EVALUATE AND PROPOSED TECHNIQUE TO GENERATE SUMMARY USING ONTOLOGY TECHNIQUE

Dissertation submitted in partial fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

REETA RANI

11604633

Supervisor

SAWAL TANDON

School of Computer Science and Engineering (14 Bold)

Lovely Professional University

Phagwara, Punjab (India)

November 2017

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0059

Supervisor Name : Sawal Tandon **UID :** 14770 **Designation :** Assistant Professor
Qualification : _____ **Research Experience :**
_____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Reeta Rani	11604633	2016	K1637	9417250625

SPECIALIZATION AREA : Programming-I **Supervisor Signature:** _____

PROPOSED TOPIC : Sentiment analysis and keyword extraction algorithm using Text Summarization.

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.67
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.67
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.67
5	Social Applicability: Project work intends to solve a practical problem.	6.67
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.67

PAC Co ıмittee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): Yes
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): NA
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: Sentiment analysis and keyword extraction algorithm using Text Summarization.

Overall Remarks: Approved **PAC CHAIRPERSON Name:** 11024::Amandeep Nagpal
Approval Date: 04 Nov 2017 11/28/2017 1:11:49 PM

ABSTRACT

The text data analysis is a very important topic of research in which text summarization and sentiment analysis are the key points. The text summarization means to understand the large data in short description. The sentiment analysis is the technique to analyse opinions towards any area. The ontology technique is applied in the base paper for the text summarization. In this research work, the weight-based algorithm will be applied which will generate the summary but with more accuracy as compared to existing ontology technique. The classification technique will be applied for the sentiment analysis which will reduce the execution time.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "EVALUATE AND PROPOSED TECHNIQUE TO GENERATE SUMMARY USING ONTOLOGY TECHNIQUES" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr./Mrs. Research Guide's Name. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

_____

Reeta Rani

11604633

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation II proposal entitled **"EVALUATE AND PROPOSED TECHNIQUE TO GENERATE SUMMARY USING ONTOLOGY TECHNIQUE",** submitted by **Reeta Rani** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

_____

(Sawal Tandon)

Date:

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of dissertation. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during this thesis work. I am sincerely grateful to them for their truthful and illuminating views on many issues related to this research.

I express my sincere thanks to my guide **Sawal Tandon** for his invaluable assistance, motivation, guidance and encouragement without which this research work will be dream. In spite of his busy schedule, he was always there to iron out difficulties which kept o aspiring at regular intervals.

I am really thankful to our **Lovely Professional University** for providing me with an opportunity to undertake this research topic in this university and providing us with all the facilities.

I am highly thankful to my friends and family for their active moral support, valuable time and advice. I am thankful to all of those, particularly the various friends, who have been instrument in cresting proper healthy and constructive environment and including new and fresh innovative ideas during project, without their help, it would have been difficult to complete dissertation within time.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF ACRONYMS / ABBREVIATIONS

ATS	Automatic Text Summarization
NLP	Natural Language Processing
SVM	Support Vector Machine
CSV	Comma Separated Values
XML	eXtensible Markup Language
SQL	Structured Query Language
IRC	Internet Relay Chats
API	Application Programming Interface

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 DATA ANALYTICS

The breaking down of data into such a form that it can useful to other users in the form of important knowledge is known as data analytics. The real scenario of the user's work can be understood better with the help of data analytics process. Better decisions can be made with the help of this process. In order to discover the useful information from the present data, various actions such as inspection, cleansing, transformation as well as modeling are performed which are collectively known as the data analytics process [1]. There are numerous facets as well as approaches present within the process of data examination. Within the different domains, numerous techniques are proposed with separate names. A specific data investigation process in which the modeling is performed, and useful information is extracted such that is can be utilized in predictive manners instead of descriptive manners is known as data mining process. However, the analysis that is performed based on aggregations which is completely dependent on business information is known as business intelligence process. The process through which the complete document is broken down such that the individual components can be investigated separately is known as investigation [2]. The raw data is converted here into useful form such that it can be easily utilized by the users during the decision-production. In order to answer various queries, to test the hypothesis or in order to disprove any theories, the data is gathered from numerous sources of the system and analyzed. There is a procedure that is performed in order to investigate the data. Further, in order to interpret the results, numerous techniques are performed and to arrange the data is proper manner, numerous approaches are also applied. This helps in analyzing the data is a more precise manner and results in dissecting the data with the help of proper measurements applied to it [3].

