SUMMARIZING CUSTOMER REVIEWS USING INCREMENTAL STS AND SENTIMENTAL ANALYSIS

Swarna C[#1],Dr. M Sharmila kumari [*2]
M.Tech Student, *Professor
Department Of Computer Science & Engineering, P. A. College of Engineering, Mangalore, 574153, India

Abstract— Now a days, social network services are popular and have become important communication platforms in our daily life. In addition to Facebook, Twitter, YouTube there are platforms available for selling and buying products online such as Amazon, flipcart, eBay etc. Due to the popularity and convenience of e-commerce, many of the users are buying products online. In order to increase the customer satisfaction and own business, vendors set up social pages for interaction with the users .Each customer can leave the message, expressing their opinions about the product. As the quantity of comments is large and generation rate is remarkably high it is a daunting task to go through the whole comment list. Moreover, companies will have high interest to understand how their customers are reacting to certain products. This paper focus on an advanced summarization technique targeting at customer reviews about a product in social network service which automatically distinguish positive and negative reviews. For classifying the comments as either positive or negative SentiWordNet algorithm is used. So the customers seeing the reviews get a quick feedback about the product. Manufacturers can also go through customer review and can take necessary actions to improve their business.

Keywords— **Term Vector, Clustering, Batch STS, Incremental STS, Sentimental Analysis, SentiWordNet**

I INTRODUCTION

Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services [1]. As e-commerce is becoming more and more popular, the number of customer reviews that a product secures increases rapidly. There will be hundreds or thousands comments for a trendy product in the web. This makes it difficult for a potential customer to read all of them to make a decision on whether to purchase the product or not. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions [1].So we want to get the summary of the reviews. We can use clustering algorithms which is a part of data mining. There are mainly two types of clustering algorithms: Partitioning and Hierarchical [2].This work is using incremental clustering algorithm to discover the top-k clusters including different groups of reviews about one product. Initially we are using batch short text summarization algorithm to form th e clusters of comments. After that for each newly coming comment we have to include it in an existing matching cluster. For that we are using incremental clustering algorithm. After that we are performing sentimental analysis to classify it as either positive comment or negative comment.

The rest of the paper is organized as follows. Section II describes the problem. Section III explains the related works. Section IV describes representation of comments as term vectors. Section V briefs the clustering definitions. Section VI illustrates the methodology. Section VII detailed study of present case. Section VIII briefs the work and concludes. Section IX describes the future work.

II PROBLEM DESCRIPTION

Here we are focusing on the customer reviews in social network and produce immediate summary of comments using incremental clustering. The problem we focus, short text summarization, is described as follows. Given a set of reviews S, and the desired number of groups k, find top-k groups {C1,C2,. . . ,Cj ,. . . ,Ck} which have top-k most reviews, and the number of reviews in Cj is larger than or equal to that of reviews in Cj+1 (i.e., $|Cj| \geq |Cj+1|$). Not all reviews in S should be included in top-k groups. Moreover, the reviews in Cj express sim ilar opinions and are a subset of S [3].

Main objective of this work is to discover top-k groups where the reviews in the same group express similar opinions while the reviews belonging to different groups express diverse points of view. Also each review in a cluster is divided in to a positive review or negative review.

Once a product is available online, users can buy it and leave comments and the number of comments may rise quickly and continuously. Manufacturers are usually unwilling to go over the complete list of comments, but they may want to see the summary of customer reviews to know the feedback about product. This indicates that the proposed approach should be able to generate the summary result at any time point of a dynamic data stream. To satisfy this requirement, this problem is modelled as an incremental clustering task. After the clusters are formed we are classifying the positive and negative reviews in the cluster using sentimental analysis. Sentimental analysis is the task of identifying the opinion expressed by a document. Sentimental analysis can be carried out at various levels-word level,s sentence level, Document level etc. Sentiment word is a sentiment lexicon associating sentiment information to each wordnet synset.In this work we use SENTIWORDNET 3.0, an enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [26].The model of the proposed work is described in the following fig 1.

