A Framework for Summarization of Online Opinion using Weighting Scheme

Mugdha More[1] and Bharat Tidke[2]

[1]Department of Computer Engineering, Savitribai Phule Pune University, Khopi, Pune-412 205, India
[2]Department of Computer Engineering, Savitribai Phule Pune University, Khopi, Pune-412 205, India

Abstract

Recently there is wide use of social media includes various opinion sites, complaints sites, government sites, question-answering sites, etc. through which customer get services, opinion, information, etc. but because of this there is more and more use of these social media right now so huge amount of data will be created, from this huge data people get confused while taking any decision about particular problem or services. For example, customer wants to purchase a product at that time he/she want the previous customer feedback or opinion about that product. But if there is lots of opinion available for particular product then that customer get confused while taking decision whether purchase that product or not. In this case there is a need of summarization concept means that only show the short and concise manner summary about service or product so that customer or organization easily understand and able to take right decision fast. Our proposed framework creating such summary which contain three main phases or steps. Firstly preprocessing is done in that stop words are removed and stemming is performed. In second phase identify frequent features using two techniques weight constraint and association rule and at the last phase it find semantics and generate the summary so that customer will able to take step without confusion.

Keywords

Association, Opinion, Preprocessing, Weighting, Summary

1.INTRODUCTION

People expose their thoughts, things with different sites and also they can communicate with each other through social media [1] [2]. They also give their views about particular services or products through web in their own language but each human thinking is different so various types of opinions are available for single thing [1] [2] [3]. From this opinion customer or organization get various way of thinking or dimensions of particular product or services but all dimensions may not be useful or may be some dimensions are fake hence automatically the opinion are fake so there is a need of identification of such opinion called opinion spam [3] [4]. Also there are some opinions which are duplicate or near duplicate [2] [5] [6]. Existing features focus on classification, summarization, spam detection, association, sentiment analysis, semantic analysis, opinion mining, etc. for identifying true opinions from fake, duplicate and near duplicate reviews or opinion [4]. Classification means categorization of whole data into different subgroups considering some constraints or rules. For example if there is a group of people and we have to classify them then we take constraint as a gender hence we get two group of people female and male [1] [2]. Summarization is nothing but summary of data i.e. short and important data from whole data. If there is ten pages data available then from summarization we get two to three page

DOI:10.5121/acii.2015.2302

proper and important data which are useful. Summarization is closely related to opinion mining concept [2]. Opinion mining is a technique which analyzes the data from different perspective and summarizes it into knowledgeable information [3]. Opinion mining is also called as sentiment analysis is nothing but analyze the data with its sentiments, also it is a study of emotions, appraisals, opinion of peoples [7]. Association technique is a rules created for frequent features identification which having some relationships between each other [2] [3] [8]. But while handling such techniques various difficulties are occurred like word sense disambiguation, negation handling, etc. because opinions are in natural language and all this techniques handle such natural language opinions [7]. There is need of deep understanding and knowledge about language rules like implicit or explicit, regular or irregular, syntactic or semantic, etc. for handling such problems [9]. Various companies, organization, markets, businesses required such techniques for their company or business improvements, decision about services, etc. In this paper we propose a framework containing three phases for creating concise summary which minimizes the customer or organization confusion.

2.RELATED WORK

Our work is related to sentiment analysis, opinion mining and opinion summarization. Existing research is on various domains and on various techniques. G. Vinodhini and RM. Chandrasekaran have been focused on performance of classification model and different features that shows sentiments [10]. Bing Liu and Lei Zhang proposed unified framework and identify mining tasks and also find out techniques and issues of opinions [11]. Polarity classification accuracy is increased through minimum cuts framework which is proposed by Bo Pong and Lillian Lee [12]. Bing Liu represents the problem of opinion mining in sentence, document and aspect-level [13]. Also there is a technique which finds opinions with its holder and topic introduced by Su Nam Kim et al. [14]. Also there is a technique for improving performance on opinion mining using syntactic dependency relationship and technique for converting dependency relation triples to composite features which are described by Mahesh Joshi et al. [15]. Eric Breck et al. proposed opinion expression identification approach with the help of conditional random field and also evaluate that approach [16]. Some research is on identification of opinion boundary and its intensity without exploiting hierarchical structure by Yejin Choi et al. [17]. For opinion statement analysis rule based approach is proposed by Alena Neviarouskaya et al. They do classification of verb and create rule for each verb class [18]. Summarization of huge data is done by using graph-based summarization framework called Opinosis by Kavita Ganeson [19]. The method for feature mining and creating summary is introduced by Minqing Hu and Bing Liu. They also find out difficulties while creating such summary [20]. By using summarization method and ranking technique Yihong Gong and Xin Liu represent the text summarization through which they increase the performance [21]. The overview of research done in sentiment analysis and summarization is shown in Table 1 with data set.