The manner in which the examination process is to be followed completely depends on the requirements of the users who will utilize the analyzed results or the investigators which are performing such tasks. An experimental unit is referred to as the general entity in which the data is gathered. There is a need to highlight and gather the particular variables related to certain population. The form in which the data is to be gathered can either be categorical or numerical. The following figure shows the analytical process.

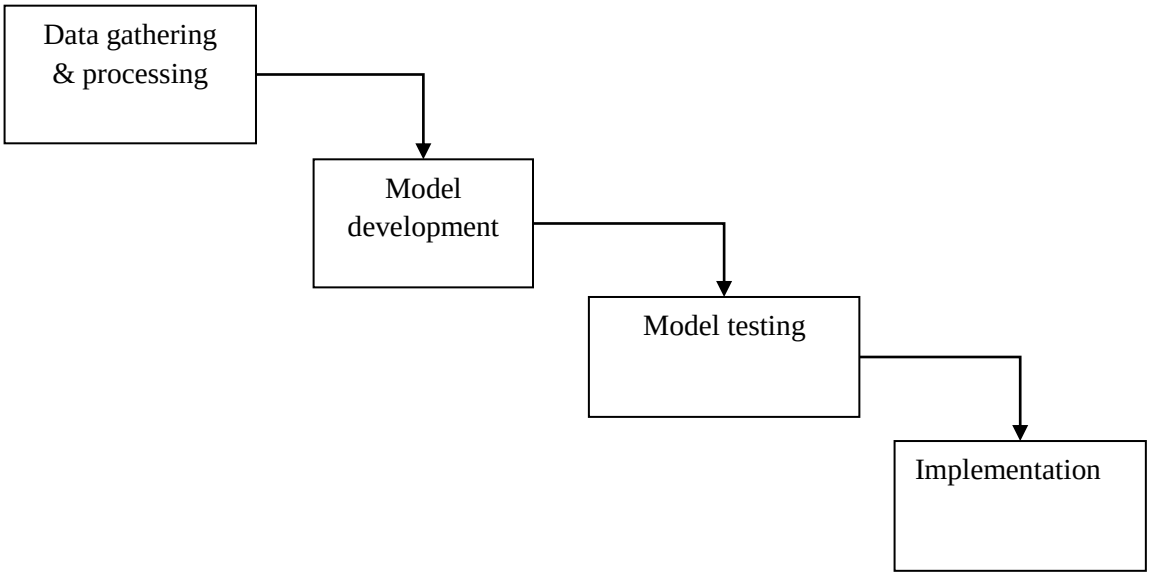

Figure 1.1: Analytics Process

1.2DIFFERENT TYPE OF DATA

Within the big data analysis, there are three major categorizations in which the data can be differentiated. They are structured, semi-structured and unstructured, which are explained further below:

1.2.1 STRUCTURED DATA

The data that can be stored within the database SQL in a tabular form which includes rows and columns is known as structured type of data. The structured data is highly organized. It is easily possible to map the pre-designed fields within this type of data which also has a relation key within it. Within the development scenario, this type of data is the one which is highly processed and the information can be managed very easily within this approach [4]. However, amongst the complete information present, the structure data comprises of only 5 to 10% of the overall data. This resulted in introducing the next category which is known as semi-structured type of data.

1.2.2 SEMI STRUCTURED DATA

The information that is not present within the relational database but still has some authoritative properties such that it can be analyzed in an easy way is known as the semi-structured type of data. The semi structured data is organized into special organizational format. That makes such data easy to handle and analyze. This type of data is stored within the relation database through some processes. There is however a need of ease of space and computations within this type of data. Numerous examples can be given within the applications that have such kind of data which include CSV yet XML and JSON documents, NoSQL documents and so on [5].

This type of category also contributes to around 5 to 10% of the overall data. Thus, the third categorization which is the unstructured data is generated.

1.2.3 UNSTRUCTURED DATA

Around 80% of the total amount of data present today is considered to be the unstructured type of data. The unstructured data does not have any specific structured. There is majorly the text and multimedia type of data present within this category. The data present in the e-mail messages, word documents, images, presentations, videos and various other documents is all included within this category. All the files involved within these documents have some type of internal structure however are considered to be unstructured

as there is no appropriate fixing of this kind of data within the database. So, it cannot be included within the structured type of data [6].

Within almost all the applications, unstructured type of data is present. Most of the applications present today include unstructured type of data within them only. The machine generated or human generated type of data is only present within the unstructured type of data [7].