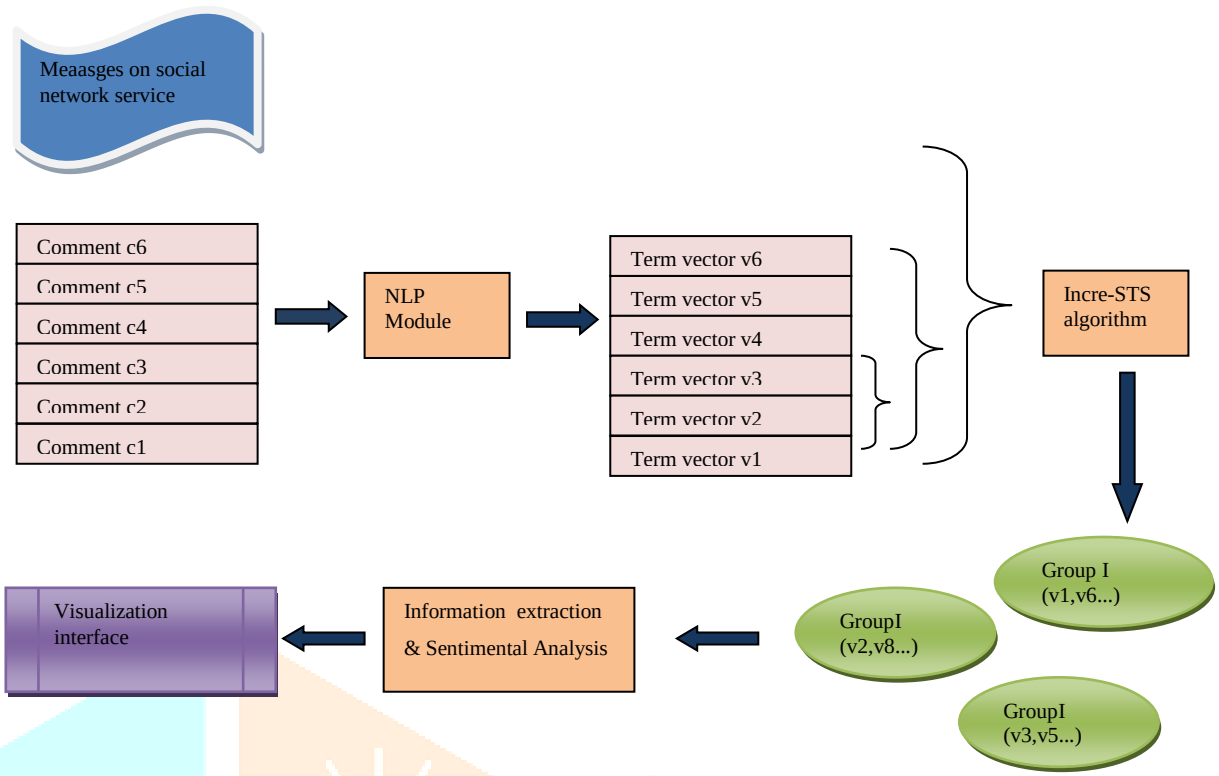

Fig 1: Proposed System Model

III RELATED WORKS

Large quantity of data is generated by user in social network services .So extracting useful knowledge from these data is a current research area. Social network services not only include Facebook, Twitter, WhatsApp etc but also include other web services where users can interact.

"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" explain different types of clustering algorithms and introduce a new clustering algorithm DBSCAN depending on a density-based notion of clusters. It is designed to discover clusters of whimsical shape [2].

In the current field of blogging and social communication services, users post millions of short messages in every minute. It is a tedious process to keep track of all the messages posted by your friends and the conversation as a whole. In "Topical Clustering of Tweets" by Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking presented a study on automatically clustering and classifying Twitter messages, also known as "tweets", into six predefined topics: News, Sports, Entertainment, Science, Technology, Money, and "Just for Fun".[4]

Kushal Dave , Steve Lawrence and David M. Pennock developed a methodology for automatically discriminate between positive and negative reviews in "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". Here a classifier uses information retrieval techniques to extract feature and scoring, and the results for various metrics and heuristics changes depending on the testing situation.[5]

Now a days large number of works are focused on micro-blogging messages. Different types of algorithms with variety of techniques are developed for summarizing. In " TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration" Adam Marcus, Michael S. Bernstein, Osama Badar,David R. Karger, Samuel Madden, Robert C. Miller presents, a module for envisaging and summing up events on Twitter. Throgh TwitInfo users can browse a large group of tweets using a timeline-based display that bring out peaks of high tweet activity.[6]

In [7], issues addressed include removing the noise, determining tweet clusters of interest bearing in mind that the methods must be online, and determining the relevant locations associated with the tweets.

In "IMASS: An Intelligent Microblog Analysis and Summarization System" the authors introduce a novel two-phase summarization scheme. In the first phase,the post plus its responses are classified into four categories based on the purpose, cross-questioning, sharing, and communicating. For each type of post, in the second phase, they apply different strategies, including opinion analysis, response pair identification, and response relevancy detection, to summarize and highlight critical information to display [8].

In [9] take out representative sentences from a blog post which represent the topics covered among its comments. The proposed solution first deduces representative words from comments and then select sentences containing representative words.

In" Comments-Oriented Document Summarization: Understanding Documents with Reader's Feedback " authors analyze the problem of comments-oriented document summarization and trying to cluster a Web document (e.g., a blog post) by considering its contents as well as comments left by its readers.[10].

Research topic on analyzing the product review has a good role in e-commerce. From these both the manufacturers and customers are getting a review about the product and according to that review they can act.[5][1].With data mining ,natural language processing and sentimental analysis are also assimilated to gain various needs of user.

