REVIEW SUMMARIZATION

PROJECT REPORT

Submitted by

KEERTHANA A CCV15CS017

SREELAKSHMI PV CCV15CS046

RANJU CT LCCV15CS063

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

COCHIN COLLEGE OF ENGINEERING & TECHNOLOGY

VALANCHERY

A P J ABDUL KALAM TECHNOLOGICAL UNIVERSITY

NOVEMBER-2018

COCHIN COLLEGE OF ENGINEERING & TECHNOLOGY

VALANCHERY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

Certified that this project report “Review Summarization” is the Bonafide work of “KEERTHANA A, SREELAKSHMI PV, RANJU CT of department of Computer Science & Engineering in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science & Engineering under the A P J Abdul Kalam Technological University the year 2015-2019.

Project Guide **Head of the Department**

Place: Valanchery
Date:

ACKNOWLEDGEMENT

Every success stands as a testimony not only to the hardship but also to hearts behind it. Likewise, the present project work has been undertaken and completed with direct and indirect help from many people and we would like to acknowledge the same.

First and foremost we take immense pleasure in thanking the **Management** and respected principal**, Mr. Sakkariya Thodungal,** for providing us with the wider facilities.

We express our sincere thanks to **Ms. Alma Mary Margret**, Head of Department of Computer Science and Engineering for giving us opportunity to present this project and for timely suggestions.

We wish to express our deep sense of gratitude to the project coordinator Ms. Alma Mary Margret, Asst. professor, Department of Computer Science and Engineering, who coordinated in right path. Words are inadequate in offering our thanks to **Guide Ms. Naeema Mohamed Kutty** Asst. professor Department of Computer Science and Engineering, for her encouragement and guidance in carrying out the project

Needless to mention that the **teaching and the non-teaching faculty members** had been the source of inspiration and timely support in the conduct of our project. We would like to express our heartfelt thanks to our beloved **parents** for their blessings, our **classmates** for their help and wishes for the successful completion of this project.

Above all we would like to thank the **Almighty God** for the blessings that helped us to complete the venture smoothly.

ABSTRACT

In the recent past, e-commerce sites have made rapid growth. There are thousands of products and various websites sell products. Massive growth in the number of reviews and their availability along with the advent of opinion-rich review forums for the products sold online, choosing the right one from a large number of products has become difficult for the users. HELP-ME- BUY is a website that assists buyers in online shopping. It is imminent for buyers to verify for genuineness and quality of products. What better way is there than to ask people who have already bought the product? This is when customer reviews come into picture. The major hitch here is popular products have thousands of reviews-we do not have the time or patience to read all thousands of them. Hence, our application eases this task by analyzing and summarizing all reviews which will help the user decide what other buyers have experienced on buying this product. We carry out this process by a number of modules that include feature extraction and opinion extraction which improves the process of analysis and helps in the formation of an efficient summary.

This paper focuses on the problem of summarization on the reviews of a specific product from the E-commerce sites. Review summarization is a process of extracting and collecting review which has been posted on the e-commerce sites. Motivated by the fact that users may desire to get a brief understanding of reviews without reading the whole review list, we attempt to group reviews with similar content together and generate a concise opinion summary for that review. Review summarization helps to gain important information about any product on a less time. This website provides a platform to see all the reviews and buy the product, it is very easy to maintain all the records of a review.

Keywords: Natural Language processing, Machine Learning, Raw text analysis, summarization.

TABLE OF CONTENTS

PAGE NO

LIST OF FIGURES

ABBREVIATIONS

DFD	Data Flow Diagram
GUI	Graphical User Interface
IDE	Integrated Development Environment
NLP	Natural language processing
RDBMS	Relational Database Management System
SVM	Support Vector Machine
WYSIWYG	What You See Is What You Get

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

With the increasing use of e-commerce websites our project improves the shopping experience for users. The domain of our project lies under Natural Language Processing (NLP) which basically includes analysis, classification and summarization of raw text obtained from customer reviews. There are thousands of products and various websites sell the products. Massive growth in the number of reviews and their availability along with the advent of opinion-rich review forums for the products sold online, choosing the right one from a large number of products has become difficult for the users. In this project, the design of a unified opinion mining and sentiment analysis framework is presented with natural language processing approach. We propose a dynamic system for feature based review summarization based on the corresponding domains of products. In this process, the reviews are extracted by web crawling. These compound sentences are broken down into individual sentences and further into words by sentence-tokenization and word tokenization respectively. The words are POS (Parts of Speech Tagging) tagged to help classify their position in a sentence that will help in extraction of features and opinions. Now, identification and extraction of the features of a product is done first. Next the opinion regarding these identified features is found and their polarity (negative/positive/neutral) is detected. Once this is done, excerpts with respect to these feature-opinion pair are extracted and further used for summarization. This summarized review provides a complete overview of opinions of users and also stresses on each feature of the product, making it easier for both customers and also the producers to know the response of the mass.

This paper focuses on the problem of summarization on the reviews of a specific product from the E-commerce sites. Review summarization is a process of extracting and collecting review which has been posted on the e-commerce sites. Motivated by the fact that users may desire to get a brief understanding of reviews without reading the whole review list, we attempt to group reviews with similar content together and generate a concise opinion summary for that review.

1.2 PROJECT SCOPE

In the recent past, e-commerce sites have made rapid growth. As the number of reviews are in terms of hundreds on certain products and in terms of thousands on popular products it is evident that the user may not read all the reviews and might miss out on some critical reviews that concern his needs. Hence we provide a solution to summarize it based on the product's features. This saves the time and energy of the users which would rather be well spent. The user will be able to decide on one look of the graphical outcome of the summarization. The users for this project would be all the customers who buy the products online.

Review summarization is a process of extracting and collecting reviews which posted on sites. Review summarization helps to gain important information about any product on a less time. This system is mainly suitable for customer and marketer. It provides a platform to see all reviews and buy the product, it is very easy to maintain all the records of a reviews.

CHAPTER 2
LITERATURE SURVEY

2.1 OPINION OBSERVER

The Web has become an excellent source for gathering consumer opinions. There are now numerous Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs. This paper focuses on online customer reviews of products. It makes two contributions. First, it proposes a novel framework for analyzing and comparing consumer opinions of competing products. A prototype system called Opinion Observer is also implemented. The system is such that with a single glance of its visualization, the user is able to clearly see the strengths and weaknesses of each product in the minds of consumers in terms of various product features. This comparison is useful to both potential customers and product manufacturers. For a potential customer, he/she can see a visual side-by-side and feature-by feature comparison of consumer opinions on these products, which helps him/her to decide which product to buy.

