb etc. For example, consider a sentence " This feature is essential" . This sentence is converted into- " This(Determinant)feature(noun)is(verb)essential(noun)" .

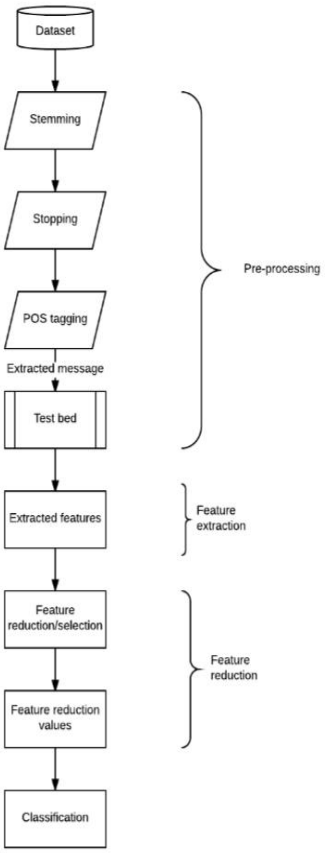

Fig 3. Method proposed by Tirath Prasad Sahu & Sanjeev Ahuja

Further after pre-processing data is extracted and then the features of the review are reduced. They used feature ranking algorithms to form the following index given in Table 1

Table 1: Feature Impact

Feature	Information Gain
Positive Sentiment Words	0.312
Positive Sentiment Bi-grams	0.26
Positive Sentiment words+Adjective	0.22
Negative Sentiment Words	0.207
Negative Sentiment Bi-grams	0.153
Negative Sentiment words+Adjective	0.116

According to these parameters SentiScore (Sentiment Score) which is used in the proposed algorithm by Sanjeev Ahuja and Tirath Prasad Sahu. We have illustrated their algorithm by using flow chart below in fig 4

We have seprated The algorithms for illustration purpose figure 4.1 illustrates Algorithm 1 and 4.2 illustrates Algorithm 2 below,

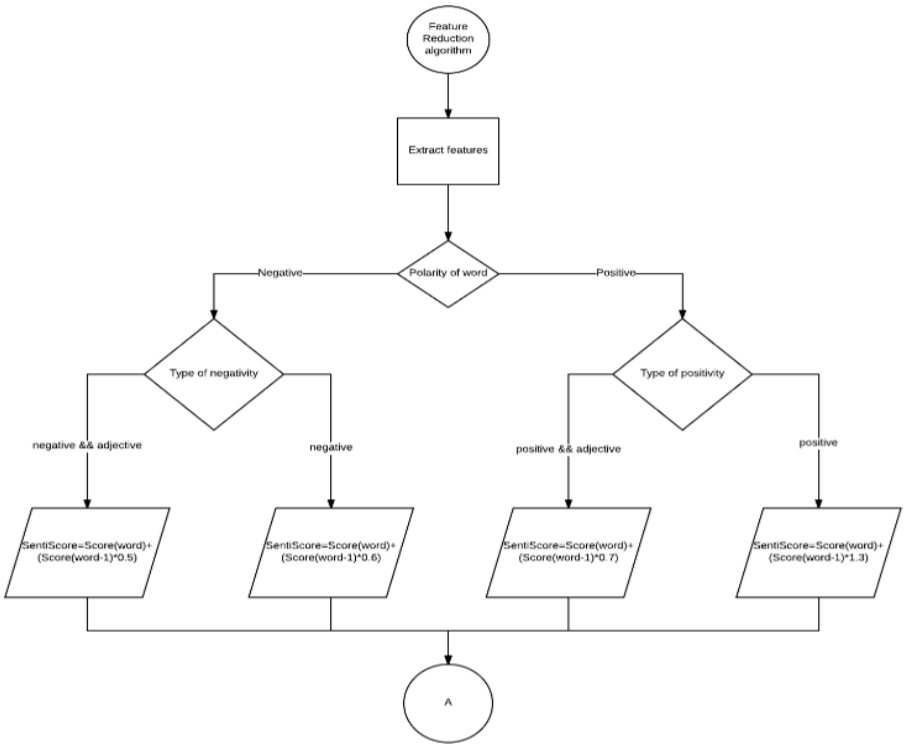

Fig 4.1 Algorithm 1

After The feature reduction by these algorithms the reduced data set is sent to classifier which classifies the review. As seen above all the algorithms and methodologies have drawbacks. Methodologies proposed by Bo Pang and Lillian Lee subjectively classifies the text, there is no incentive for further classification by polarity.