Table 1. Existing research technique with its features and dataset.

Author	Technique	Feature	Dataset
G. Vinodhini et al.	Sentiment Analysis	Sentiment method, technique, challenges, classification	Movie Review, Product Review
Bing Liu et al.	Opinion Mining, Sentiment Analysis	Framework to unify research directions	Product Review
Bo Pong et al.	Sentiment polarity, Minimum Cuts in Graphs	Machine Learning Method, Text Categorization Techniques	Movie Review
Su Nam Kim et	Semantic role labelling	Target word, Phrase Type,	FrameNet, The

al.	method	Head word, Parse tree path, Position, Voice, Frame name	New York Times, BBC News, etc.
Mahesh Joshi et al.	Syntactic dependency relation triples	Composite Back-off, Full Back-off, Ngram Back-off	Amazon.com, CNET.com
Eric Breck et al.	CRF	Lexical, Syntactic, Dictionary based	MPQA Corpus
Yejin Choi et al.	Hierarchical parameter sharing technique, CRF	Per-token, Transition	MPQA Corpus
Alena Neviarouskaya et al.	Rule-based approach, Verb classification	Opinion Type, Strength, Confidence level, Verb class	VerbNet
Kavita Ganesan et al.	Abstractive Summary	Opinosis Framework	Hotel Review, Car Review, Product Review
Minqing Hu et al.	Semantic orientation, Term Phrases	Bootstrapping Technique, Feature Extraction Method	Amazon Reviews
Yihong Gong et al.	Semantically Important Sentences, Improve Summarization Performance	Standard IR Model, Latent Semantic Analysis, VSM Weighting Schemes	CNN Worldview News Program

3.PROPOSED FRAMEWORK

Figure 1 gives proposed framework for online opinion summarization. The inputs to the framework are date and time, product name and review of that product. The output is summary of the review in short and concise manner. The system perform the summarization in three steps: (1) Product feature get mined which is given by customer; (2) Identify frequent features in each opinion sentence in each review; (3) Finding out whether feature are opinionated and also identify orientation of the words and finally summary will be created. Multiple sub-steps are involved for performing these three steps.

3.1.Dataset

We conducted our experiments using customer reviews of hotels, tourism, hard disk and baby product. Hotels include reviews of two hotels, Hard disk contain the reviews of two companies product, tourism have two location and in baby product two companies product are included. The reviews were collected from Amazon.com. Product in these sites has large number of reviews.

3.2.Pre-processing

The framework first downloads all reviews and put them in the review database. We use the parser to parse each review to split text into sentence to produce part of speech tags for each word, reason behind is that the original opinion is in unstructured format means that they are written in users natural language [22]. Preprocessing task is also performed on these documents. It removes stop word or unimportant words in filtration step for that it maintain a stop word list. Words are matched with this list, if match found then remove that word from opinion and if new word is found as a stop word then add that word in the stop word list. It mainly includes a, an, the, etc. words that was only used for grammatically complete the sentence. After that stemming is performed, in this process root form of inflected words is find out and also eliminate the prefixes and suffixes of the words. If there is a words like developed, written then after stemming the words like develop and write.