For a product manufacturer, the comparison enables it to easily gather marketing intelligence and product benchmarking information. Second, a new technique based on language pattern mining is proposed to extract product features from Pros and Cons in a particular type of reviews. The system focused on one type of opinion sources, customer reviews of products. And the system is used to compare consumer opinions of multiple products. To support visual analysis, we designed a supervised pattern discovery method to automatically identify product features from Pros and Cons in reviews of format. A friendly interface is also provided to enable the analyst to interactively correct errors of the automatic system, if needed, which is much more efficient than manual tagging. The system is highly effective. Here the system doesn’t evaluate the strength of opinions, and It is not suitable to simply compare the set of extracted features (no duplicates) from all reviews of a product with the set of manually identified features as it does not measure how effective the extraction is for individual reviews.

2.2 MINING THE PEANUT GALLERY

The web contains a wealth of product reviews, but sifting through them is a daunting task. ideally, an opinion mining tool would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good). We begin by identifying the unique properties of this problem and develop a method for automatically distinguishing between positive and negative reviews. These results perform as well as traditional machine learning methods. We then use the classifier to identify and classify review sentences from the web, where classification is more difficult. Our classifier draws on information retrieval techniques for feature extraction and scoring, and the results for various metrics and heuristics vary depending on the testing situation. The best methods work as well as or better than traditional Machine learning. When operating on individual sentences collected from web searches, performance is limited due to noise and ambiguity. But in the context of a complete web-based tool and aided by a simple method for grouping sentences into attributes, the results are qualitatively quite useful. The number of issues that make this problem difficult are Rating inconsistency, Ambivalence and comparison, Sparse data, Skewed distribution.

These challenges may be why traditional machine learning techniques (like SVMs) and common metrics (like mutual information) do not do as well as our bias measure with n-grams on the two tests. Few refinements improved performance in both cases. Encouragingly, two key innovations metadata substitutions and variable length features were helpful. Extraction proved more difficult. It may be that features that are less successful in classification, like substrings, do better in mining because they are more specific. More work is needed on separating genre classification from attribute and sentiment separation. The best methods work as well as or better than traditional Machine learning. When operating on individual sentences collected from web searches, performance is limited due to noise and ambiguity. But in the context of a complete web-based tool and aided by a simple method for grouping sentences into attributes, the results are qualitatively quite useful. The number of issues that make this problem difficult are Rating inconsistency, Ambivalence and comparison, Sparse data, Skewed distribution.

2.3 MINING AND SUMMARIZING CUSTOMER REVIEWS

Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. As ecommerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. For the manufacturer, there are additional difficulties because many merchant sites may sell the same product and the manufacturer normally produces many kinds of products. In this research, we aim to mine and to summarize all the customer reviews of a product.

This summarization task is different from traditional text summarization because we only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. We do not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Our task is performed in three steps: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results.

The objective is to provide a feature-based summary of a large number of customer reviews of a product sold online. Our experimental results indicate that the proposed techniques are very promising in performing their tasks. We believe that this problem will become increasingly important as more people are buying and expressing their opinions on the Web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. We believe that this problem will become increasingly important as more people are buying and expressing their opinions on the Web.

2.4 LEARNING SIMILARITY METRICS

Social media sites (e.g., Flickr, YouTube, and Facebook) are a popular distribution outlet for users looking to share their experiences and interests on the Web. These sites host substantial amounts of user-contributed materials (e.g., photographs, videos, and textual content) for a wide variety of real-world events of different type and scale. By automatically identifying these events and their associated user-contributed social media documents, which is the focus of this paper, we can enable event browsing and searching state-of-the-art search engines. To address this problem, we exploit the rich ”context” associated with social media con- tent, including user-provided annotations (e.g., title, tags) and automatically generated information (e.g., content creation time).

Using this rich context, which includes both textual and non-textual features, we can define appropriate document similarity metrics to enable online clustering of media to events. As a key contribution of this paper, we explore a variety of techniques for learning multi-feature similarity metrics for social media documents in a principled manner. Our similarity metric learning techniques yield better performance than the baselines on which we build. Our classification- based techniques show significant improvement over traditional approaches that use text-based similarity. It is focuses on event documents only. By automatically identifying these events and their associated user-contributed social media documents, which is the focus of this paper, we can enable event browsing and searching state-of-the-art search engines. As a key contribution of this paper, we explore a variety of techniques for learning multi-feature similarity metrics for social media documents in a principled manner.

2.5 EDDI

Twitter streams are on overload: active users receive hundreds of items per day, and existing interfaces force us to march through a chronologically ordered morass to find tweets of interest. We present an approach to organizing a user’s own feed into coherently clustered trending topics for more directed exploration. Our Twitter client, called Eddi, groups tweets in a users feed into topics mentioned explicitly or implicitly, which users can then browse for items of interest. To implement this topic clustering, we have developed a novel algorithm for discovering topics in short status updates powered by linguistics syntactic transformation and call outs to a search engine. An algorithm evaluation reveals that search engine call outs out perform other approaches when they employ simple syntactic transformation and back off strategies. Active Twitter users evaluated Eddi and found it to be a more efficient and enjoyable way to browse an overwhelming status update feed than the standard chronological interface.

Twitter streams are on overload: active users receive hundreds of items per day, and existing interfaces for ceus to march through a chronologically-ordered morass to find tweets of interest. We present an approach to organizing a user's own feed into coherently clustered trending topics for more directed exploration. Our Twitter client, called Eddi, groups tweets in a users feed into topics mentioned explicitly or implicitly, which users can then browse for items of interest. To implement this topic clustering, we have developed a novel algorithm for discovering topics in short status updates powered by linguistic syntactic transformation and callouts to a search engine. An algorithm evaluation reveals that search engine callouts outperform other approaches when they employ simple syntactic transformation and back off strategies. Active Twitter users evaluated Eddi and found it to be a more efficient and enjoyable way to browse an overwhelming status update feed than the standard chronological interface.

CHAPTER 3
DESIGN METHOD

The main objective is to cluster reviews with content similarity, semantic similarity and generate a concise opinion summary for the review. There is a need to discover how many different group opinions exist and provide an overview of each group to make users easily and rapidly understand. Therefore, here the goal is developing an efficient and effective technique to identify the clusters of these comments. We model a novel incremental clustering problem based on the requirements of review summarization. The summarization algorithm starts when the user selects a product from the list. NLP module that transforms each comment into a set of n-gram terms. The process of n-gram terms extraction is carried out to extract terms that are used for representing this comment.