- Satellite images: As per the surveillance of satellites, there are numerous images received by the governments and the scientists. Such type of data is involved within these types of sources.
- Scientific data: The seismic images, atmospheric data, and high energy physics are involved within this type of data.
- Photographs and video: The security, surveillance as well as traffic videos generate numerous amounts of data as well.
- Radar or sonar data: The vehicular, meteorological as well as oceanographic seismic profiles generate numerous data.

The human-generated unstructured type of data is created by various sources which are mentioned below [8]:

- The text present within the organizations: The various documents, logs, survey results as well as emails of the organizations comprise of numerous amount of text which falls within this category. This type of data comprises of large part of the text information which is existing today.
- Social media data: Through the social media stages, numerous amounts of data are generated as well.
- Mobile data: The various types of data, text messages and information are included within such sources.
- Website content: Any kind of site that delivers unstructured content is categorized as the website content.

As compared to the growth of data within other categories, the unstructured data grows at the highest rate. Within the business decisions, the exploitation of such data is very helpful.

1.3 Natural Language Processing

Natural language Processing (NLP) is an application of the computational linguistics [9]. NLP is used to interpret the text and make it analyzable. NLP is the area of Computer Science and Artificial Intelligence. It deals with the interaction and interpretation of computer and human natural language. In the area of Sentiment analysis NLP has been primarily used. NPL is an important tool in area of artificial intelligence as it helps in interaction of robots in human natural language with humans [10]. It encompasses various techniques for automatic generation, manipulation as well as analysis of the natural human language. Earlier NLP approach is rule based nowadays it is based on deep learning approach that is more flexible for algorithms to learn and identify as well as interpret human language.

1.4 Chat Summarization

With the advancement in the technology, the internet is accessible through various devices like smartphones, smartwatches and within the reach of common people. The communication of users through the social media has been exponentially increased [11]. A considerable piece of information exchange happens as online conversations like Internet Relay Chats (IRC), Facebook and Twitter streams. Among them, we concentrate on conversations from the online support gatherings which plan to examine and resolve user-related issues. Such discussions contain a considerable measure of information which can benefit associations and additionally information-seeking users [12]. However, these gatherings suffer from the problem of information overload and redundancy, where comparable topics get discussed multiple times by different users.

Synopsis is a proven effective approach to tackle these problems. An effective outline provides the fundamental topics of dialog by removing redundant and unwanted information from the conversation [13]. This saves users' time by giving the essence

of the document. These summaries can be easily used to analyze the effect of virtual socialinteractions and virtualauthoritative culture on software or item development.Synopsis has been applied to different text genres, for example, news articles, scientific articles and sites. Be that as it may, very little work has been done on outlining conversations. Conversation synopsis differs from the conventional document rundown in its informational need and structure. Written conversations as emails and chats have features like acronyms, hyperlinks, nicknames and spelling mistakes which make conventional Natural Language Processing (NLP) techniques hard to apply. Text as news articles and books is a monolog whereas conversations fall in the genre of correspondence and requires discourse examination [14]. Real-time conversations comprise a sequence of exchanges between multiple users that might be synchronous or offbeat, and may traverse different modalities. This poses more challenges in examining conversations.

1.5 Sentiment Analysis

Sentiment Analysis also known as the opinion mining. It uses the NLP in order to categorize the opinions of people about the products or the reviews. Sentiment analysis deals with opinions and perspective of human related to emotions and attitude about some occurrent or the event [15]. Opinion mining is most useful in various fields like commercial product reviews, social media analysis and movie reviews etc. the semantic analysis is a valuable technique in creation of recommender systems. The user gives the text reviews like online reviews, comments or the feedbacks on the social media sites, e-commerce websites. This text is an important source of user’s opinions. The sentiment analysis is done to check the positive, negative and neutral opinion of users about products to check its popularity or importance in the market [16].

CHAPTER 2

REVIEW OF LITERATURE

Dan Cao, et.al, *Analysis of Complex Network Methods for Extractive Automatic Text Summarization*. 2016

The Automatic Text Summarization is an important research areain the domain of information systems. It intends to make a compressed version of documents, which should cover all the significant contents and general information. In extractive text summarization, sentences are scored on a few of features. A large number of features network based have been proposed by researchers in the past literatures. This paper reviews every one of the features that utilization metrics and idea of complex network for scoring sentences [17]. The experiment results on single component and combinations of different features we proposed are discussed. Quantitative and qualitative aspects were considered in our assessment performing on the DUe 2002 data sets. Shortest ways demonstrated astounding for summarization, which got the highest scores for the quality of generated summary. Another contribution was the discovery of results that features combinations with a similar kind property of network indicated incredible influence to choose sentences. About sentence relationship between sentences which turned out to be an essential element in the extraction of good rundowns, cause may concern about the structure of text document it inferred well.