"Extracting descriptions of problems with product and services from twitter data" present a system that filters tweets related to an enterprise and extracts descriptions of problems with their product/service[11].
"Selecting Quality Twitter Content for Events "explore approaches for finding representative messages among a set of Twitter messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information.[12]
Elham Khabiri and James Caverlee and Chiao-Fang Hsu proposes (i) a clustering-based approach for identifying correlated groups of comments;and (ii) a precedence-based ranking framework for automatically selecting informative user-contributed comments in "Summarizing User-Contributed Comments".[13].Though this work is more important for this project work it is not handling it in real time.
"Short and Tweet: Experiments on Recommending Content from Information Streams" authors are focusing on URL recommendations from Twitter Information Streams.
"Short Text Classification in Twitter to Improve Information Filtering " aims at filtering messages for classification .This work effectively classifies the text to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages[14].
In [15] authors propose to explore a variety of text sources for summarizing the Twitter topics, including the tweets, normalized tweets via a dedicated tweet normalization system, web contents linked from the tweets, as well as integration of different text sources.
"Opinion Mining and Sentiment Analysis " deals with the computational treatment of opinion, sentiment, and subjectivity in text. In general sentimental analysis is used to classify messages in to predefined labels such as positive and negative .And this project work also uses sentimental analysis to form positive and negative review about a product.[25].
In "Relevance Modeling for Microblog Summarization" the authors focus on a new summarization technique which aims to synthesize content from multiple microblog posts on the same topic into a human-readable prose description of fixed length [16].Zi Yang, Keke Cai*y*, Jie Tang, Li Zhang*y*, Zhong Su*y* and Juanzi Li proposes a a dual wing factor graph (DWFG) model, which utilizes the mutual reinforcement between Web documents and their associated social contexts to generate summaries[24].
"Incremental Clustering for Mining in a Data Warehousing Environment " by Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, Xiaowei Xu modifies the DBSCAN algorithm using incremental clustering[18].
"SIMFINDER: A Flexible Clustering Tool for Summarization " by Vasileios atzivassiloglou, Judith L. Klavans, Melissa L. Holcombe,Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown present a statistical similarity measuring and clustering tool, SIMFINDER, that organizes small pieces of text from one or multiple documents into tight clusters. By placing highly related text units in the same cluster, SIMFINDER enables a subsequent content selection/generation component to reduce each cluster to a single sentence, either by extraction or by reformulation[19].
In [20] BIRCH: An Efficient Data Clustering Method for Very Large Databases Tian Zhang Raghu Ramakrishnan Miron Livny" incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources .
"Opinion Mining and Sentiment Analysis "by Bo Pang and Lillian Lee is a survey paper which describes the techniques and approaches for opinion-oriented information-seeking systems [25]. "Online new event detection and tracking" ia a paper which describes how to detect a new event in the broadcast stories and finding the relationship between the stories. James Allan and John Papka says that event detection and tracking is a part of Topic Detection and Tracking initiative [21].

IV REPRESENTATION OF COMMENTS AS TERM VECTORS

This section describes how NLP module is converting a comment in to a set of n-grams. Fig 2 Illustrate how to process the comment "Liking this product soooooo much!" and grams will be extracted. At the beginning, for every word unnecessary punctuation marks will be removed by the punctuation removal function. So here the exclamation mark at the end of the comment will be removed. So the comment will become "Liking this product soooooo much". Then redundant characters are also eliminated for assuring the clarity of the message. Hence the word "soooooo" will be replaced by "so". Also all upper case letters will be transformed to lowercase letters. Now the comment is transformed to "like this product so much". We use the standard Porter Stemming Algorithm [47] for stemming process. So in this example the word "liking" is converted to "like". Now n-gram extraction is taking place. For this example the value of n is set as 3.So comment string will be traversed from left to right to produce all 1-gram, 2-gram and 3-gram terms. At the end we are executing stopwords removal process to remove all grams containing only stopwords. In the fig II it is shown as red striked words. So in 1-gram this, so, much are removed. In the 2-gram , term 'so much' is removed.

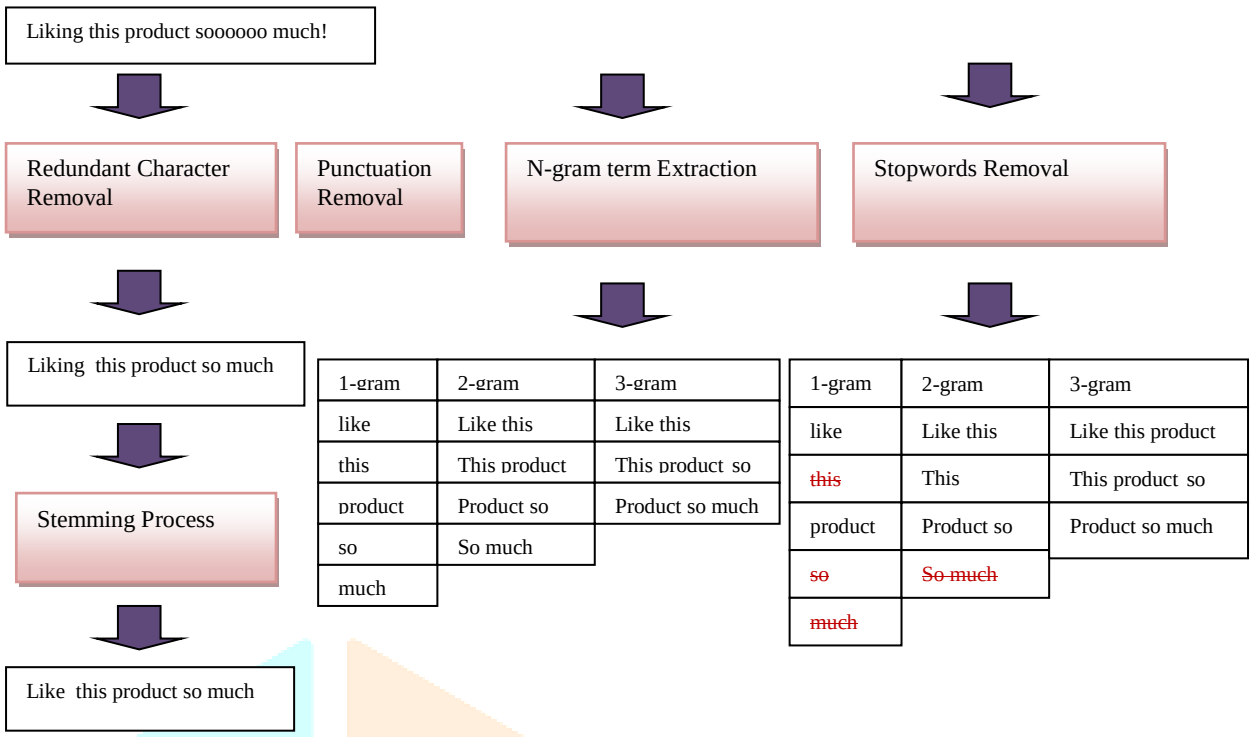

Fig 2:n-gram extraction system model

V DETAILS OF CLUSTERING

Consider two comments represented in the term vector model, $v_a = (t_{1;a}, t_{2;a}, ..., t_{N;a})$ and $v_b = (t_{1;b}, t_{2;b}, ..., t_{N;b})$. Each dimension corresponds to a separate term, and N is the number of dimensions. Since we define that the weights of terms are equal, if the term t_i occurs in the comment v_a, $t_{i;a}$ will be set to 1. Otherwise, $t_{i;a}$ will be set to 0. Thus, we define the modified cosine similarity of two comments as follows [1].

