IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

Improved Text Summarization Method for Summarizing Product Reviews

B. Batra[1*], S. Sethi[2], A.Dixit[3]

[1] Computer Engineering, YMCAUST, Faridabad, India
[2] IT & CA, YMCAUST, Faridabad, India
[3] Computer Engineering, YMCAUST, Faridabad, India

[*]*Corresponding Author: batrabhumika1992@gmail.com, Tel.: 9873178913*

Available online at: www.ijcseonline.org

Accepted: 03/Jun/2018, Published: 30/Jun/2018

Abstract— Text Summarization is an active and interesting research area that has emerged to in recent times .There is an increase in trend of online shopping. Before buying any product or service online users prefer to read opinions about that product of service. But problem is that there are millions of reviews for same product is available over different websites and users do not have the time to read all the reviews. So need of review summarization is there. Text summarization method summarizes the content of reviews of people with help of similarity and clustering methods and guide them whether to purchase that product or not. Text summarization can be of many types. This research work proposes an extractive improved text summarization method which performs better than existing text summarization methods. This is done by including some improvements in the text summarization method like in this research work, those sentences are retained which have at least one noun, one adjective, one verb and one adverb. This stops the elimination of some of the important reviews. Also this research work is combining STASIS similarity and LDA technique for calculating semantic and context similarity respectively. After that similarity score generated by both the techniques is combined and an overall similarity score is calculated. Sentences are assessed using this similarity score and conflicting sentences are eliminated. And most important improvement of this research work is to improve k-means clustering by including Levenshtein distance instead of Euclidean distance. After doing this improvements both the existing and improved text summarization methods are applied on datasets of reviews and their performance is compared using factors like Rand Measure, Precision, Recall, F-measure, Review Importance Factor and it is proved that proposed method is better than existing method.

Keywords— Text Summarization, Data Mining,Extractive Summarization,Text Mining

I. INTRODUCTION

Recent developments in the world of internet led to increase in online shopping which is an electronic platform to shop products and services online. For comparing the products most important criterion which is used is reading reviews or opinions of the users who have already used that particular product or service. Reviews or opinions of people about products and services are available on various e-commerce sites where that particular product is sold, social media platforms, blogs etc. But the problem is that millions of reviews are available for same product or service. People don't have time to read each and every review. So review summarization systems are need of today.

Text summarization systems have various applications ,for example can be used in search engine as summarizer to give users a summarized information of webpages or can be used as a tool to give summarized view of a document to users so that they can decide whether to read the full document or not. In addition to this text summarization systems can also be used by newsgroups to merge the most vital information available in different documents but discussing the same topic. Text summarization systems can also be used to summarize letters and other important documents used in any organization. Text summarization systems are more commonly used in areas where there is a need to transfer less information ,for instance people who use their phone to check emails like to do that but by using less data while being connected to internet.

These systems must be able to extract valuable contents from millions of reviews about a product. Opinion mining or analyses the opinions of people using a very popular method known as text summarization method. Text summarization method summarizes the content of reviews of people with help of similarity and clustering methods and guide them whether to purchase that product or not.

Rest of the paper is organized as follows, Section II contain the literature survey, Section III contain the problems identified in existing work, Section IV contain proposed methodology , Section V describes results and discussion and Section VI concludes research work with future directions.

II. LITERATURE SURVEY

The Internet is basically known as the interconnection of networks which is spread worldwide and operated by the various governments and private agencies. Initially, internet was used only by government and helped them in interconnecting laboratories used for the government research work. The exponential growth of the Internet can be seen daily, its usage has been increased tremendously over last two decades. In 1995 only 16 million users used internet which was merely 0.4% of the total population. In comparison to that, December 2017 recorded a total of 4,157 million users using internet which is around 54.4 % of the world population. The main reason for the growth in internet usage is wireless technology. With the advent of wireless networks for connection, use of internet increased to manifolds as there was no need of having wired internet connection for each computer or laptop. Multiple computers and laptops can be connected with a single Wi-Fi connection. Moreover, users can take their laptop anywhere and connect to Wi-Fi available in that area.

Revolution in mobile technology further increased internet usage manifolds. Cheap internet connections are available in mobile nowadays. People don't even need to carry their laptops for accessing the internet. They can do so on their personal mobile devices with their personal internet connection. All these developments led to increase in online shopping which is an electronic platform to shop products and services online. In addition to the cheaper internet connection, various other factors contribute to increase in online shopping like the convenience of online shopping and the wide assortment of products available over the internet.

Online shopping is convenient for people because products get delivered on their door-steps, several variants of same product or service which is available over the internet can be compared before buying that particular product or service. If product is not up to the expectations of customers, customers can even get that product exchanged without any penalty or fine suffered by customers and even if customers want to return the product, they can easily do so and get their money back. For comparing the products most important criterion which is used is reading reviews or opinions of the users who have already used that particular product or service.