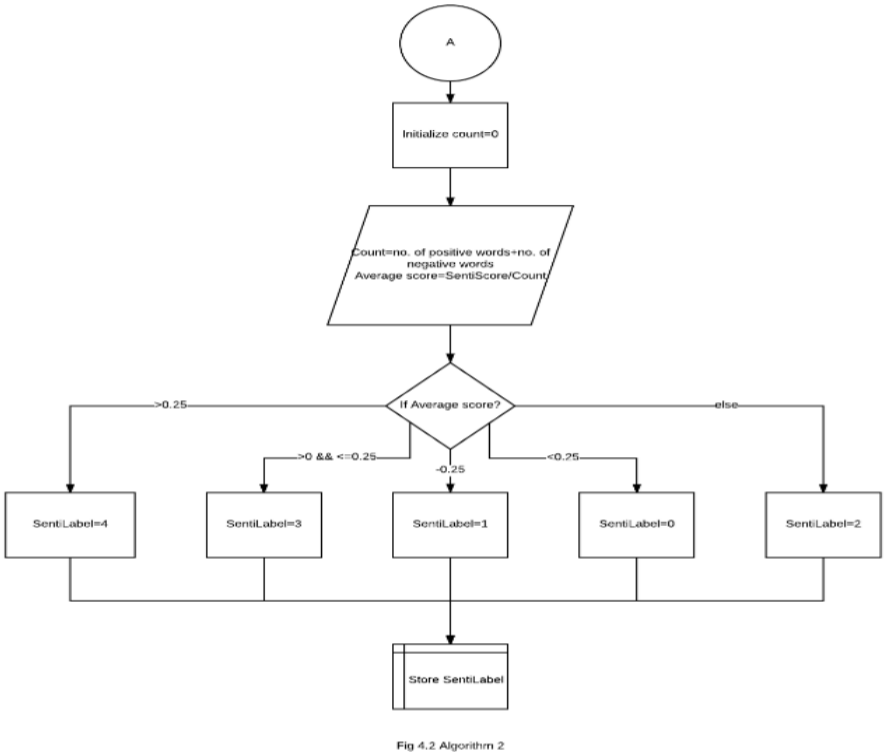

Fig 4.2 Algorithm 2

Conversely in methodology and algorithm proposed by Tirath Prasad Sahu and Sanjeev Ahuja classifies the extracted review text further by their polarity and labels them and is accordingly sent to classifier but there is no incentive for subjective classification of text.

IV. PROPOSED METHODOLOGY

What we propose to do is use the methodologies of the research paper we have reviewed and find a better method which will remove limitation from the methodologies proposed by the author's research paper [2][3].
The changes are illustrated in the flowchart fig 5
5.1 illustrates common preprocessing.
5.2 illustrates merging of method and algorithm by authors of [2][3].

In our proposed methodology, the data set is sent for subjective analysis. In subjective analysis, the subjective sentences of the extracted review text are selected, objective sentences (like product specification) which are of no use are discarded. Then the subjective data is preprocessed i.e. Porter stemming, stopping, POS tagging. The data preprocessed is then further classified based on the polarity. These selected datasets are reduced using the feature reduction algorithms (1&2) as proposed by Tirath Prasad Sahu and Sanjeev Ahuja. The data is then classified by classifier algorithms.