$$sim(v_a, v_b) = \begin{cases} (v_a . v_{b)/D} & \text{if } v_a . v_b \le D \\ 1 & \text{if } v_a . v_b > D \end{cases} \quad (1)$$

where $v_a . v_b$ is the inner product of two vectors, and D is a positive integer constant.
comment-comment distance:-.Distance between two comments v_a and v_b is defined as:

$$distance(v_a, v_b) = \begin{cases} [1/(\sin(v_a,v_b)]-1 & \text{if } sim(v_a, v_b) \neq 0 \\ 1 & \text{if } sim(v_a, v_b) = 0 \end{cases} \quad (2)$$

center of cluster:- Let $\{v_1; v_2, ... v_i, ,...,v_n\}$ be the set of comments belonging to cluster Ce , where $v_i = (t_{1,i}; t_{2,i},..., t_{j,i},..., t_{N,i})$. The center vc_e of cluster Ce is defined as:

$$vce = (tc_{1,c}; tc_{2,c},..., tc_{j,c},..., tc_{N,c}) \quad (3)$$

$$tc_{j,e} = \sum_{p=1}^{n} tj,p \quad (4)$$

Comment-cluster distance:-The similarity between comment v_i and cluster C_e (whose center is vc_e) is defined as:

$$sim(vi,Ce) = \begin{cases} f(vi,Ce)/T & \text{if } f(vi,Ce) \le T \\ 1 & \text{if } f(vi,Ce) > T \end{cases} \quad (5)$$

$$f(V_b, C_s) = \sum_{p=1}^{N} \begin{cases} 2 & if\, t(p,i) * tC(p,e) > 2 \\ t(p,i) * tC(p,s) & otherwise \end{cases}$$

Accordingly, the distance between comment vi and cluster C_e is defined as:

$$\text{distance}(_{vi},C_e) = \begin{cases} [\,1/(\sin(v_i,C_e)]\ -1 & \text{if } \text{sim}(v_i,C_e) \neq 0 \\ 1 & \text{if } \text{sim}(v_i,C_e) = 0 \end{cases} \quad (7)$$

In Equation 5, T is a positive integer constant.

cluster-cluster distance: The distance between two clusters is defined as the distance between two cluster centres derived from Equation 2.
radius of cluster: The radius r_a of cluster Ca is defined as the farthest distance between the center of Ca and any comment in this cluster. Based on the perspective of clustering problem, the short text summarization task is defined as follows.
short text summarization on comment streams: Given a set of comments S, and a desired number of cluster k, find top-k clusters C1,C2,. . . ,Cj ,. . . ,Ck which have top-k most comments, and the number of comments in Cj is larger than or equal to that in Cj+1 .Not all comments in S should be included in top-k clusters, and Cj _ S. In addition, the radius of each cluster should be smaller than the radius threshold Θ_r.

VI METHODOLOGY

This paper aims to provide up-to-date and quick summary of customer opinion and also whether the opinion is positive or negative. Here for short text summarization Incremental clustering method is used. Short text summarization, is described as follows. Given a set of comments S, and the desired number of groups k, find top-k groups {C1,C2,. . . ,Cj ,. . . ,Ck} which have top-k most comments, and the number of comments in Cj is larger than or equal to that of comments in Cj+1 (i.e., |Cj | ≥ |Cj+1|). Not all comments in S should be included in top-k groups[3]. Moreover, the comments in Cj express identical opinions and are a subset of S .

Here each comment is converted in to a set of n-gram terms by the NLP module. After that each n –gram is transformed to its corresponding term vector. Incre-STS Algorithm will find the top-k groups of reviews in real time. Then we will extract the information and perform sentimental analysis using SentiWordNet 3.0 to classify it as positive and negative comments.

6.1. Batch Short text Summarization

Here we are finding the connected components of comment set CS first .If the distance between the cluster and comment is not infinite then this comment will be added to such a cluster. Otherwise a new cluster will be formed with this comment. Then we are also checking the distance between two clusters. If it is not infinite we are merging it in to one. This algorithm also ensures the radius of each cluster is less than the threshold value θ_r.

Algorithm BatchSTS
1. Initialize *C* = Ø;
2. **for** each element *vi* of *CS*
3. **if** there exists any cluster *Cj* where *distance(vi,Cj)* is not infinite
4. Add *vi* into anyone of these clusters;
5. **else**
6. Form a new cluster *Cnew* with the comment *vi*;
7. *C* = *C* ∪ *Cnew*;
8. **for** each non-single-point element *Ci* of *C*
9. **for** each non-single-point element *Cj* of *C* where $i \neq j$
10. **if** *distance(Ci,Cj)* is not infinite
11. Merge *Ci* and *Cj*;
12. **for** each non-single-point element *Ci* of *C*
13. **while** the radius of *Ci* is larger than or equal to θ*r*
14. **for** each comment *vj* in *Ci*
15. **if** *distance(vj,Ci)* ≧ θ*r*
16. Exclude *vj* from *Ci*;
17. Check whether *vj* can be merged with other excluded comments;
18. Output top-k clusters in *C* which have top-k most comments;
End

6.2. Incremental Short text Summarization with sentimental analysis

Second Algorithm we are using is Incremental STS algorithm, for incrementally updating the clustering result with the newly incoming comment. Here the set of previous clustering result C is given as input. Newly incoming comment is denoted by Vnew. And radius threshold is denoted by θ*r*.