We propose IncreSTS algorithm that can incrementally update clustering results with incoming reviews. The primary concept of IncreSTS is to maintain the clustering result of the previous phase, and to incrementally update the clustering result with the reviews. Key-term extraction is a way to extract the top-k terms with the k most frequency counts from the cluster center. However, such strategy will lead to the problem that 1- gram terms dominate over n-gram terms, where n is an integer and larger than or equal to 2. To remedy this defect, in each set of n-gram terms, top-k terms with k most counts are extracted. We design an at-a-glance presentation, which is concise, informative, and impressive, to help users easily and rapidly get an overview understanding of a review.

3.1 FEASIBILITY STUDY

3.1.1 TECHNICAL FEASIBILITY

A study of function, performance and constraints may improve the ability to create an acceptable system. Technical Feasibility is frequently the most difficult area to achieve at the stage of product engineering process. Our system needs a browser. Browser is used to open our website, search the E commerce sites, view product, view reviews. So the system is technically feasible.

3.1.2 OPERATIONAL FEASIBILITY

The purpose of the operational feasibility study is to determine whether the new system will be used if it is developed. There is no difficulty in, implementing the system and the proposed system is so effective, user friendly and functionally reliable so that the users can view the summarized comments from any e commerce sites. If the user of the system is fully aware of the internal working of the system then the users will not be facing any problem in running the system. So the system is operationally feasible.

3.1.3 ECONOMIC FEASIBILITY

Proposed system is developed with the available resources. Since cost input for the software is almost nil the output of the software is always a profit. Hence software is economically feasible. In the existing system, storage of the records and summarization is very difficult. In the proposed system, only needs a browser, no need of extra charges. So the system is economically feasible.

3.2 SYSTEM DESIGN

3.2.1 SYSTEM ARCHITECTURE DESIGN

A system architecture is the conceptual model that defines the structure, behaviour and more views of a system. Once a review is posted on e-commerce sites, users can leave reviews immediately and the number of reviews may rise quickly and continuously. Moreover, readers are usually unwilling to go over the whole list reviews, but they may request to see the summary at any moment. This indicates that the IncreSTS approach should be able to generate the summary result at any time point of a dynamic data stream. To satisfy this requirement, here model this problem as an incremental clustering task. The System architecture of INCRESTS adopt the term vector model, and there for each review is transformed into a set of n-gram terms by the NLP module.

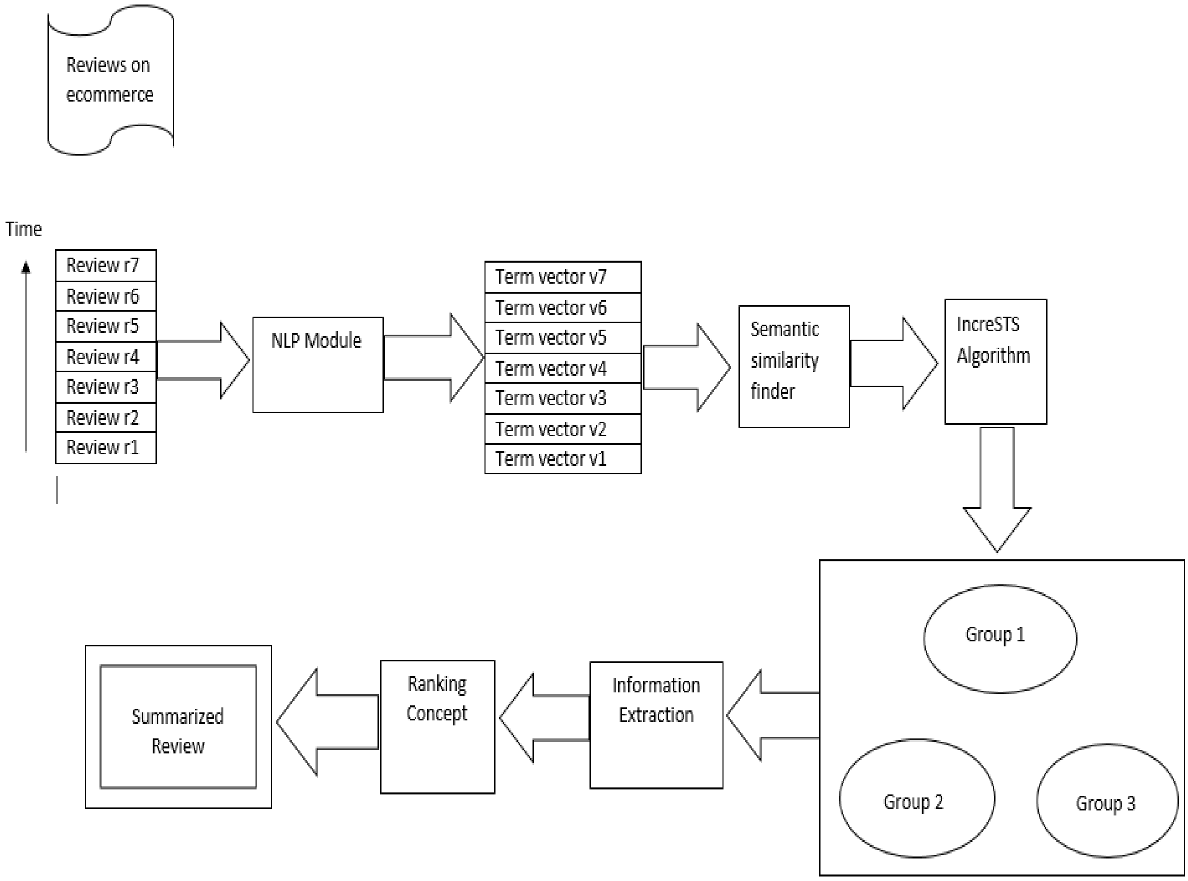

Fig 3.1 System architecture

Since informal and unstructured texts are widely used on the reviews, and also apply some heuristics to enhance the quality of n-gram terms that can better represent each comment. Here a semantic similarity finder is used to check the semantic similarity between reviews. To decide whether two words are semantically similar, it is important to know these mantic relations that hold between the words. For example, the words horse and cow can be considered semantically similar because both horses and cows are useful animals in agriculture.

To compute the semantic similarity between two words are relational model is used. First, using snippets retrieved from a web search engine, an automatic lexical pattern extraction algorithm is used to represent the semantic relations that exist between two words. Whenever a request is received, the INCRESTS algorithm incrementally producing latest clustering results and simultaneously outputting significant reviews that are closest to the centre of each cluster. Finally, for the visualization interface, representative terms will be extracted to form a key-term cloud for each group. Thus, users will be provided a concise, informative, and at-a-glace presentation that can help them easily comprehend the main points of responses to one review.