Reviews or opinions of people about products and services are available on various e-commerce sites where that particular product is sold, social media platforms, blogs etc. But the problem is that millions of reviews are available for same product or service. People don't have time to read each and every review. Most of the people read few reviews which come at the top and decide whether to purchase the product or not. But sometimes even top reviews are not meaningful and fail in giving correct guidance to the people. So review summarization systems are need of today. These systems must be able to extract valuable contents from millions of review about a product.

Data mining can be used to achieve this task. Broadly it is classified into two types structured (for relational tables) and unstructured. Since reviews are in form of unstructured text, unstructured data mining i.e. text mining is used. Although text mining is able to extract useful information from the unstructured text but it is not able to extract opinions of people about a particular product or service. So text mining is extended [9] and a new technology known as opinion mining or sentimental analysis came into existence to analyse the opinions of people. In opinion mining this task is done by a very popular method known as text summarization method which summarizes the content of reviews of people with help of similarity and clustering methods and guide them whether to purchase that product or not.

Data mining is defined as the procedure of finding patterns in huge data sets including methods of machine learning, statistics, and database systems. Data mining is a field of computer science where techniques are applied to extract information and patterns from a dataset and convert it into a form which can be easily understood by users. It is a new technology which has high scope and potential in helping organizations in extracting important information from their databases and data warehouses. Only important information which is required for use in applications of the organizations is extracted so that their application become more efficient and time of engineers is saved which is otherwise wasted to handle data which includes unimportant information as well. This is required because nowadays a huge amount of data is readily available on internet which cannot be easily handled [10]. Moreover after applying data mining techniques to existing data more refined data can be found. Tools which are used in data mining can easily answer business questions of the organizations which were too time-consuming before data mining came into existence. Hence organizations prefer to use data mining techniques while designing business models.

Structured data mining is defined as the technique of finding useful information from structured and semi-structured data sets. Traditionally data mining was only related to structured data in form of tables, but all data that is available on the internet cannot be converted to relational databases which resulted in the growth of semi-structured data.

Unstructured data mining or text mining is defined as the process of giving the structure as per your requirement to unstructured data mining. It is the process of deriving important and useful information from text [1]. This requires both linguistic and statistical techniques to evaluate unstructured data. Unstructured data files consist of text and multimedia content. For example e-mails, word processing documents, videos files, presentations, audio

files, web pages ,photos and many types of business documents. But this technique is not efficient because due to advances in internet technology people can exchange their information instantly on various social media platforms but it merely extracts useful information from the text but does not tell about the opinions of people. For that, an advanced technology was required for which text mining was extended. This led to the development of technology which is called as opinion mining. Opinion mining is defined as a kind of natural language processing for keeping track of the mood of the general people regarding particular product or service. Opinion mining is also known as sentimental analysis and used to identify and categorize opinions about text and then uses text summarization method to find summarized opinions of text.

Text summarization [2] is a technique used in opinion mining to summarize the opinions or reviews about a particular product or service. Summarization is needed because there can be millions of opinions or reviews of a single product by a number of people on various social media platforms. There is no time available to people that they read all the reviews before purchasing the product or service, so they need summarized reviews or opinions. Summarized reviews can assist customers in providing most important information that can help them in purchasing the product or service. Text summarization can be of various types, some of which are discussed in this section.

Extractive summarization works by choosing a subset of already existing and important words, phrases or sentences from the original document in order to form summary. This method is usually easy to implement because it is not based on the semantic relation between sentences and are more successful. Extractive summaries generally consider most important information is used as first sentence of the summary [3]. Also the summaries generated using these techniques are generally longer than average. The summaries generated from this technique suffer from the drawbacks of inconsistency, lack of balance and lack of cohesion. In this study, text summarization will be done using this technique.

Abstractive summarization works by generating a sentence from data having semantic representation and after that generates a summary by using natural language processing (NLP) methods. It includes interpretation of original information in shorter version. Abstractive summaries generated using this technique may include words which are not actually there before summarizing the original . They are difficult to generate as they require deep knowledge of NLP tasks but are more concise and accurate than extractive summaries. These summaries are required in cases where opinions of people are very diverse. Summaries generated using this technique have less compression ratio and less redundant data but cost of generating these summaries is usually high.

Informative summarization is used as an alternative to the original text. It generate summaries which give the user, the brief information regarding the original text and motivates Length of these summaries is nearly 20 to 30 percent of the original text. But in some cases these summaries even leads to miscommunication that the text is not worth reading because they do not have detailed information about the text. These summaries are easier to produce.

Indicative summarization is applied for quickly viewing a long text. It is used in cases when the user needs to know what the main idea of the original text is. Summaries generated using this technique generally small 5 to 10 percent of the original text and it helps users in deciding whether they want to read the full document or not. It does not contain the actual data but contains the metadata which can include scope of the document, methodology used in the document, purpose of the document etc. For example, before purchasing a book or novel, a buyer in most cases first reads the summary given in the front and back side of the novel and then continues with content later.