1. *Ca* = {*Ci* | *Ci* is an element of *C* ∩ *distance(vnew,Ci)* is not infinite};
2. *Cb*= {*Cj* | *Cj* is an element of *Ca* ∩ *distance(vnew,Cj)* < θ*r*};

3. **if** *Cb* is not empty
4. Add *vnew* into *Cadded* which have most comments in *Cb*;
5. Initialize *Cchanged* = Ø ;
6. **for** each element *Ci* of *Ca* where *Ci* ≠ *Cadded*
7. **for** each comment *vj* in *Ci*
8. **if** *ditances(vj,Cadded)* < θ*r*
9. Add *vj* into *Cadded*;
10. Exclude *vj* from *Ci*;
11. *Cchanged* = *Cchanged* ∪ *Ci*;
12. **for** each element *Ci* of *Cchanged*
13. **while** *V* = {*vj* | *distance(vj,Ci)* ≧ θ*r*} is not empty
14. Exclude all elements in *V* from *Ci*;
15. Try to add each comment in *V* into other clusters
from large to small sizes;
16. **else**
17. Form a new cluster *Cnew* with the comment *vnew*;
18. *C* = *C* ∪ *Cnew*;
19. Output top-k clusters in *C* which have top-k most comments;

End

Newly entered comment will be added to an existing cluster. For that we are finding out the comment to cluster distance .If it is not infinity such clusters are added to $C_{a.}$.From $C_{a.}$ find all clusters whose radius is below Θ_r then add those clusters to $C_{b.}$ Then add the newly entered comment into C_{added} which is the cluster from C_b which has maximum comments. Further we are checking whether comments from other clusters are can be added to $C_{added..}$For that we are checking whether the radius restriction is satisfied. As output we will get the top-k clusters which have top-k comments.

6.3. Generate Summary

Only top-k clusters will be generated by the proposed system .Key terms which are regularly cited will be extracted to construct a key-term cloud. While extracting the top-k terms the problem faced is 1-gram term will prevail over n-gram terms where n is greater than or equal to 2.To solve this first we will check each set of n-gram terms. For particular term t_i in a set, ti will be dropped if there exist another term t_j whose count is larger than or equal to that of t_i, and more than Θ % of words in t_i is present in t_j, where Θ %is the threshold of overlapping percentage. This algorithm will extract the set of representative key terms key-terms denoted as $S_{\text{key-terms}}$

1. Initialize *Skey-terms* = *vc*;
2. **for** each set of n-gram terms in *Skey-terms*
3. Eliminate the terms whose counts do not rank top *k* in this set;
4. **for** each term *ti* in *Skey-terms*
5. **if** there exists any term *tj* where (*tj.ngram == ti.ngram && tj.count >= ti.count*)
6. **if** there are over θ% of words in *ti* also contained in *tj*
7. Eliminate *ti* from *Skey-terms*;
8. **for** each term *ti* in *Skey-terms*
9. **if** there exists any term *tj* where (*tj.ngram* > *ti.ngram*)
10. **if** there are over θ% of words in *ti* also contained in *tj*
11. Eliminate *ti* from *Skey-terms*;
12. Output the set *Skey-terms* of representative key-terms;

End

6.4. Sentimental Analysis

Sentimental analysis is the task of identifying the opinion expressed by a document. Sentimental analysis can be carried out at various levels-word level, sentence level, Document level etc. Sentiment word is a sentiment lexicon associating sentiment information to each wordnet synset.In this work we use SENTIWORDNET 3.0, an enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications [26]. For each wordnet synset the following information is available in SENTIWORDNET.

- Positive Score Pos(s)
- Negative Score Neg(s)
- Objective Score Obj(s)

SentiWordNet system was built in two main steps. In the semisupervised learning step 8 classifiers were established to decide for each synset belonging to WordNet if it is negative, positive or objective. This provides on the one hand a higher generalization factor and a low risk of overfitting, on the other hand the different classification results help to give the synsets a tendency of being more positive or negative, rather than just one opportunity: namely that a synset can positive, negative and objective to a certain extend. In the second step, the randomwalk step the scores for the positive and negative scores are due to the "defiens-defiendum" relationship. Through averaging of all the classification results a value between 0.0 and 1.0 can be obtained for each

category for each synset. If all classifiers will decide on the same category, this sentiment will have the maximum value, which is 1.0.

VII RESULTS AND DISCUSIION

Here we demonstrate the real case the summarization and extraction of positive and negative comments. Also We will get the summary of comments and keyterms used by the customers.