3.2.2 USE CASE DIAGRAM

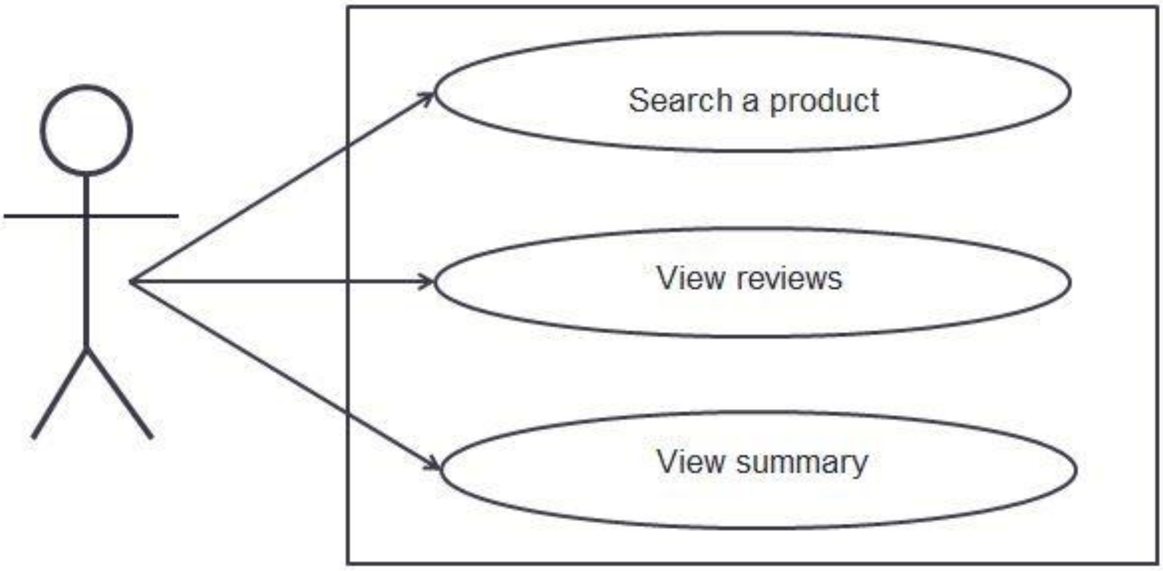

Fig 3.2 Use Case Diagram

A use case diagram is a graphic depiction of the interactions among the elements of a system. The actors, usually individuals involved with the system defined according to their roles. A use case represents a part of the functionality of the system and enables the user (modeled as an actor) to access this functionality. An association is a connection between an actor and a use case. An association indicates that an actor can carry out a use case. Several actors a tone use case mean that each actor can carry out the use case on his or her own and not that the actors carry out the use case together. An include relationship is a relationship between two use cases.

In this system, there is one user and the use cases are search a product, view reviews and view summary. After login to the system user can search the e-commerce sites and search products, then get the reviews and view summary.

3.2.3 DATA FLOW DIAGRAM

A dataflow diagram (DFD) illustrates how data is processed by a system in terms of inputs and outputs. .It is usually beginning with a context diagram as the level 0 of DFD diagram, a simple representation of the whole system. To elaborate further from that, it is drilled down to a level 1 diagram with lower level functions decomposed from the major functions of the system. This could continue to evolve to become a level 2 diagram when further analysis is required. Progression to level 3, 4 and so on is possible but anything beyond level 3 is not very common.

In level 0, User request for the summarization of reviews from e-commerce sites through IncreSTS. System gives response to the user. In level1 the user is the entity and the processes are search page, view review and summarization. User search pages from filpkart and the reviews are stored in review table. User gets the summary of the comments and then summaries are stored in the summary table. In level 1.1, comments fetch through batchSTS and IncreSTS.