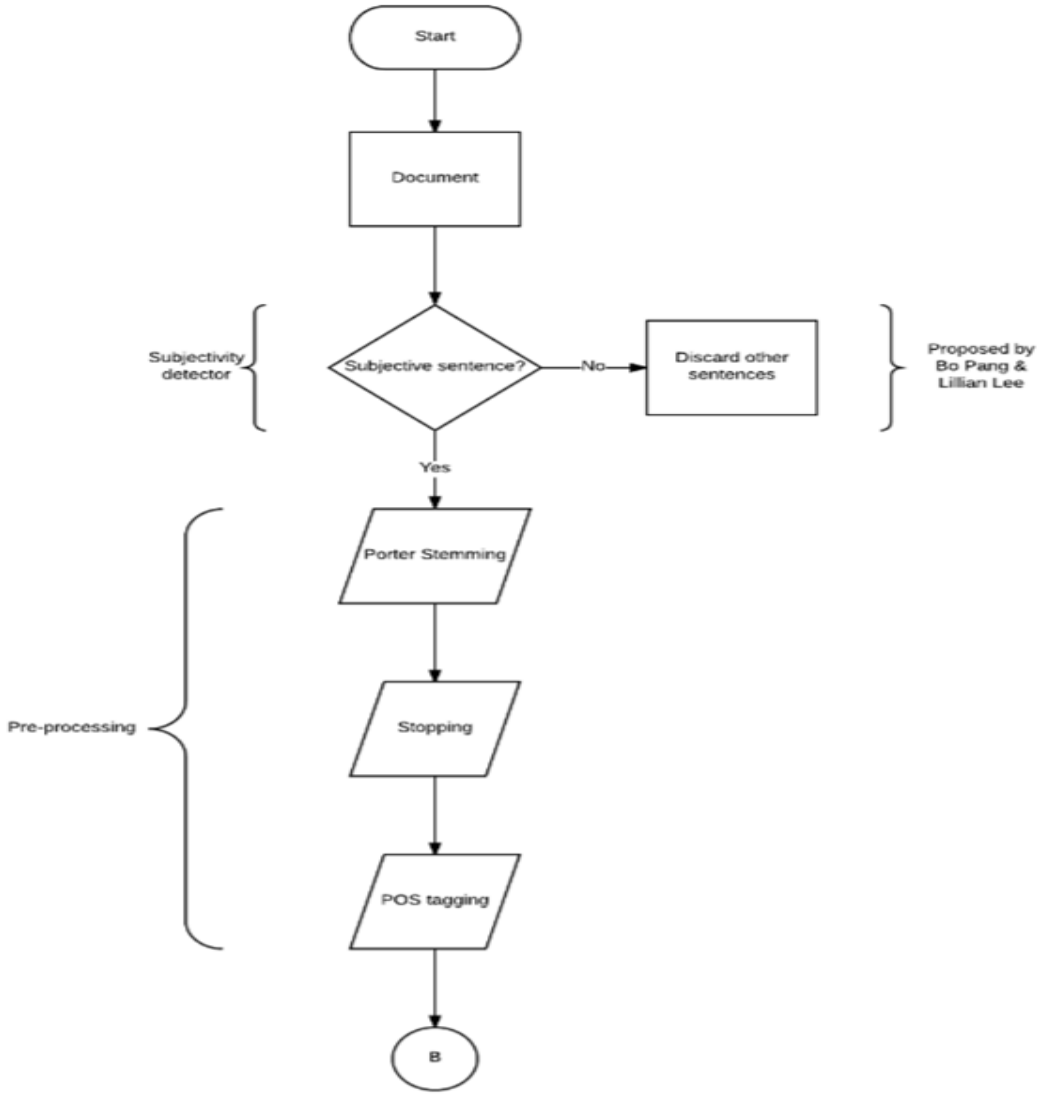

fig 5 (a): common preprocessing

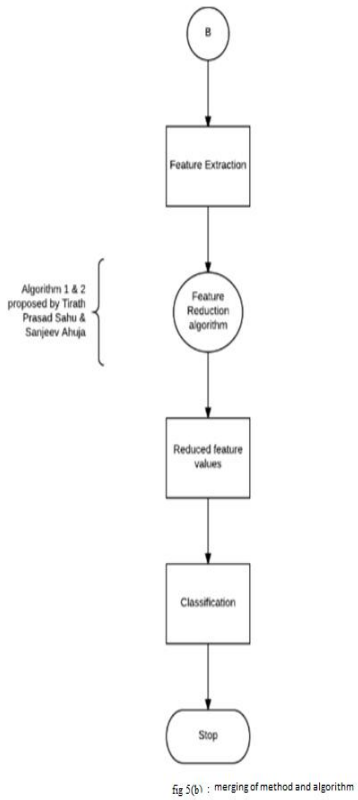

fig 5(b) : merging of method and algorithm

V. CONCLUSION

Opinion mining is one the most widely spread technique to validate customer products and what we have tried is to find an optimal way to do the same. For our proposed algorithm we have inherited the ongoing base algorithm to mine for positive-negative comments for customer products and have tried to customize an algorithm that best describes a way for completely extracting a way to find the real review a customer makes for a product using less computational power and more reliability.

REFERENCES

[1]. P.Kalarani, Dr.S. Selva Brunda (2015)“ An Overview on Research Challenges in Opinion Mining and Sentiment Analysis “ Published in International Journal of Innovative Research in Computer and Communication Engineering.

[2]. BoPang and LillianLee "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts" Published in proceeding ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.

[3]. Tirath Prasad Sahu and Sanjeev Ahuja “Sentiment Analysis of Movie Reviews: A study on Feature Selection & Classification Algorithms “ 2016 International Conference on Microelectronics, Computing and Communications (MicroCom) Year: 2016.

[4]. Keerthi Lingam, E Ramalakshmi and Srujana Inturi. Article: English to Telugu Rule based Machine Translation System: A Hybrid Approach. International Journal of Computer Applications 101(2):19-24, September 2014.

[5]. Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." AAAI. Vol. 4. No. 4. 2004.

[6]. V. S. Jagtap, Karishma Pawar “ Analysis of different approaches to Sentence-Level Sentiment Classification” International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 [1 April 2013]

[7]. Swati N. Manke ,Nitin Shivale "A Review on: Opinion Mining and Sentiment Analysis based on Natural Language Processing" International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 4, January 2015.

[8]. G.Vinodhini, RM.Chandrasekara "Sentiment Analysis and Opinion Mining: A Survey" Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.