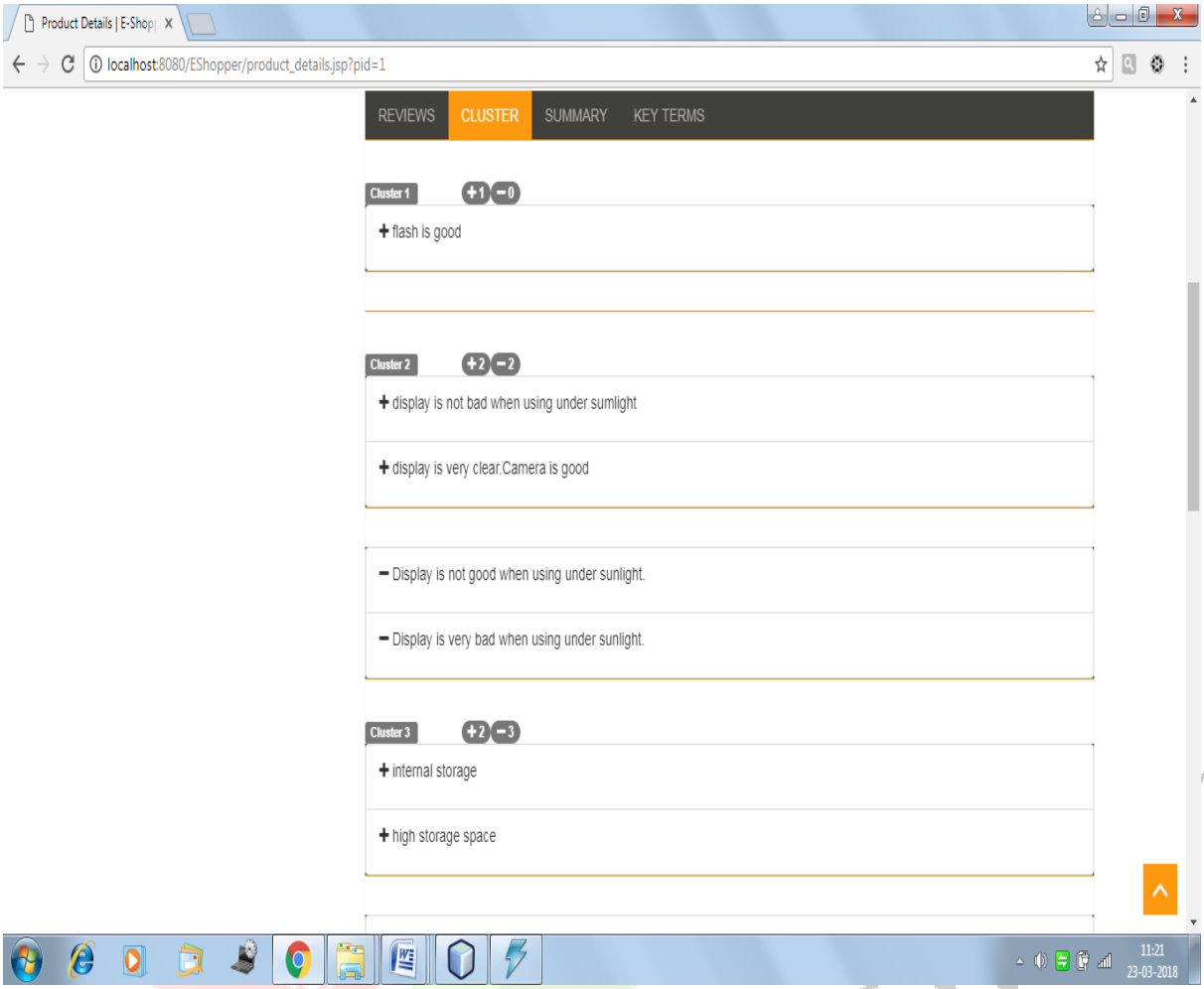

Fig 3: Cluster of comments with sentimental Analysis

Whenever a customer clicks on a product in the website this page will be displayed. Here he can view all reviews about the product. Here in Fig 3 three clusters are shown with sentimental analysis. Each cluster stands for a group of comments expressing similar opinion. Positive comments are marked with '+' sign and negative comments are marked with '-'sign. Total number of positive and negative comments is displayed on the top of the cluster. While clicking the summary and Key terms on the top it will display the summary of comments and will extract the key terms.