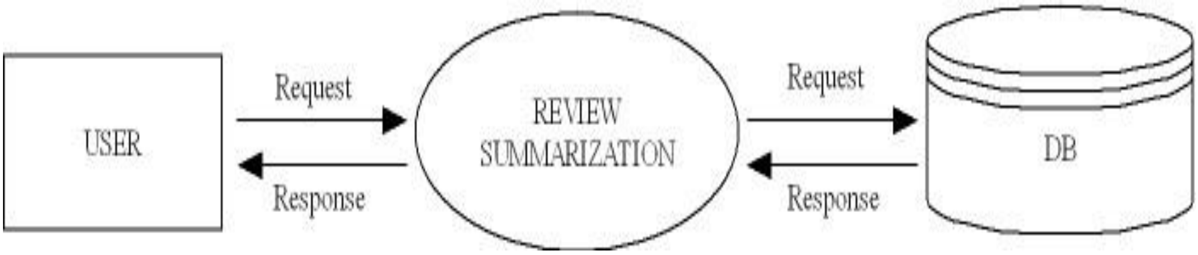

Fig 3.3 Level 0 Data Flow Diagram

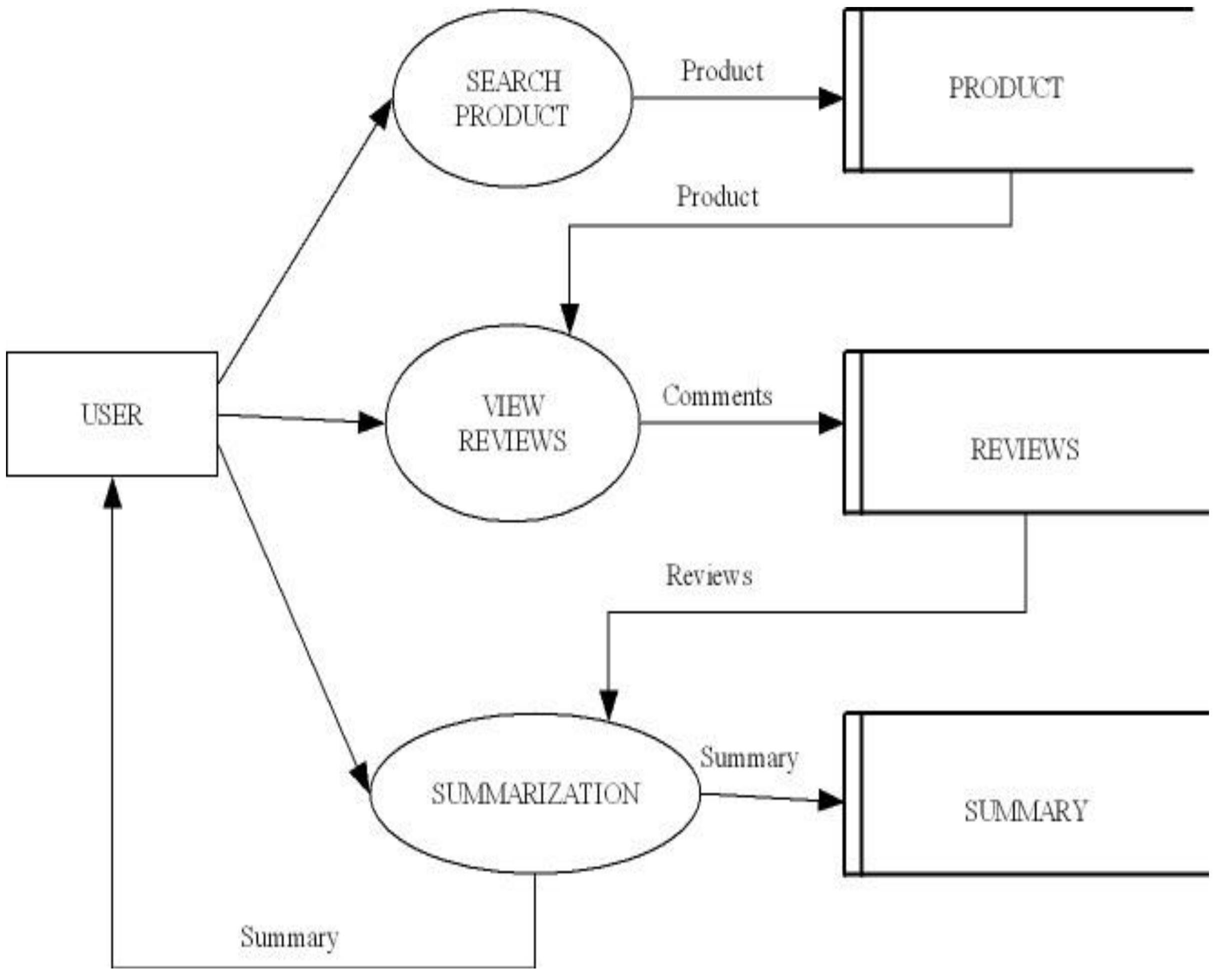

Fig 3.4 Level 1 Data Flow Diagram

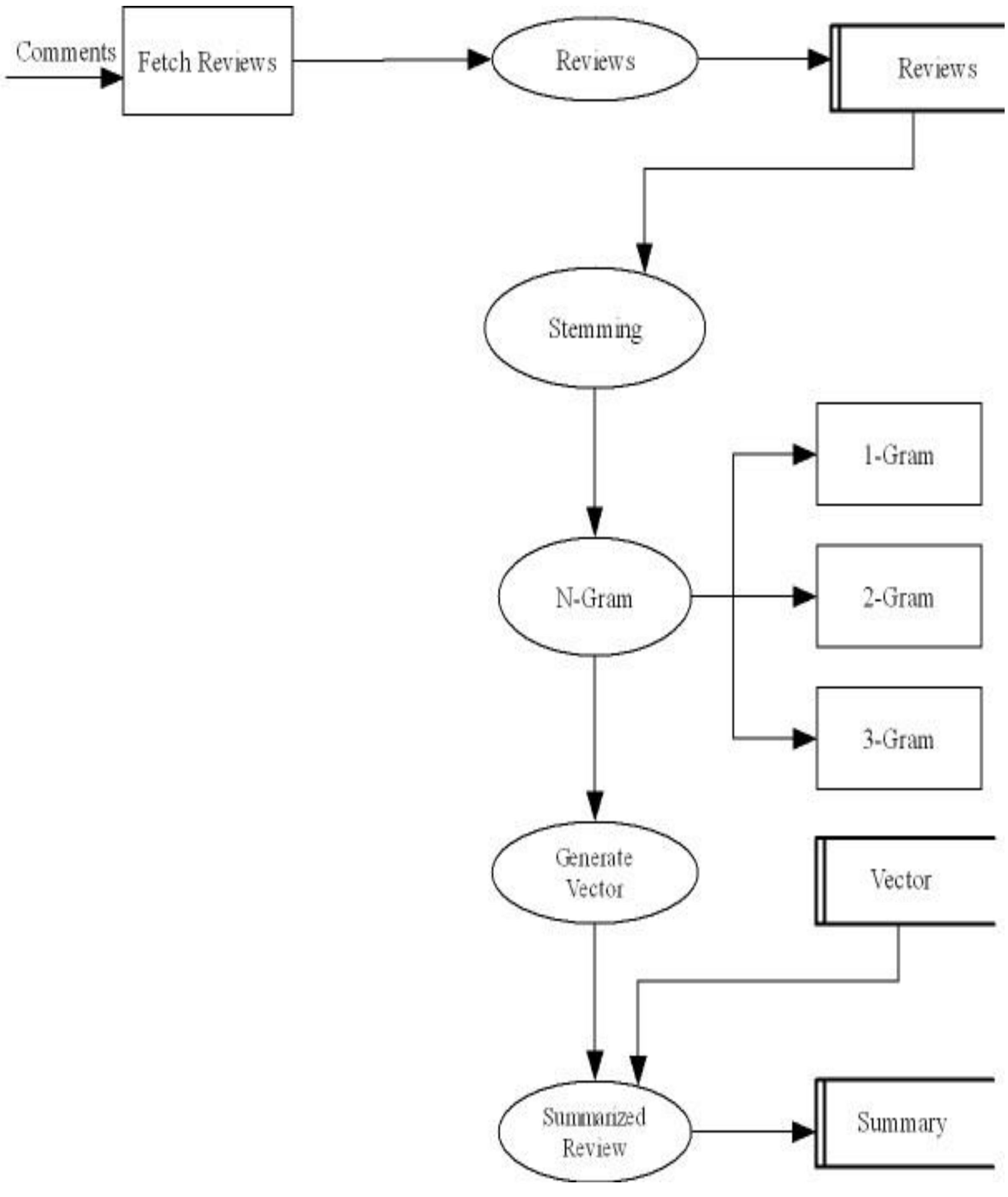

Fig 3.5 Level 1.1 Data Flow Diagram

3.2.4 ER DIAGRAM

An entity relationship model (ER model) describes inter-related things of interest in a specific domain of knowledge. The main components of ER models are entities (things) and the relationships that can exist among them. Entities may be characterized not only by relationships, but also by additional properties (attributes), which include identifiers called ”primary keys”. An entity relationship model is usually the result of systematic analysis to define and describe what is important to processes in an area of a business. An ER model is commonly formed to represent things that a business needs to remember in order to perform business processes. It does not define the business processes; it only presents a business data schema in graphical form. It is usually drawn in a graphical form as boxes (entities) that are connected by lines (relationships) which express the associations and dependencies between entities. Diagrams created to represent attributes as well as entities and relationships may be called entity-attribute-relationship diagrams, rather than entity relationship models.