Rasim Alguliyev, et.al, *A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization*. 2016

In this paper text summarization is represented as a sentence scoring and selection process. The process is displayed as a multi-objective optimization issue. As a result of the large amounts of text documents are created in the web and e-government and their volume increments exponentially along years. In result, expanding the volume of text documents has made troublesome for clients to read and extract helpful information from them. In this way, with proceeding with increment of the text documents ATS has turned out to be important research direction and therefore attracted the consideration of numerous researchers throughout the previous couple of years [18]. This paper is centered on the extractive text summarization where a

summary is generated by scoring and choosing the sentences in the source text. At first it assesses the score of each sentence and afterward chooses the most representative sentences from the text by considering that semantic similarity between chose sentences will be low. For scoring the sentences another formula is introduced. The proposed show endeavors to find balance amongst coverage and redundancy in a summary. For taking care of the optimization issue a human learning optimization algorithm is used.

Narendra Andhale, et.al,*An Overview of Text Summarization Techniques,* 2016

The process is which the condensed type of document can be generated that can help in recording the significant information and provides importance to the source text is known as text summarization. An important method through which the related information can be identified from huge documents is known as automatic text summarization method. The extractive and abstractive methods are two categories of text summarization techniques. The comprehensive survey of both of the techniques present within the text summarization is presented in this paper [19]. There are various extractive and abstractive types of summarization methods which are studied in this paper. An effective summary is to be generated by summarization method which has less redundancy and involves correct sentences which are grammatically correct. Good results are achieved within the extractive and abstractive methods which can be utilized further by the users. The testing for hybridization is studied within this paper which helps in generating the information which is compressed and readable by the users.

Rupal Bhargava et.al,*MSATS: Multilingual Sentiment Analysis via Text Summarization.*2017

Sentiment Analysis has been a sharp research area for recent years. In any case, a significant part of the exploration that has been done supports English dialect as it were. This paper proposes a strategy utilizing which one can break down various languages to find sentiments in them and perform sentiment analysis. The strategy leverages diverse techniques of machine learning to dissect the text. Machine

translation is utilized as a part of the system to give the component of dealing with various languages. After the machine translation, text is processed for finding the sentiments in the text [20]. With the coming of blogs, forums and online reviews there is substantial text present on internet that can be utilized to break down the sentiment about a specific subject or an object. Thus to reduce the processing it is beneficial to extract the important text present in it. So the system proposed utilizes text summarization process to extract important parts of text and after that utilizations it to examine the sentiments about the specific subject and its aspects. Experiment demonstrates that proposed strategy can deliver promising results.

Archana N.Gulati, et.al, *A novel technique for multi-document Hindi text summarization.* 2017

A text summary is a reduction of original text to condensed text by choosing what is important in the source. Over a period of years, the World Wide Web has expanded with the goal that tremendous measure of data is created and accessible on the web. Text summarization is required when individuals need a essence of a specific topic from at least one sources of information accessible on the web. Thinking about the above issue a novel procedure for multi document, extractive text summarization is proposed [21]. Additionally, considering the normal dialect in India being Hindi, a summarizer for a similar dialect is assembled. News articles on games and governmental issues from online Hindi newspapers were utilized as contribution to the system. Fluffy inference motor was utilized for the extraction process utilizing eleven important features of the text. The system accomplishes an average precision of 73% over multiple Hindi documents. The summary generated by the system is discovered near summary generated by humans. The Precision, Recall and F-score values demonstrates good accuracy of summary generated by the system.

Manisha Gupta, et.al, *Text Summarization of Hindi Documents using Rule Based Approach.*2016

Automatic summarization assumes an important part in document processing system and information recovery system. Era of summary of a text document is an important piece of NLP. There are a number of situations where automatic construction of such

outlines is helpful. Text summarization is that process which converts a larger text into its shorter shape keeping up its information. Summary of a more drawn out text saves the reading time as it contain lesser number of lines yet exceedingly important information of the original text document. In this paper we present a novel approach for text summarization of Hindi text document based on some linguistic principles [22]. Dead wood words and phrases are likewise removed from the original document to generate the lesser number of words from the original text. Proposed system is tested on different Hindi sources of info and accuracy of the system in type of number of lines extracted from original text containing important information of the original text document. Info text size can be decreased to 60% - 70 % with the assistance of proposed system. System generates the extractive summary given by the client i.e. it doesn't generate the summary of the text on the premise of the semantics of the text.