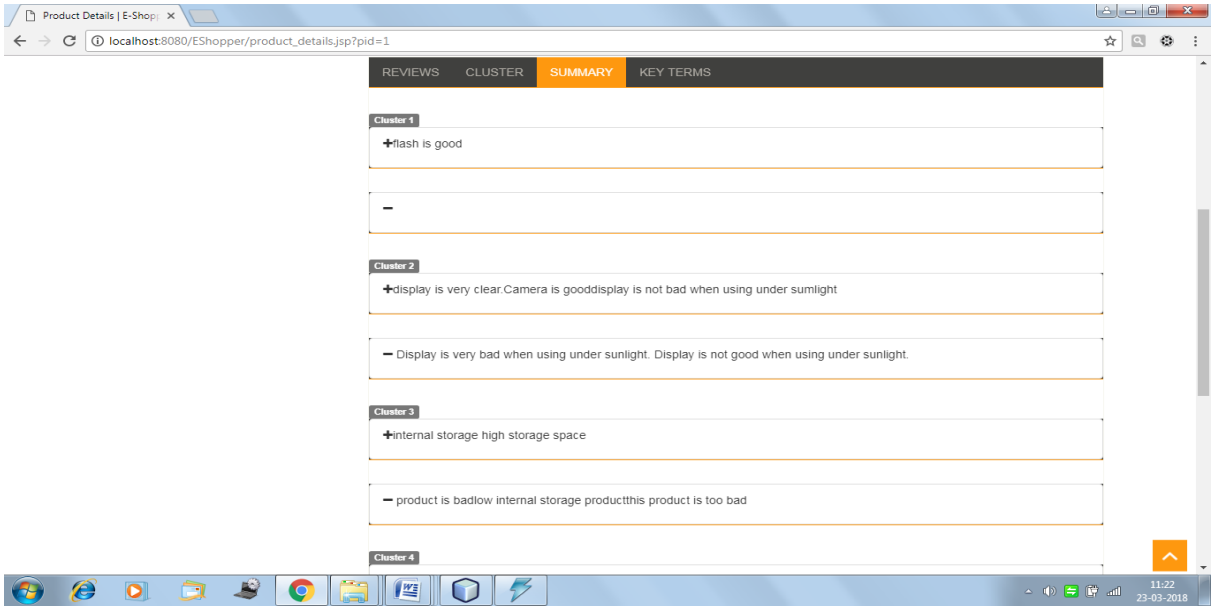

Fig 4 Summary of Comments

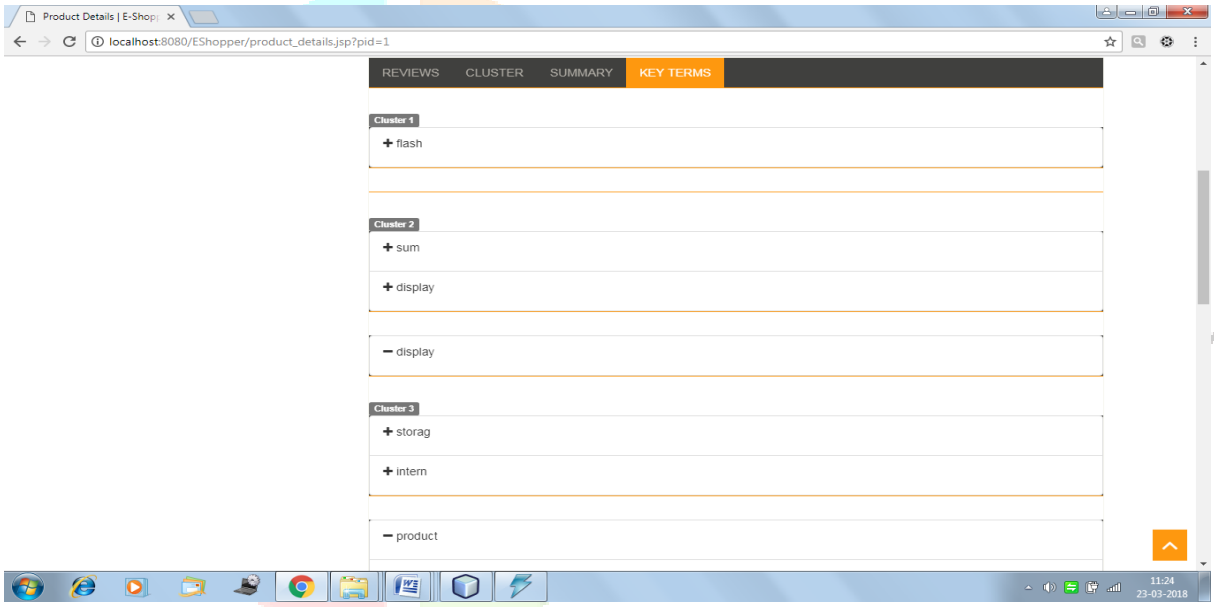

Fig 5 Display of Key Term

VIII CONCLUSION

This paper focus on an advanced summarization technique targeting at customer reviews about a product in social network service which automatically distinguish positive and negative reviews .For classifying the comments as either positive or negative SentiWordNet algorithm is used. So the customers seeing the reviews get a quick feedback about the product. Also Companies have a great interest in their customers' feedback and if many of those customers express their opinion online in text format, companies prefer to find and analyse them automatically. With the help customers' opinions they can adapt their future plans and needs and consequently increase their profit. For adding the newly entered comments to a cluster in real time ,incremental clustering algorithm is used. Sentimental analysis is performed on term vectors to identify whether that particular comment represents a positive comment or negative comment. Also key terms are extracted and displayed for users who require a quick overview about the product.

IX FUTURE WORK

This paper is using only basics of Natural Language Processing to extract the n-grams. And the separation of positive and negative comments are done using SentiWordNet 3.0 algorithm. This system can modify using detailed natural language processing techniques. Also SentiWordNet with machine learning techniques can be incorporated which will analyse the comments entered by the user and separate the positive and negative comments more efficiently.

REFERENCES

[1]. Mining and Summarizing Customer Reviews **:**Minqing Hu and Bing Liu Department of Computer Science University of Illinois at Chicago, KDD'04, August 22–25, 2004, Seattle, Washington, USA

[2]. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, Published in Proceedings of 2nd nternational Conference on Knowledge Discovery and Data Mining (KDD-96)

[3]. IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services Cheng-Ying Liu, Chi-Yao Tseng, Ming-Syan Chen, Fellow, IEEE, DOI 10.1109/TKDE.2015.2405553, IEEE Transactions on Knowledge and Data Engineering.