The entities in this system ER diagram are review, bstsum, product, short text, summary and cluster element. These two entities are associated with each other. The next entity is cluster and its attributes are cluster id,msg id and cluster center. These two entities are associated with each other. The next entity is cluster element and the cluster and cluster element are associated with each other. The next entity is review having attributes like review_id, product_id, review. The review and cluster element are associated with each other. The entities product and review are associated with each with the relationship. The next entity is bst sum and its attributes are review id, product id and wrds. This entity is associated with review and product entities.

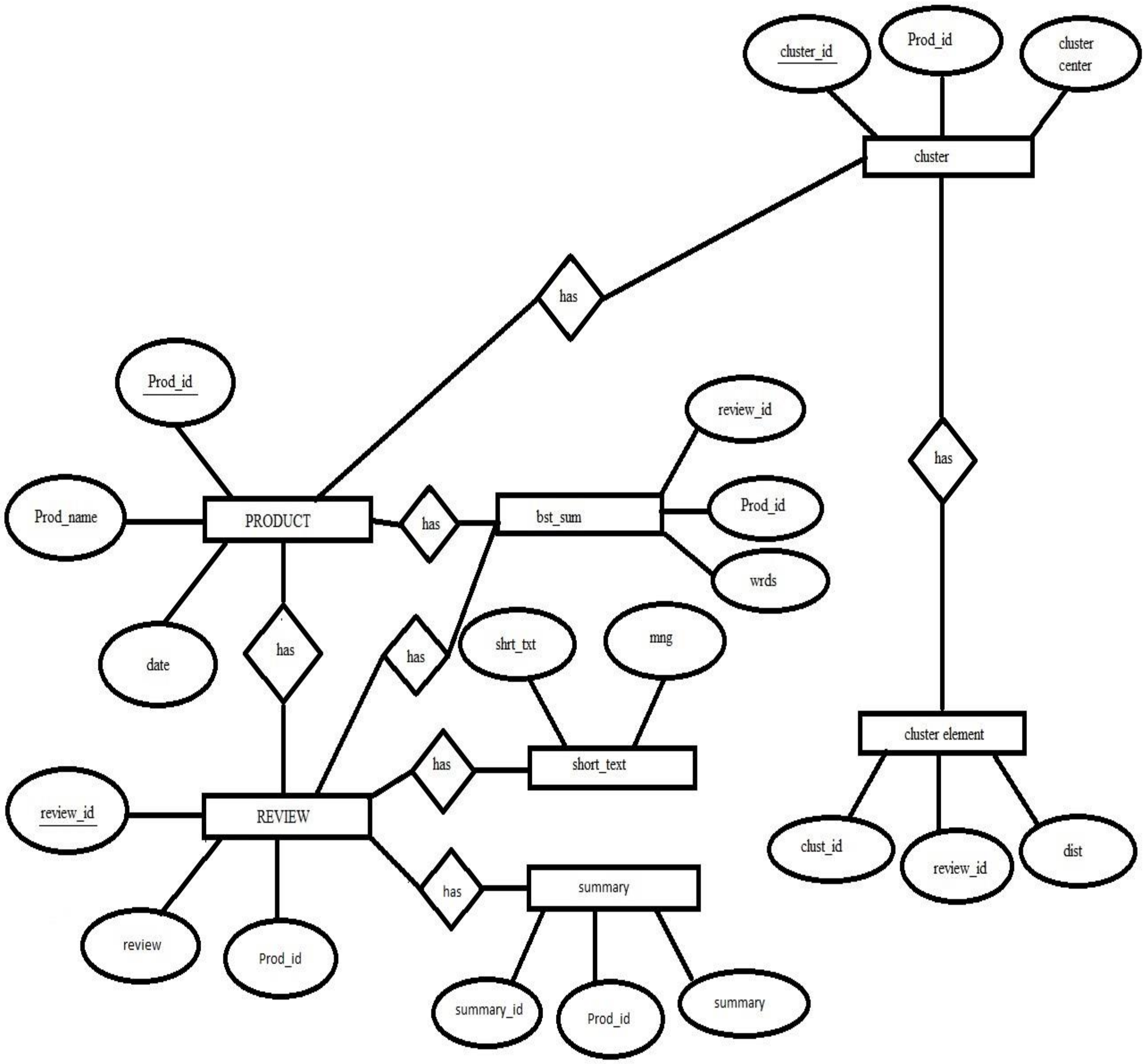

Fig 3.6 ER diagram

3.3 SYSTEM MODULES

3.3.1 MODULE SPLIT UP

3.3.1.1 NLP Module

NLP module that transforms each review into a set of n-gram terms. Initially, for each word, the process of punctuation removal will be applied to eliminate unnecessary punctuation marks connected with this word. Moreover, develop the heuristic process of redundant character removal, designed for restoring words on SNS. It can be observed that casual language style is commonly used on SNS.

3.3.1.2 Clustering Module

The similar words are clustered and generate the summarized review of the given product.

3.3.2 ALGORITHM

3.3.2.1 Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In this paper, the problem of product review summarization is modeled as an clustering task. Consider two reviews represented in the term vector model, a = (t1,a, t2,a, ... tN,a) and b =(t1,b, t2,b,....,tN,b). Each dimension corresponds to a separate term, and N is the number of dimensions. Here define that the weights of terms are equal, if the term ti occurs in the review a, ti,a will be set to 1. Otherwise, ti,a will be set to 0.

The reason for this design is that the length of each review is usually very short compared to other text documents. Then find out the content similarity score of the reviews using modified cosine similarity equation.

3.3.2.2 BatchSTS Algorithm

BatchSTS takes the whole review set as the input. These input is the radius threshold theta used for determining how similar the reviews are in a cluster. The aim of this algorithm is to find all components of the review set. The points belonging to the same connected component will be merged as a cluster. This algorithm outputs different clusters of reviews.

3.3.2.3 IncreSTS Algorithm

It is an iterative version of BatchSTS algorithm. This is aiming to provide immediate and instant summary of product reviews. The primary notion of this algorithm is to maintain the clustering result of the previous phase, and to incrementally update the clustering result with the newly-incoming review. Here first check whether the last review that is considered in the BatchSTS algorithm is equal or not to the newly incoming comment. If it is not equal then clear the previous term vectors, clusters, cluster elements. Then call the BatchSTS algorithm for clustering reviews.