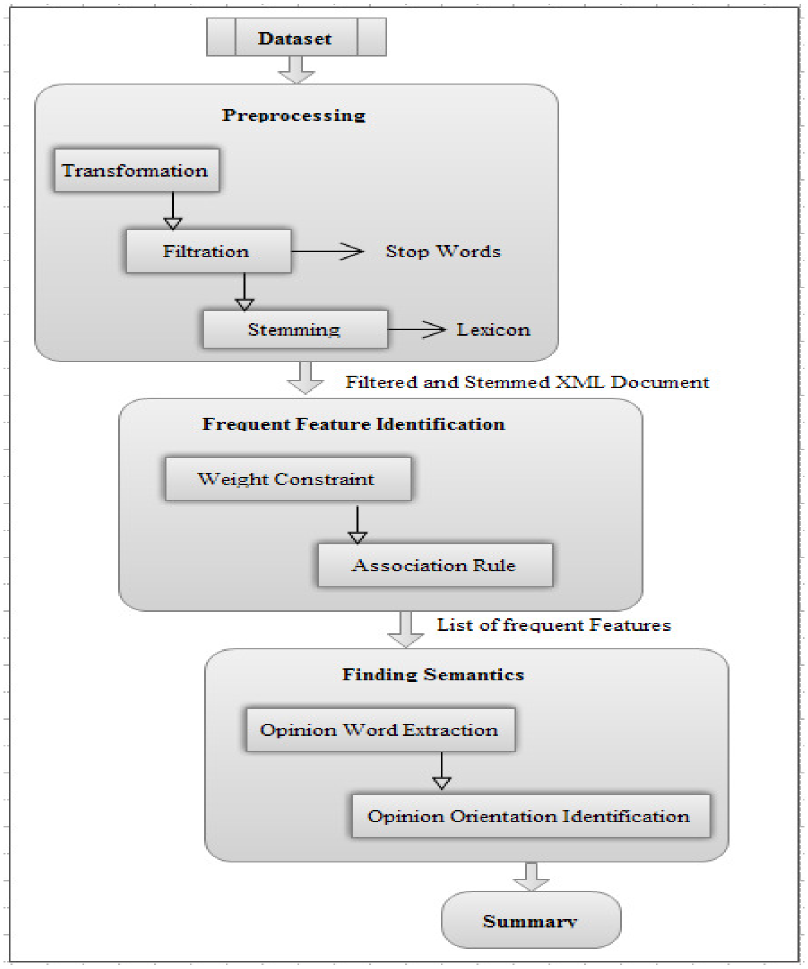


Figure 1. Feature Based Online Opinion Summarization Framework

3.3.Frequent Feature Identification

The In this step we identify product feature which is present in opinion that express by users. Before going to actual technique we firstly introduced that how difficult to extract feature from opinion which are in natural language. Some opinions are directly express on entity explicitly so that extract feature from these is not so much difficult. For example, “Sound quality of Samsung phone is good” but there are some opinion which expressed implicitly so it difficult to find out features. For example, “Samsung phone does not fit in pocket” means here user indirectly talks about size of the phone. In our work, we used two techniques for identifying frequent features. Firstly we use weighting scheme on filtered and stemmed XML document to index the feature. These process is possible by manually and automatically also but here we use TF-IDF weighting scheme for assigning weights to keywords [8]. The formula for TF-IDF weighting scheme is as follows:

$$w(a,b) = tfidf\left(d_{a,}t_b\right) = \begin{cases} Nd_{a,}t_b * \log_2 \frac{|C|}{Nt_b} & if\ Nd_{a,}t_b \geq 1 \\ 0 & if\ Nd_{a,}t_b = 0 \end{cases}$$

where, $w(a, b) \geq 0$, $Nd_{a,}t_b$ number of terms t_b occurs in document d_a, Nt_b denotes the number of document in collection C in which t_b occurs at least once and $|C|$ denotes the number of document in collection C.

After indexing to features by weighting schemes, association rule is applied on it to find all frequent item sets. Association mining is nothing but identifying relationship between the words hence we get frequent feature set [6]. In our case, we apply association on frequent features to get more accurate frequent feature set. These frequent features are getting by calculating support formula. This support must be greater than minimum support which is given by user. Support formula is as follows:

$$Support\ (w_a,\ w_b) = \frac{Support\ count\ of\ w_a{,}w_b}{Total\ number\ of\ document}$$

where, w_a, w_b are the frequent features. We combine these two techniques for getting more accuracy in identifying frequent features.