Generic summarization is kind of summarization in which summaries are generated for any kind of user and in addition to that these summaries also do not rely on the theme of text. Summaries are generated from author's perspective and are not user-specific. As summaries can be used by any kind of user all of the information is given the same kind of importance. No prior knowledge of text is available while generating the summaries. Also there is no deed to keep track of different interests of different users while generating these summaries.

Query-based summarization is a kind of question answer summarization in the sense that user have general information about a particular interest and ask for special information about it. In this technique summaries generated are result of user queries. These summaries provide user specific view and rely on the type of query given by the user which gives idea about users' interests. These summaries can be used only by users who have interests related to these summaries and therefore they are difficult to generate as there is need to track different interests of different users.

Genre specific summarization is used by systems which accept only special kind of input. For example input can be in form of manuals, newspaper articles, stories etc. It solves the problem of summarizing heterogeneous documents. But only few real life systems use this technique.

Domain independent summarization is a kind of summarization technique uses systems that can accept different kinds of text. This technique generates summaries which do not depend on domain. These summaries can be used by any kind of user and are not dependent on any type of input received. But these summaries are difficult to generate because different criteria is required for pre-processing different kinds of input received each time. Most of the real life systems are domain independent.

Single-document summarization is a kind of summarization accepts only one document at a time as input. These summaries are generally easier to generate as only a single document is required to summarize and they have less overhead. Almost all systems using single document summarization technique generate summaries using monolithic structure of the document. For example, for writing single document summaries, take first sentence of each paragraph and arrange them together in the original sequence. Main drawback of this technique is that summaries of related topics cannot be generated using it.

Multi-document summarization is a kind of summarization accepts several documents at a time as input. These summaries are generally difficult to generate as only a multiple documents are required to summarize. Systems using multi document summarization technique for generating summaries usually do not depend upon the structure of the document because structure of different documents used for summarization and are not readily available as in case of a single document. A multi-document summarization system is efficient if it organize the information around the most important aspects so that variant views can be represented easily and as a result users get a good overview of a particular topic whose documents are summarized.

Mono-lingual summarization is a kind of summarization technique uses systems which accepts only those documents which have a specific language and generates output on the basis of that language only. Summaries generated using this technique have less overhead and are easier to implement as only one language needs to be processed. But these are used in very less areas as nowadays most of the companies and organizations are multinational and require handling of different languages. In most cases these summaries require translation in the language required by the user which itself is a tiresome process which make systems using this technique less efficient.

Multi-lingual summarization *is a* kind of summarization technique [12] is used by systems which accept documents having different languages and generates output on the basis of different languages accepted. In today's world most of the text summarization systems use these techniques because multinational companies work across different countries using different languages. Also these systems are more efficient and have less overhead as they can be used by different users in different language without the need of translation of the document in their own language. Summaries generated using this technique are difficult to produce as documents in different languages need to be handled in the same system.

Efficient text summarization methods use clustering to group and integrate similar sentences together. Clustering is also used to remove noisy sentences because noisy data can influence text summarization method in a bad way. Hence clustering is to make text summarization methods more efficient.

Clustering [5] is defined as a technique in which groups of similar objects are created by dividing the data. Each group which is known as the cluster comprises of objects which are similar among themselves but they are not similar to objects contained in other groups. When data is represented in clusters although some data is lost simplification is achieved. Clustering can be done using various techniques some of which are discussed in this section.

Single linkage clustering [6] is also known as connectedness or minimum clustering method. In this technique distance between one cluster and another cluster is considered same as shortest distance which can be calculated from any item of any cluster to any item of other cluster.

Complete linkage clustering is also known as diameter or maximum clustering method. In this technique distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster.

Average linkage clustering is a technique in which distance between one cluster and another cluster is considered same as greatest distance which can be calculated from any item of any cluster to any item of other cluster.

K-means clustering is a popular method used for clustering in data mining [7]. It groups given set of data objects into clusters on the basis of their proximity to each other using square error function. For which first mean value or centre is found by randomly selecting a data object. After that most similar objects are assigned in same clusters on the basis of mean value. This is done by finding data objects pairs which have minimum squared error value than all the other data objects pair. This algorithm finds k partitions i.e. clusters which minimizes square error function.

K-medoids clustering [8] is used to reduce sensitivity of k-means clustering to noisy data and outliers. It uses absolute error function to group two sets of data objects. Actual objects are used instead of mean values which is known as representative of data object i.e. medoid which at initial stages is selected randomly. After that each representative data object is replaced by non-representative data object. This process continues till quality of clustering stops improving. K-medoids clustering technique is robust because it very sensitive to noise and outlier data objects and do not get much affected by their presence.