3.3.2.4 Key-Term Extraction Algorithm

Finally for the design of visualization interface, representative terms will be extracted to form a key-term cloud for each group. For this here propose a key-term extraction algorithm. Which intent to provide a concise at-a-glance visualization interface that enables users to quickly get an a overview understanding of review

3.3.3 EXPERIMENTATION PLATFORM

3.3.3.1 Software Requirements

- OPERATING SYSTEM: Windows 7
- TECHNOLOGY: Java
- FRONT END: NetBeans IDE 8.0.1
- DATABASE : MySQL

3.3.3.2 Hardware Requirements

- PROCESSOR: intel CORE i3
- SPEED: 2.80GHZ
- RAM: 2GB
- HARD DISK: 20GB
- KEYBOARD: Standard Windows keyboard

3.3.4 TOOLS AND TECHNIQUES

3.3.4.1 MySQL

SQLyog is the most powerful manager, admin and GUI tool for MySQL, combining the features of MySQL Query Browser, Administrator, phpMy Admin and other MySQL Front Ends and MySQL GUI tools in a single intuitive interface. SQLyog is a fast, easy to use and compact graphical tool for managing your MySQL databases. SQLyog was developed for all who use MySQL as their preferred RDBMS. Whether you enjoy the control of handwritten SQL or prefer to work in a visual environment, SQLyog makes it easy for you to get started and provides you with tools to enhance your MySQL experience. SQLyog is a GUI tool for the RDBMS MySQL. It is developed by Webyog, Inc. based in Bangalore, India and Santa Clara California.

3.3.4.2 Adobe Dreamweaver

Adobe Dreamweaver is a software application that allows you to create and develop Websites. Dreamweaver is considered WYSIWYG (What You See Is What You Get), meaning that when you format your Web page, you see the results of the formatting instead of the mark-ups that are used for formatting. HTML is not WYSIWYG, whereas Microsoft Word is WYSIWYG. However, Dreamweaver allows you to hand code HTML as well. Dreamweaver also supports CSS and JavaScript as well as other languages including ASP and PHP. Dreamweaver makes it easy to upload your entire Website to a Webserver. You can also preview your site locally. Dreamweaver also lets you create templates for your Website that you can use again and again by modifying certain unrestricted areas within the template.

3.3.4.3 NetBeans IDE

NetBeans IDE is a free, open source, integrated development environment (IDE) that enables you to develop desktop, mobile and web applications. The IDE supports application development in various languages, including Java, HTML5, PHP and C++. The IDE provides integrated support for the complete development cycle, from project creation through debugging, profiling and deployment. The IDE runs on Windows, Linux, Mac OS X, and other UNIX-based systems.

The IDE fully supports JavaEE using the latest standards for Java, XML, Web services, and SQL and fully supports the GlassFish Server, the reference implementation of Java EE. The NetBeans IDE is primarily intended for development in Java, but also supports other languages in particular PHP,C/C++ and HTML5.