3.4.Finding Semantics

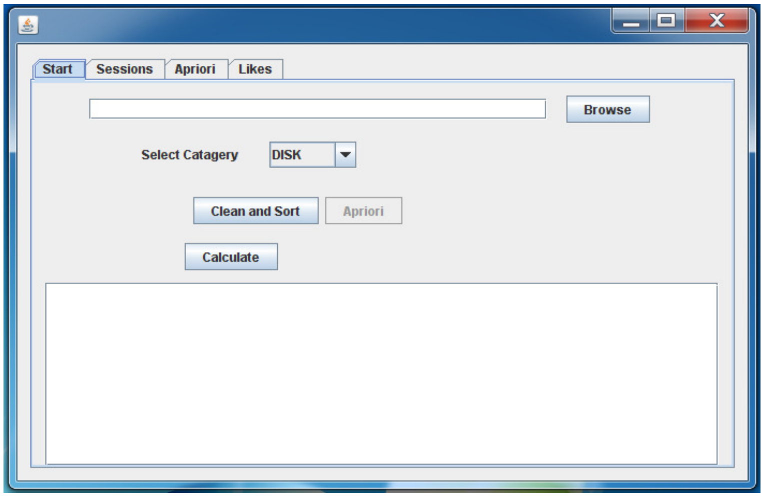
In this step, firstly remember one thing is that we are only interested in customer opinion. In the first sub-step we extract the opinion word that express the subjective opinion. In some cases presence of adjectives is more useful for identifying subjective opinion. In previous step we assign the tags to word using part-of-speech tag technique. With the help of these tags and its position in the sentences we get adjective, adverb, noun and all so extracting the opinion words from these words are easy. Identification of a semantic orientation of a word is useful for identifying semantic orientation of sentences. Each word having semantic orientation which is directed to its semantic group. If the word shows positive orientation that means word gives desirable state and if the word shows negative orientation that means word gives undesirable state. To predict the semantic orientation we maintain the synonym set and antonym set. Each new word shows or has either positive or negative orientation but for predicting that orientation we have to search its synonym or antonym set so according to that we get its orientation. Initially we manually create a list called seed list by considering very common adjectives like good, nice, fast, etc. as a positive word and bad, slow as a negative word. For predicting the orientation of that word we search that word in WordNet. If found the orientation then add it in to our list. According to this all procedure semantic orientation of a word is find out.

3.5.Summary

After performing all previous steps, we are ready to generate feature based online opinion summary which is short and in concise manner. Previously we finding out frequent features after that we got its orientation for particular product. After all this we get frequent opinion about particular product then we convert that all frequent opinion into single well defined opinion without changing its meaning so that customer easily understand and easily get the meaning of opinion in short time.

4.EXPERIMENTAL EVALUATION

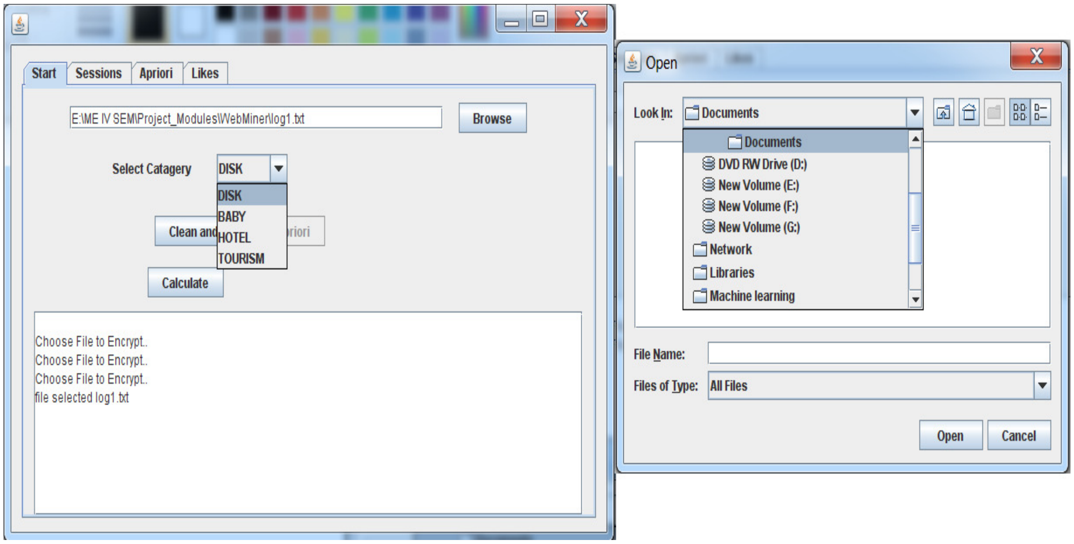
A Framework called feature based online opinion summarization based on proposed techniques has been implemented in Java. We conducted our experiment using the customer review of four products having different company. For each product, firstly we crawled and download the reviews. After that, by using parser we generate part of speech tags then to perform summarization our framework is applied on it. User interface of our framework is shown in figure 2.