III. PROBLEM IDENTIFICATION

There are many problems associated with existing text summarization approaches used for summarizing the reviews and showing the users top reviews. In many cases, some of the most important reviews are not shown to the users due to which in some cases users are misguided. Some of these problems are-

- In preprocessing steps earlier researches focused on sentences having at least one noun and one adjective. They do not retain sentences having noun and verb or noun and adverb combination which leads to the elimination of some of the important reviews.
- Similarity calculation by earlier text summarization methods was not efficient as most of the text summarization methods avoid context similarity between sentences and a few systems which considered context similarity use a bogus technique which was efficient only for few sentences.
- Earlier techniques used k-means, k-medoids and similar techniques for selecting top k sentences. Each of them has their disadvantages and do not give accurate results in some cases.

IV. PROPOSED METHODOLOGY

The proposed text summarization method for summarizing product reviews consists of five components (as shown in figure 1)-

- Review Extractor
- Review Pre-processor
- Sentence Importance Calculator
- Review Similarity Calculator
- Summarized Review Generator

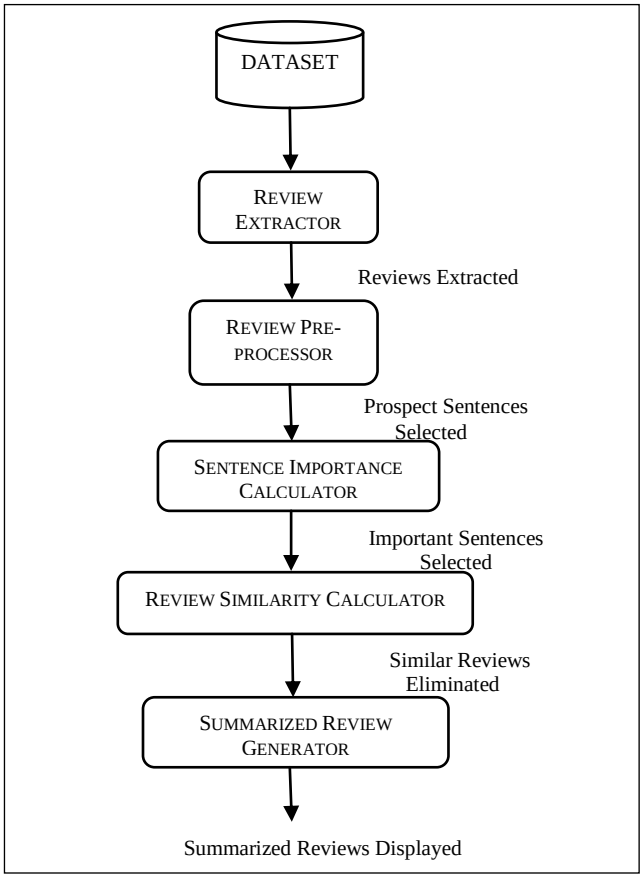

Figure 1. Architecture of Proposed Text Summarization System

The work in this research collects dataset from the website – http://jmcauley.ucsd.edu/data/amazon/

Data sets Dj= {d1, d2, …dj} as input

- d1- Hyundai Accent GLS Sedan 2012
- d2-The Bing Bang Theory(Amazon Instant Video)
- d3-Epiphone Casino(Guitar)

And Data sets Dn= {d1, d2, …dn} as output

These reviews are in json format as-

{"reviewerID": "A3BC8O2KCL29V2", "asin": "B000H00VBQ", "reviewerName": "Carol T", "helpful": [0, 0], "reviewText": "I highly recommend this series. It is a must for anyone who is yearning to watch \"grown up\" television. Complex characters and plots to keep one totally involved. Thank you Amazon Prime.", "overall": 5.0, "summary": "Excellent Grown Up TV", "unixReviewTime": 1346630400, "reviewTime": "09 3, 2012"}

A) Review Extractor

Reviews are collected and extracted in this module. Most of the reviews available on the internet are not in the format that can be used for pre-processing by a language. Datasets of three products collected which are in json format. Main review text is needed for further evaluation. As this study uses java implementation this is done by converting json objects to map objects using hash map. This module follows following algorithm (as shown in figure 2)–

Review Extractor ( )

Input: Review datasets in json format Dj
Output: Extracted review dataset in map format dm

Method-
Step 1.Repeat for each dataset dj €Dj
1.1. Input the dataset dj

Step 2.For each review rj€dj={r1,r2…..rj} reviews
2.1. Convert rj to map object rm
2.2 Store rm to dm

Step 3.Call review pre-processor (dm)

Figure 2. Algorithm of Review Extractor

Now the extracted reviews are passed on to review pre-processor module for pre-processing of reviews.

B) Review Pre-processor

Sentences are considered as the basic unit of analysis in this module. So when it receives reviews in the form of paragraphs as input from review extractor it first splits the paragraphs into sentences. After that it eliminates stopwords like ‘is’,’am’,’the’ etc. because there is no need of these words in summarized reviews as they do not hold a significant purpose in a sentence, they are only added to

make a sentence grammatically correct. This study uses a list of 119 stopwords in implementation.