CHAPTER 4

DESIGN EXAMPLE

4.1 MODEL DESIGN AND EXAMPLE

4.1.1 SEARCH PAGE

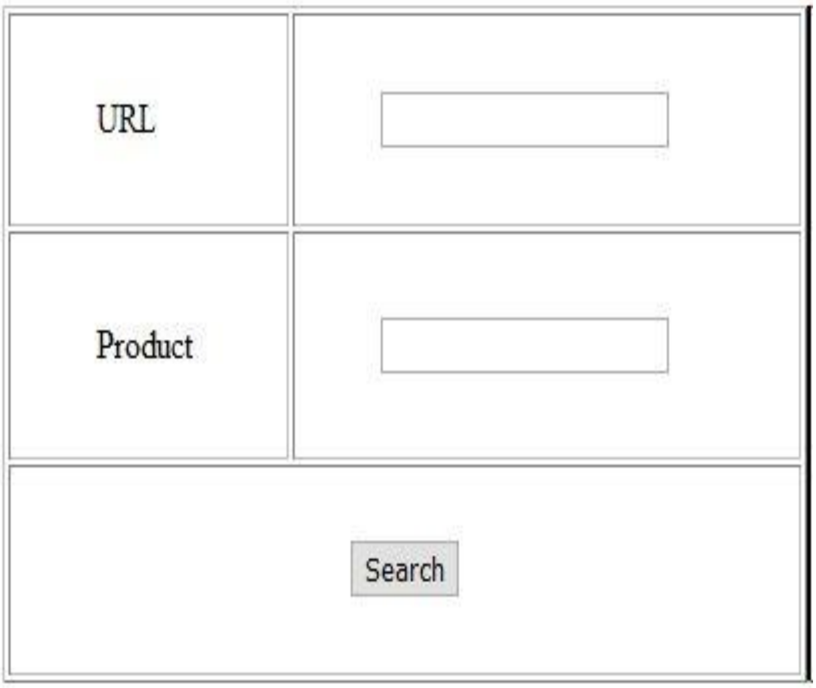

Fig 4.1 Search page

4.1.2 PRODUCT INFORMATION

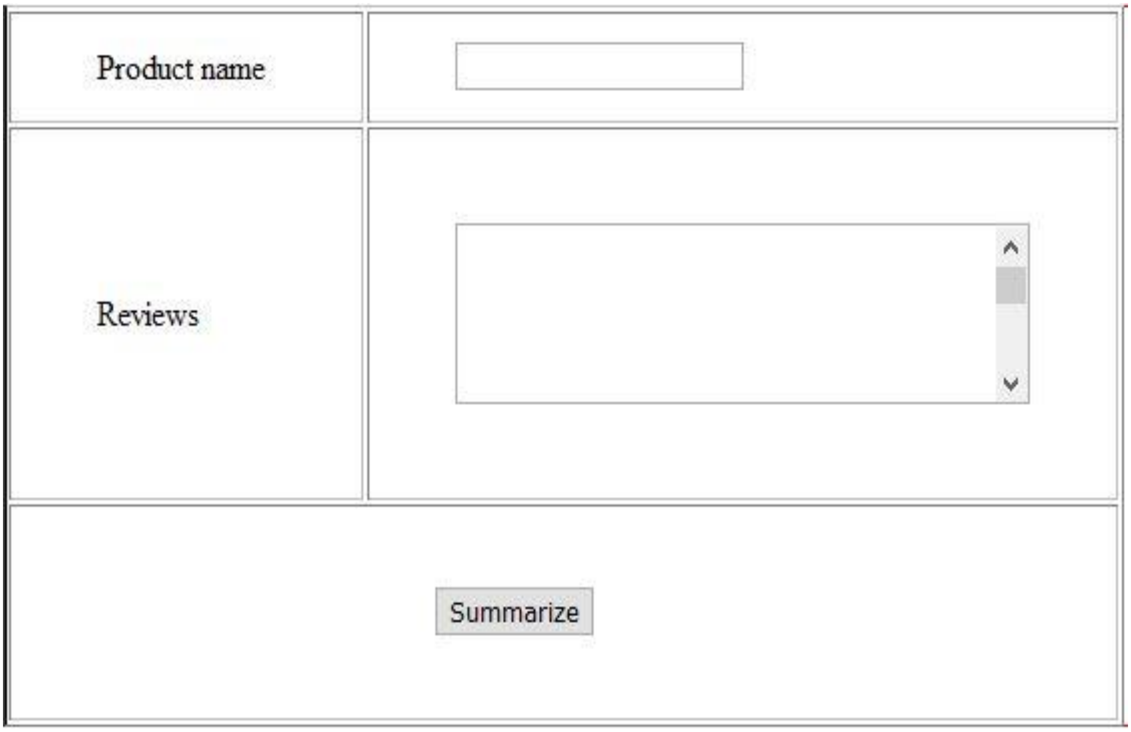

Fig 4.2 Product information

4.1.3 SUMMARIZED REVIEW

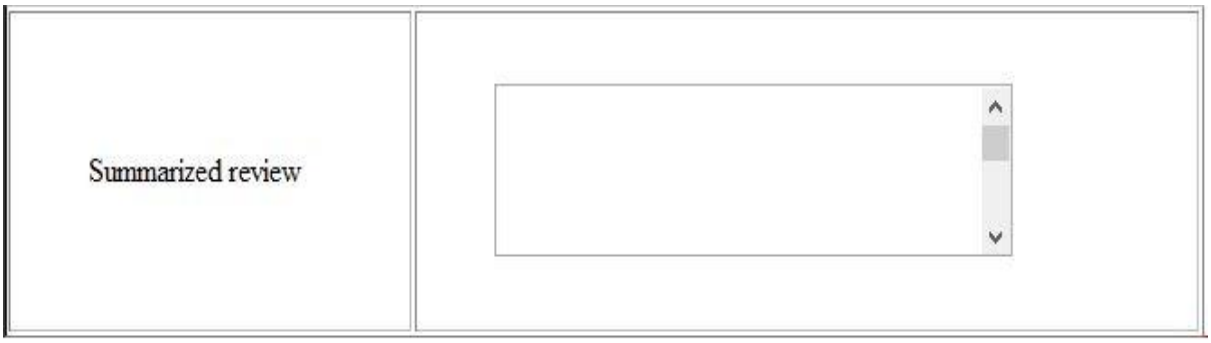

Fig 4.3 Summarized review

CHAPTER 5

CONCLUSION

In this project, to enable the capability of review summarization, we model a novel incremental clustering problem and propose the algorithm IncreSTS, which can incrementally update clustering results with latest incoming reviews in real time. With the output of IncreSTS, we design a visualization interface consisting of basic information, key-term clouds, and representative reviews. This at-a-glance presentation enables users to easily and rapidly get an overview understanding of a product reviews. In this research we incorporated a dataset from XML into a more manageable and efficient format in MySQL database. Using queues from past research, we pursued Support Vector Machine as an avenue for machine learning to automatically predict review helpfulness. As proven by our results, our hypothesis of creating specialized machine learners to increase performance on partitioned datasets in an effective way to classify product reviews. The proposed system used hybrid classifiers SVM and Naïve Bayes with fuzzy logic. Among the classification methods used, the hybrid classifier method was highly robust in nature. The accuracy of hybrid classifier method can be increased by increasing the number of classifiers. We used this technique to improve the system performance and reduce the error rating.

After studying the existing systems, we conclude that our solution provides a more realistic and efficient summarization of user opinions. Also building an application that uses this solution has brought about its use in an apt manner. This can be scaled into other domains data analytics applications that concern analysis of raw text data and summarization.

REFERENCES

[1]. Priya pawar, Nikita patil “Online product review summarization”, IEEE Conference on innvocations in information, 2017.

[2]. Akkamahadevi R Hanni “Summarization on customer reviews for a product on a website using natural language processing”, IEEE Conference on advances on computing, sept 2016.

[3]. Cheng-Ying Liu, Chi-Yao Tseng “incrests: towards real time incremental short text summarization on comment streams from social network services”, IEEE Transactions on knowledge and data engineering, 2015.

[4]. Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In Proceedings of the third ACM international conference on Web search and data mining, pages 291– 300. ACM, 2010.

[5]. Michael S Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H Chi. Eddi: interactive topic-based browsing of social status streams. In Proceedingsofthe23ndannualACMsymposiumonUser interface software and technology, pages 303–312. ACM, 2010.

[6]. Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web, pages 342–351. ACM, 2005.

[7]. Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519–528. ACM, 2003.

[8]. Minqing Huand Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACMSIGKDD international conference on Knowledge discovery and data mining, pages 168–177. ACM, 2004.

[9]. Xing Fang and Justin Zhan ,”sentiment analysis using product review data”, Journal of Big Data,(2015).

[10]. http://ijarcet.org/wp-content/uploads/IJARCETVOL-, 4ISSUE-7- 32329690- last accessed on:13/10/2016.

[11]. Giuseppe Carenini, Raymond Ng, and Adam Pauls, “Multi- Document Summarization of Evaluative Text” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 7, July 2015.

[12]. Vijay B. Raut et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, 2014.

[13]. Scorrt Bolter,” Predicating Product Review Helpfulness Using Machine Learning And Specialized classification Models”, an JoseState .

[14]. Bing Liu,” Mining and Summarizing customer Reviews“, University of llinois at Chicago .

[15]. R. Bharathi,” Online Shopping Product Aspect and Ranking Using Support Vector Machine Algorithm”, International Journal of computer Techniques, (Sep-oct 2015).

[16]. www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/Aulas/naïve_bayes_classifier.pdf

[17]. https://en.wikipedia.org/wiki/Support_vector_ machine/ last accessed on 3/2/2017.

[18].Santhi Chinthala, Ramesh Mande, Suneetha Manne and Sindhura Vemuri, Sentiment Analysis on Twitter Streaming Data, Springer International Publishing Switzerland, (2015).