Select Catagery DISK

Figure 1. GUI for summarization framework.

In that we provide a facility for selecting a file which contains all reviews. Also we can select the category whatever we want. We show these all functionality in following figure 3.



E:\ME IV SEM\Project_Modules\WebMiner\log1.txt

Select Catagery DISK

```
Choose File to Encrypt..
Choose File to Encrypt..
Choose File to Encrypt..
file selected log1.txt
```

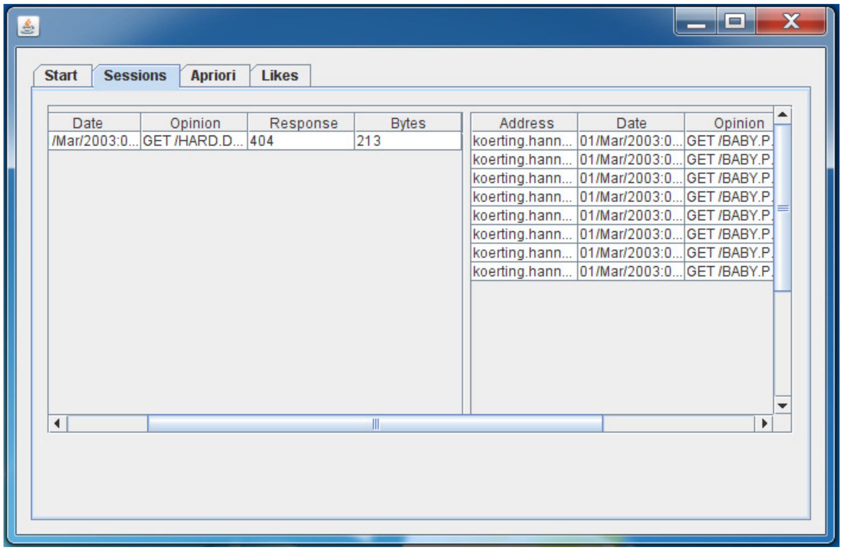
Look In: Documents

File Name:

Files of Type: All Files

Figure 2. Functionality in the framework.

After clean and sort we see the results in session tab called user wise session which is shown in following figure 4.