Moreover this module is also responsible for removing the first identified problem of this study which is that in most of the text summarization methods only those sentences which have atleast one nouns and one adjectives are retained while eliminating the other sentences. This leads to elimination of some of the important reviews. But this study retains those sentences which have atleast one noun, one verb, one adjective and one adverb are retained.

This is done by giving POS (part-of speech) tag to each word of the sentence by using StandFord POS tagger which is a very popular tagger used for tagging texts in java.

The tags are assigned to each word of all the sentences.After this only those sentences which contain one noun, one verb, one adverb and one adjective combination are retained and tags are removed from all the words of all the reviews.

This module follows following algorithm (as shown in figure 3)-

Review Pre-processor ( )

Input: Review set in map format dm
Output: Pre-processed review dataset dp

Method-
Step1-Each review rm €dm ,rm is converted to rp
1.1.rm is split into sentences.

1.2-Stop words like 'a','an','the' etc. are removed from sentences because there is no relevance of these words in most of the cases. This study uses a list of 119 stopwords.

1.3-POS (part-of-speech) tagging is done and each word of the sentence is tagged.

1.4-Filtering of sentences is done. Only those sentences are retained which have atleast one noun, one adjective, one verb and one adverb and called as prospected sentences.

1.5–Tags are removed from the prospected sentences.

Step 2.Store rp in dp.
Step 3.Call sentence importance calculator (dp).

Figure 3. Algorithm of Review Pre-processor

Pre-processed reviews are then passed to sentence importance calculator module.

C) Sentence Importance Calculator

This module is responsible for calculating the importance of the sentences. Several factors influence sentence importance so this module calculates importance of prospected sentences. Various factors are studied and then decision is taken which factors influence the sentence importance the most. In this study following two factors are considered for calculating sentence importance.

- Sentence weight (S_w)
- Location of sentence in review paragraph (S_{loc}).

Sentence weight is calculated by summing the term frequency of each word and then dividing it by the length of the sentence. This factor is important because it include those terms which occur more frequently in the collection of reviews. And the thing which is said by most of the people is more important and should not be eliminated.

Location of the sentence is also important because most of the people give their overall reaction to a product or service in the first or last sentence of the review. In middle of the paragraph details of their opinions are included which matters but not as much as overall reaction. So first and last sentences are assigned location score as 1. And other sentences are assigned location score in the range (0, 1)

After that overall sentence importance is calculated by multiplying sentence weight and location score factors. Experiments are performed to calculate the threshold value of sentence importance score. In this study threshold value for sentence importance score is considered as 70. Sentences having importance score less than 70 are eliminated.

Sentence Importance Calculator ( )

Input: Pre-processed review dataset dp
Output: Important review dataset di

Method-
Step1-Each review rp €dp, rp is converted to ri
1.1.rm is split into sentences.

1.2- For calculating sentence weight first frequency of each term is calculated, summed and divided by length of the sentence. This is done using following equation-

$St_{freq} = \sum_{i=0}^{n} tfreqi$ (1)

$S_{w} = St_{freq}/n$ (2)

where n is number of terms in a particular sentence and $tfreqi$ is frequency of ith term in the sentence and St_{freq} is sum of term frequencies.

1.3- Sentence location factor is considered as 1 for first and last sentence of the review. Rest of the sentences are scored in the interval (0, 1) and division of this interval is done and sentence is assigned a sentence location factor value. This is done using equations-

$S_{loc} = 1$

interval=1/(n-1) (3)

For 2nd sentence to 2nd last sentence

S_{loc} =-interval (4)

$S_{loc} = 1$

1.4-Overall sentence importance is calculated by multiplying S_w and S_{loc} as-

$$S_{imp} = S_w * S_{loc} \quad (5)$$

1.5- Threshold value of S_{imp} is calculated on the basis of experiments. In this study threshold value is 70. Sentences having S_{imp} below 70 are eliminated.

1.6- Sentences of each review are combined to form a single sentence ri are sent to review similarity calculator.

Step 2.Store ri in di.
Step 3.Call review similarity calculator (dp).

Figure 4. Algorithm of Sentence Importance Calculator
These reviews are then passed as input to review similarity calculator module.

D) Review Similarity Calculator

This module is responsible for calculating similarity between two reviews and eliminating one of the two reviews which are similar to each other. This is done by calculating semantic similarity core using STASIS similarity technique and context similarity score using LDA technique. After that both the similarity scores are combined and sentence similarity score is calculated. One of the sentences in sentence pair having sentence similarity score as 0.9 or 1 are eliminated as they are nearly similar to each other.

This module follows following algorithm (as shown in figure 5)-

Review Similarity Calculator ( )

Input: Review dataset di
Output: Review dataset ds

Method-
Step1-Each review ri €di, ri is converted to rs

1.1-Calculation of semantic similarity between words is done on the basis of path length.

1.2-Semantic similarity between two reviews is calculated by calculating cosine coefficient of two vectors which are representations of two sentences whose similarity need to be calculated.