[4]. Topical Clustering of Tweets :Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking Language Technologies Institute Carnegie Mellon University 5000 Forbes Ave. Pittsburgh, PA, USA SWSM'10, July 28, 2011, Beijing, China. Copyright 2011 ACM 1-58113-000-0/00/0010

[5]. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews Kushal Dave, NEC Laboratories America,Steve Lawrence,NEC Laboratories America David M. Pennock WWW 2003, May 20–24, 2003, Budapest, Hungary.ACM 581136803/03/0005

[6]. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration Adam Marcus, Michael S. Bernstein, Osama Badar,David R. Karger, Samuel Madden, Robert C. Miller CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.Copyright 2011 ACM 978-1-4503-0267-8/11/05

[7]. TwitterStand: News in Tweets Jagan Sankaranarayanany,jagan@cs.umd.edu,Hanan Samety hjs@cs.umd.edu, ACM GIS '09, November 4.6, 2009. Seattle, WA, USA

[8]. IMASS: An Intelligent Microblog Analysis and Summarization System:Jui-Yu Weng Cheng-Lun Yang Bo-Nian Chen Yen-Kai Wang Shou-De Lin, Proceedings of the ACL-HLT 2011 System Demonstrations, pages 133–138,Portland, Oregon, USA, 21 June 2011 58

[9]. Comments-Oriented Blog Summarization by Sentence Extraction Meishan Hu, Aixin Sun and Ee-Peng Lim Centre for Advanced Information Systems School of Computer Engineering Nanyang Technological University, Singapore CIKM'07, November 6–8, 2007, Lisboa, Portugal. Copyright 2007 ACM 978-1-59593-803-9/07/0011

[10]. Comments-OrientedDocument summarization:Understanding Documents with Readers' Feedback Meishan Hu, Aixin Sun, and Ee-Peng Lim School of Computer Engineering, Nanyang Technological University, Singapore SIGIR'08, July 20–24, 2008, Singapore.Copyright 2008 ACM 978-1-60558-164-4/08/07

[11]. Extracting descriptions of problems with product and services from twitter data Narendra K. Gupta,AT&T Labs - Research, Inc Florham Park, NJ 07932 – USA

[12]. Selecting Quality Twitter Content for Events, Hila Becker , Columbia University hila@cs.columbia.edu, Mor Naaman, Rutgers University,mor@rutgers.edu, ,LuisGravano,Columbiauniversity,gravano@cs.columbia.edu 12

[13]. Summarizing User-Contributed Comments Elham Khabiri and James Caverlee and Chiao-Fang Hsu Copyright _c 2011, Association for the Advancement of Artificial Intelligence

[14]. Short Text Classification in Twitter to Improve Information Filtering:Bharath Sriram, David Fuhry,Engin Demir, Hakan Ferhatosmanoglu Computer Science and Engineering Department,Ohio State University, Columbus, OH 43210, USA ,Murat Demirbas Computer Science and Engineering Department,University at Buffalo, SUNY, NY 14260, USA,demirbas@cse.buffalo.edu SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

[15]. Why is "SXSW" trending? Exploring Multiple Text Sources for Twitter Topic Summarization Fei Liu1 Yang Liu1,FuliangWeng Computer Science Department, The University of Texas at Dallas,Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 66–75,Portland, Oregon, 23 June 2011.

[16]. Relevance Modeling for Microblog Summarization,Sanda Harabagiu University of Texas at Dallas Richardson, Texas USA *sanda@hlt.utdallas.edu*,Andrew Hickl Language Computer Corporation Richardson, Texas USA *andy@languagecomputer.com* ,Copyright _c 2011, Association for the Advancement of Artificial Intelligence

[17]. Social Context Summarization :Zi Yang, Keke Cai*y*, Jie Tang, Li Zhang*y*, Zhong Su*y* and Juanzi Li*, SIGIR'11,* July 24–28, 2011, Beijing, China.Copyright 2011 ACM 978-1-4503-0757-4/11/07

[18]. Incremental Clustering for Mining in a Data Warehousing Environment Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, Xiaowei Xu Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 München, Germany ,Proceedings of the 24th VLDB Conference New York, USA, 1998

[19]. SIMFINDER: A Flexible Clustering Tool for Summarization Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe,Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown Department of Computer Science Columbia University 1214 Amsterdam Avenue,New York, NY 10027, USA

[20]. BIRCH: An Efficient Data Clustering Method for Very Large Databases Tian Zhang Raghu Ramakrishnan Miron Livny" SIGMOD '96 6/96 Montreal, Canada IQ 1996 ACM 0-89791 -794-4/96/0006

[21]. Online new event detection and tracking, James Allan, Ron Papka, Victor Lavrenko,Center for Intelligent Information Retrieval,University of Massachusetts

[22]. Learning Similarity Metrics for Event Identification in Social Media, Selecting Quality Twitter Content for Events, Hila Becker , Columbia University hila@cs.columbia.edu, Mor Naaman, Rutgers University,mor@rutgers.edu, ,Luis Gravano Columbia University,gravano@cs.columbia.edu WSDM'10, February 4–6, 2010, New York City, New York, USA.Copyright 2010 ACM 978-1-60558-889-6/10/02

[23]. OPTICS: Ordering Points To Identify the Clustering Structure Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander Institute for Computer Science, University of Munich, Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[24]. Short and Tweet: Experiments on Recommending Content from Information Streams Jilin Chen*, Rowan Nairn†, Les Nelson†, Michael Bernstein, Ed H. Chi April 10–15, 2010, Atlanta, Georgia, USA.Copyright 2010 ACM 978-1-60558-929-9/10/04 17

[25]. Opinion Mining and Sentiment Analysis:Bo Pang, Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089, USA, and Lillian Lee,Computer Science Department, Cornell University, Ithaca, NY 14853,USA, llee@cs.cornell.edu, Foundations and Trends in Informatio n Retrieval Vol. 2, Nos. 1–2 (2008) 1–135

[26]. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche Via Giuseppe Moruzzi 1, 56124 Pisa, Italy E-mail: hf irstname.lastnamei@isti.cnr.it