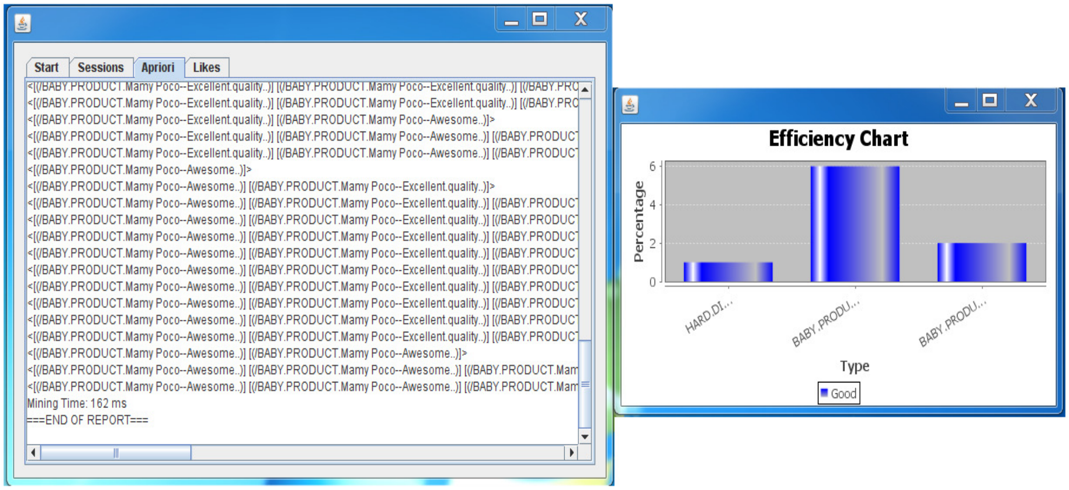


Date	Opinion	Response	Bytes
/Mar/2003:0...	GET /HARD.D...	404	213

Address	Date	Opinion
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P
koerting.hann...	01/Mar/2003:0...	GET /BABY.P

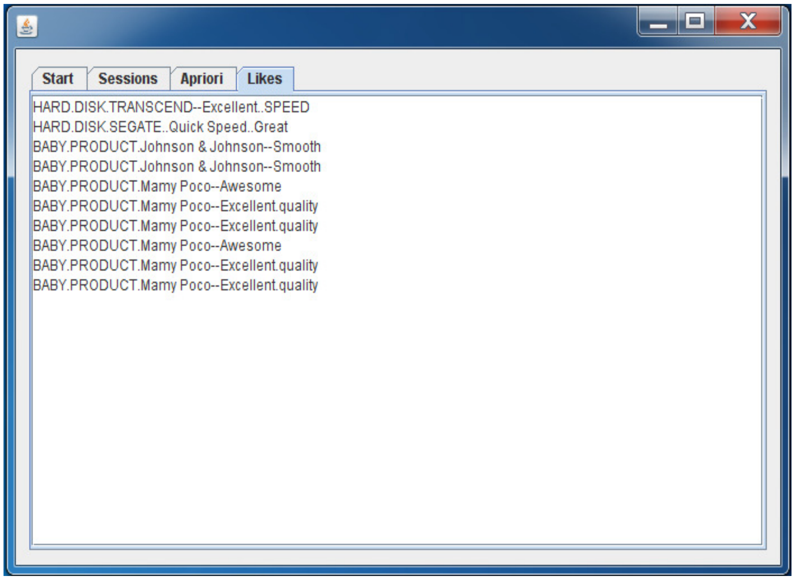
Figure 3. User wise session of a dataset.

To apply the apriori technique, we provide apriori functionality in that we manually give the support and calculate it. The result is in apriori tab with the time required for calculation i.e. mining time in milliseconds with the efficiency chart and theoretical representation of this chart is in likes tab which is shown in following figure 5 and figure 6 respectively.



```
<[(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRO
<[(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRO
<[(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)]>
<[(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)]>
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)]>
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Excellent.quality..)] [(/BABY.PRODUC
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)]>
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mam
<[(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mamy Poco--Awesome..)] [(/BABY.PRODUCT.Mam
Mining Time: 162 ms
===END OF REPORT===
```

Figure 4. Apriori result with efficiency chart.



```
HARD.DISK.TRANSCEND--Excellent..SPEED
HARD.DISK.SEGATE..Quick Speed..Great
BABY.PRODUCT.Johnson & Johnson--Smooth
BABY.PRODUCT.Johnson & Johnson--Smooth
BABY.PRODUCT.Mamy Poco--Awesome
BABY.PRODUCT.Mamy Poco--Excellent.quality
BABY.PRODUCT.Mamy Poco--Excellent.quality
BABY.PRODUCT.Mamy Poco--Awesome
BABY.PRODUCT.Mamy Poco--Excellent.quality
BABY.PRODUCT.Mamy Poco--Excellent.quality
```

Figure 5. Theoretical representation of efficiency chart.