1.3-Overall similarity between two reviews is calculating by combining similarity score obtained from step 1 and step 2.

1.4. One of the reviews from the review pair which have similarity store is eliminated and

Step 2. Retained review rs is stored in ds.
Step 3.Call summarized reviews generator (ds).

Figure 5. Algorithm of Review Similarity Calculator
These reviews are then passed to summarized review generator module.

E) Summarized Review Generator

This module is responsible for selecting top reviews which are important from all the other reviews and are displayed to the users as an output of text summarization system. This technique is known as clustering. This study tries to improve already existing k-means clustering technique because for applying k-means is generally on strings or sentences vectorization of the sentences is needed because it involves calculating Euclidean distance which cannot be applied to string data. Hence this study uses Levenshtein distance instead of Euclidean distance to calculate distance between sentences as it can be directly applied on string data without converting it into vectors. This leads to an improved k-means clustering algorithm for clustering string data. Its performance is compared to original k-means on various factors in next section.

This module works on following algorithm (as shown in figure 6)-

Summarized Review Generator ( )

Input: Review dataset ds
Output: Review dataset dn
Method-
Step1-Each review rs €ds, rs is converted to rn

1.1. Initial set of mean values is initialized.

1.2. Each observation is assigned to a cluster which has least Levenshtein distance.

1.3. Mean values are updated.

1.4. Steps 3& 4 are repeated until there is no change in any cluster.

1.5. Top reviews from each of the k clusters are selected and is known as rn

Step 2. Review rn is stored in dn.
Step 3.Each of the review from dataset dn is displayed.

Figure 6. Algorithm of Summarized Review Generator
Similarly dn for each dj is generated and Dn consisting of all datasets dn is generated.

After that performance analysis of this proposed text summarization method for summarizing product reviews is done in comparison to existing text summarization method in the next section.

V. RESULTS AND DISCUSSION

In this section performance comparison of proposed text summarization and existing text summarization system is done using following factors-

- Rand Measure
- Precision
- Recall
- F-measure
- Review Importance Factor

All the factors mentioned above are analysed for both the proposed and existing techniques using all the three products but with 1000 reviews and average percentage difference is calculated for all the factors which tells which technique is better for summarization of text reviews.
Definition of these factors and criteria to find values of those factors is discussed in following section.

1) Performance Analysis using Rand Measure(R)-It is defined as percentage of right decisions made by a clustering algorithm while making clusters. It can also be known as accuracy of a clustering algorithm.

It is calculated using equation-

$$R = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

Where

- TP is defined as number of true positives
- TN is defined as number of true negatives
- FN is defined as number of false negatives
- FP is defined as number of false positives

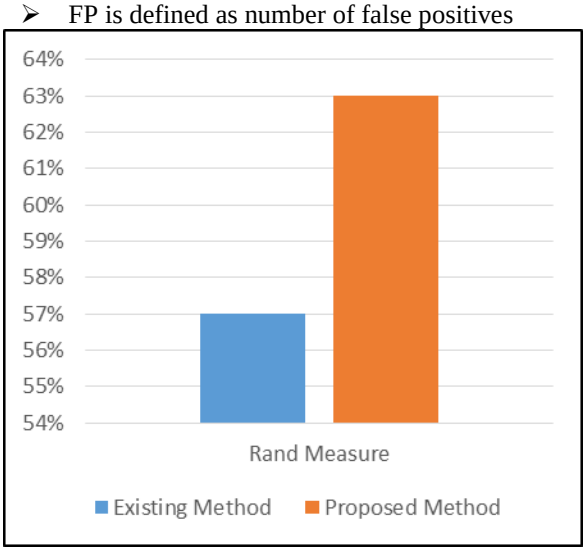

Figure 7. Comparison of Rand Measure Values

Rand Measure for existing technique is 57% and for proposed technique is 63%. Increase of 6% in rand measure i.e. accuracy is there as shown in figure 7.

2) Performance Analysis using Precision (P)-Precision in the terms of clustering is defined as fraction of pairs which are correctly put up in the same clusters by a clustering algorithm.

It is calculated using equation-

$$P = \frac{TP}{TP+FP} \quad (7)$$

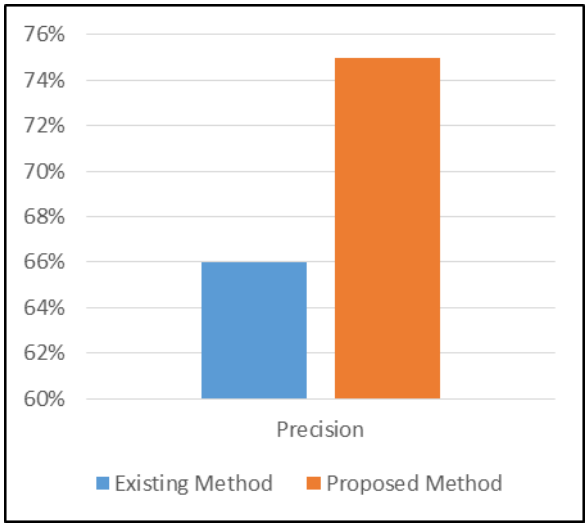

Figure 8. Comparison of Precision Values

Precision for existing technique is 66% and for proposed technique is 75%. Increase of 9% is there as shown in figure 8.

3) Performance Analysis using Recall (Re)-Precision in the terms of clustering is defined as fraction of correct pairs which are identified by a clustering algorithm.

It is calculated using equation-

$$Re = \frac{TP}{TP+FP} \quad (8)$$

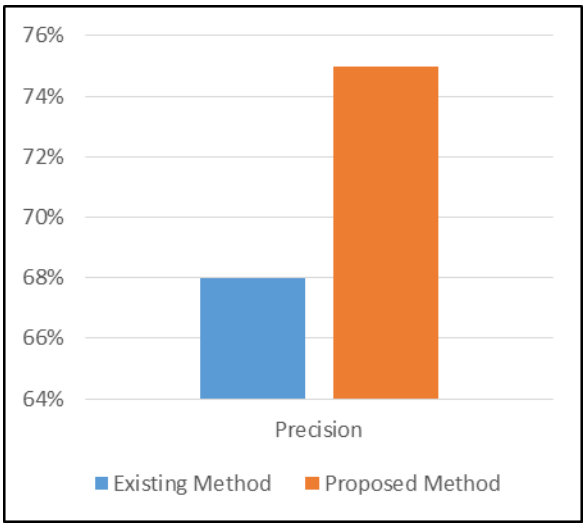

Figure 8. Comparison of Recall Values

Recall for existing technique is 68% and for proposed technique is 75%. Increase of 7% is there as shown in figure 9.

4) Performance Analysis using F-measure (F)-F-measure is calculated for balancing the effect of false negatives in a clustering algorithm.

It is calculated using equation-

$$F = \frac{2*P*Re}{P+Re} \quad (9)$$

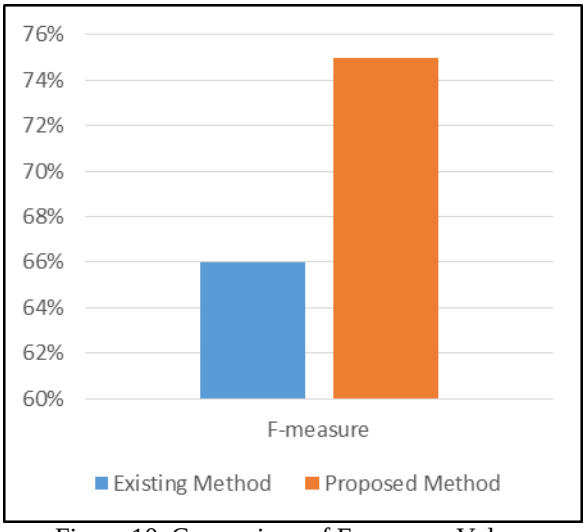

Figure 10. Comparison of F-measure Values

Recall for existing technique is 66% and for proposed technique is 75%. Increase of 9% is there as shown in figure 10.

5) Performance Analysis using Review Importance Factor (Rif)-Review important factor is defined as the ration of number of unimportant reviews eliminated by the text summarization method to total number of reviews excepted by the reviews.

It is calculated using equation-

$$Rif = \frac{U}{T} \quad (5)$$

where

- T is defined as total number of reviews
- U number of unimportant reviews

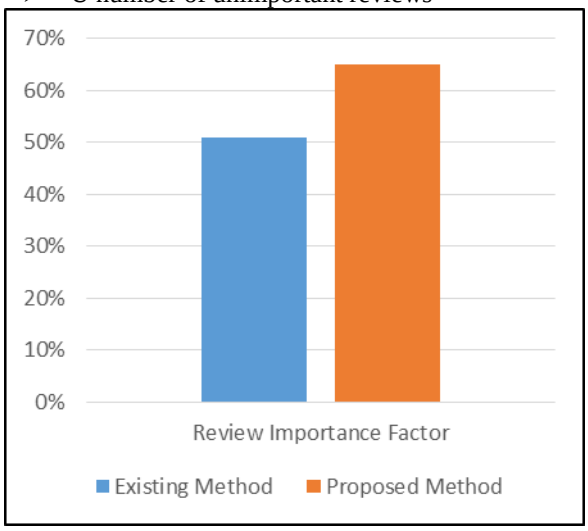

Figure 11. Comparison of Review Importance Factor Values

Recall for existing technique is 51% and for proposed technique is 65%. Increase of 14% is there as shown in figure 11.

V. CONCLUSION AND FUTURE SCOPE

Most of the text summarization methods developed for summarizing product reviews are not as efficient as they should be because now a days users want results as soon as possible and with high level of accuracy which is not there in existing text summarization systems. Hence this research work proposes an improved text summarization method for mining product reviews which is better than all the existing methods.