5.CONCLUSIONS

In this paper, we study the various methods and techniques related to different domains like opinion summarization, sentiment analysis, etc. This framework collects the different set of opinions and applies all techniques which we already explained for getting summary. This summary is shown in the form of graphs and also in text which are easily understandable.

In future work, we mainly focus on words which gives different orientation in different situations. Also we are trying to increase accuracy in finding out features which is implicit. We also plan to visualize our final result in more informative and accurate text summary.

REFERENCES

[1] B. A. Tidke, R. G. Mehta, D. P. Rana, “A Novel Approach for High Dimensional Data Clustering”, in International Journal of Engineering Science and Advanced Technology (IJESAT) ISSN 22503676 Vol.02(3) May-Jun 2012.

[2] Mugdha More, Bharat Tidke, "Social media online opinion summarization using ensemble technique," Pervasive Computing (ICPC), 2015 International Conference on , vol., no., pp.1,6, 8-10 Jan. 2015.

[3] Nitin Jindal and Bing Liu, “Review Spam Detection”, Proceedings of 16th International World Wide Web conference, WWW '07. Banff, Alberta, Canada 2007.

[4] Nitin Jindal and Bing Liu, "Opinion Spam and Analysis", Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Feb 11-12, 2008, Stanford University, Stanford, California, USA.

[5] Snehal Dixit and A. J. Agrawal, “Survey On Review Spam Detection”, International Journal of Computer & Communication Technology ISSN, 2013.

[6] Minqing Hu and Bing Liu, "Opinion Extraction and Summarization on the Web", Proceedings of 21st National Conference on Artificial Intelligence (AAAI-2006, Nectar paper), July 16.20, 2006, Boston, Massachusetts, USA.

[7] Qingliang Miao, Qiudan Li, Ruwei Dai, “AMAZING: A sentiment mining and retrieval system”, Expert Systems with Applications 36 (2009) 7192–7198.

[8] Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey, “A Text Mining Technique Using Association Rules Extraction”, World Academy of Science, Engineering and Technology, Vol: 2 2008-06-21.

[9] Long-Sheng Chen, Cheng-Hsiang Liu, Hui-Ju Chiu, “A neural network based approach for sentiment classification in the blogosphere”, Journal of Informetrics 5 (2011) 313–322.

[10] G Vinodhini, RM Chandrasekaran, “Sentiment analysis and opinion mining: a survey”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[11] Bing Liu and Lei Zhang, “A Survey of Opinion Mining and Sentiment Analysis”.

[12] Bo Pang and Lillian Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”, 2004 ACL.

[13] Bing Liu, “Sentiment Analysis and Opinion Mining”, Morgan Claypool Publishers, 2012.

[14] Mahesh Joshi and Carolyn Penstein-Rose, “Generalizing Dependency Features for Opinion Mining”, In ACL, 313–316, 2009.

[15] Xiangyang Lan, Stefan Roth, Daniel Huttenlocher and Michael J. Black, “Efficient Belief Propagation with Learned Higher-Order Markov Random Fields”, In ECCV, 269–282, 2006.

[16] Eric Breck, Yejin Choi and Claire Cardie, “Identifying Expressions of Opinion in Context”, In IJCAI, 2683–2688, 2007.

[17] Yejin Choi and Claire Cardie, “Hierarchical Sequential Learning for Extracting Opinions and Their Attributes”, In ACL, 269–274, 2010.

[18] Alena Neviarouskaya, Helmut Prendinger and Mitsuru Ishizuka, “Semantically distinct verb classes involved in sentiment analysis”, In IADIS, 27–35, 2009.

[19] Ganesan Kavita, Zhai ChengXiang and Han Jiawei , “Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions”, Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), (2010).

[20] Minqing Hu and Bing Liu, "Mining and summarizing customer reviews", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, 2004.

[21] Yihong Gong and Xin Liu, “Generic Text Summarization using Relevance Measure and Latent Semantic Analysis”, SIGIR‟01, September 9-12-2001, New Orleans, Louisiana, USA.

[22] Chris Manning and Hinrich Schütze, “Foundations of Statistical Natural Language Processing”, MIT Press. Cambridge, MA: May 1999.