Supremacy of the proposed technique can be observed when the existing and proposed methods were analysed for same datasets and it solved the problems identified as-

1) Importance Reviews Elimination: Earlier methods did not consider verbs and adverbs while pre-processing the reviews which led to elimination of some of the importance

reviews and less review importance score of 51%. The proposed method solves this problem using module 2 and review importance score is increased to 65%. Hence there is an increase of 14% in review importance score by the proposed method in comparison to existing method.

2) Similarity Calculation Inefficient: Earlier methods did not consider context similarity while calculating similarity between sentences and as a result there is duplicity of similar reviews and system waste time in processing of those reviews. But proposed method solves this problem using module 4 and duplicate or almost similar reviews are not processed again and again by the system.

3) k-means Clustering was not Efficient for Clustering Text Data: Original k-means is not efficient for clustering text data because data first needs to be converted into vectors which is a tiresome process with Rand Measure 57%, Precision 66%, Recall 68%, F-measure 66%. Proposed method solves this problem using module 5 and applies Levenshtein distance instead of Euclidean distance in k-means clustering and increases the efficiency of clustering sentences with Rand Measure 63%, Precision 75%, Recall 75%, F-measure 75%. Hence there is an increase of 6% in rand measure value, increase of 9% in precision value, increase of 7% in recall value, increase of 9% in f-measure value in comparison to the existing method.

In future more improvements needs to be done as trend of online shopping and speed of internet is increasing at a rapid rate hence need of more efficient text summarization method will be there in near future. For that efficiency of existing clustering and similarity technique needs to be improved. Evaluation of emoticons can also be added in this system as most of the people nowadays while expressing their opinions use emoticons along with the text. Personal recommendation systems for different users can also be added so that user can get summarized reviews on the basis of factors which they want.

REFERENCES

[1] Abhishek Kauirik, Sudhanshu Naithani, "*A comprehensive study of text mining approach*", IJCSNS International Journal of Computer Science and Network Security, Vol. 16,Isuue 2,Feb 2016

[2] Alaa F. Alsaqer, Sreela Sasi. "*Movie review summarization and sentiment analysis using rapidminer*", 2017 International Conference on Networks & Advances in Computational Technologies, ISSN: 978-1-5090-6590-5, July 2017

[3] D. Gaikwad and C. Mahender, "*A review paper on text summarization*", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5,Issue 3, March 2016

[4] A. Shetty, R. Bajaj, "*Auto text summarization with categorization and sentiment analysis*", International Journal of Computer Applications,Vol.130, Issue 7 ,2015

[5]Xiaoye Wanga, Xiaorui Chaia, Ching-Hsien Hsua, Yingyuan Xiaoa, Yukun Lia,"*cluster analysis based on opinion mining*", 8th International Conference on Ubi-Media Computing,ISSN:978-1-4673-8270-0,2015

[6]https://onlinecourses.science.psu.edu/stat555/node/86

[7] A.Dharmarajan, T.Velmurugan, "*Applications of partition based clustering algorithms: a survey*" , IEEE International Conference on Computational Intelligence and Computing Research, ISSN:978-1-4799-1597-2,2013

[8] Preeti Arora, Dr. Deepali, Shipra Varshney,"*Analysis of k-means and k-medoids algorithm for big data*", International Conference on Information Security & Privacy, Dec 2015

[9]S. Sethi, A. Dixit," *An adaptive web search system based on web usages mining*" International journal of computer engineering and application , Vol. X , Issue. 1, ISSN: 23213469 ,2016

[10] S. Sethi, A. Dixit, "*An Automatic User Interest Mining Technique for Retrieving Quality Data*" International Journal of Business Analytics. Vol. 4, Issue. 2, pp 62-79, ISSN: 2334-4547, 2017

Authors Profile

Bhoomika Batra is currently pursuing M. Tech. in Computer Engineering YMCA University of Science & Technology, Faridabad Haryana. She is currently a research scholar and doing research in the field of data mining. She has received her B. Tech degree from MD University ,Rohtak, in the year 2014.

Shilpa Sethi has received her Master in Computer Application from Kurukshetra University, Kurukshetra in the year 2005 and M. Tech. in Computer Engineering from MD University Rohtak in the year 2009. She is currently pursuing PhD in Computer Engineering and serving as Assistant Professor in the department of computer engineering at YMCA University of Science & Technology, Faridabad Haryana. She has published more than ten research papers in various International journals and conferences. Her area of research includes Internet Technologies, Web Mining and Information Retrieval System.

Ashutosh Dixit received his PhD and M. Tech. in Computer Engineering from MD University Rohtak, in the years 2010 and 2004 respectively. He is presently serving as Associate Professor in the department of computer engineering at YMCA University of Science & Technology, Faridabad Haryana. He has published around 80 research papers in various International journals and conferences. His research interests include Internet Technologies, Data Structures and Mobile and Wireless networks.