Akshi Kumar et al. ***Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization.*** **2017**

There are various different algorithm that are used in the text summarization. The algorithms are differentiated according to the extractive text summarization and abstractive text summarization approach [23]. In this paper author has analyze and compare the performance of three different algorithms. Firstly, the different text summarization techniques explained.Extraction based techniques are used to extract important or keywords to be included in the summary.Abstraction based techniques generates its own sentences for text summary. For comparison three keyword extraction algorithms namely TextRank, LexRank, Latent Semantic Analysis (LSA) were used. Three algorithms are explained and implemented in paython language. The ROUGE-1 is used to evaluate the effectiveness in extracted keywords. The result of all algorithms compares with the handwritten summaries to evaluate the performance of the algorithms at the end the performance of TextRank Algorithm is much better than other algorithms.

N. Moratanch et al. ***A Survey on Extractive Text Summarization.*** **2017**

The text summarization has various techniques that are classified as the extractive and abstractive approaches of summarization of the text [24]. In this paper the author has

presented the comprehensive review of extraction based text summarization techniques. In this paper the author provides survey on extractive summarization approach by categorized them in: Supervised learning approach and Unsupervised learning approach. Then different methodologies, the advantages are presented in the paper. The author also includes various evaluation methods, challenges and future research direction in the paper.

Mihai Dascălu, et.al, *Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing.*2011

With the expanding fame and advancement of Computer Supported Collaborative Learning frameworks, the requirement for developing a tool that automatically surveys instant informing discussions has turned out to be basic. The primary reasons are the high volume of information and the increased amount of time spent for physically evaluating discussions. We propose an automated examination framework in view of Natural Language Processing (fixated on Latent Semantic Analysis and Social Network Analysis) and enhance its runtime performance by methods for distributed figuring [25]. Besides, we give an exceptional grading component in light of a multilayered engineering and initiate an increase of speedup by sending a Replicated Worker design. Load balancing and fault tolerance represent key parts of this approach, other than the genuine increase in performance. We presented comes about demonstrating that the framework is fit for assessing a corpus of chats in a timely way, conceding quick access to feedback for participants. We likewise presented usage subtle elements on a distributed form of the instrument, a solution that significantly increases the general performance, permitting bigger corpuses to be investigated in a littler amount of time. The framework performs and scales well under a wide assortment of conditions and loads.

Wen Hua, et.al, *Understand Short Texts by Harvesting and Analyzing SemanticKnowledge.* 2016

There are numerous challenges being faced during the presence of short messages within various applications. The punctuation of the composed language is not checked within the sort messages. The connection amongst the natural language processing tools and the part-of-speech tagging for dependency parsing is not seen within these

applications. There are no appropriate statistical signs present within the short messages which might provide support to the various content mining approaches being utilized within these systems [26]. There are large numbers of short messages generated which are mainly questionable as well as noisy which is also a major challenge as it is very difficult to be handled. A prototype framework is proposed in this paper in order to understand the short messages. This will help in providing semantic knowledge which can further be utilized in order to provide automatic harvesting of the web content generated. The performance is evaluated here with the help of various simulation experiments. On the basis of results achieved it can be seen that the proposed technique provides better results and helps in analyzing the short messages in better way.

Pierre Ficamos et el. ***A Naïve Bayes and Maximum Entropy approach to Sentiment Analysis: Capturing Domain Data in Weibo*****. 2017**

As the social media become more popular nowadays, the more researches have been focusing on automatic processing and extracting the sentiment information from the large data [27]. In this paper the author proposed a feature extraction method that relays on Part Of Speech (POS) tags. That helps in selection of the unigram and bigram features. The paper focuses on the sentiment analysis of the Chinese social media. The grammatical relations between the different words are used in construction of the bigram and unigram features. The experiment shows that the proposed method provides the better results with the Naïve Bayes.

Ankur Goel et el. ***Real time sentiment analysis of tweets using Naive Bayes.*** **2016**

Twitter is the most popular micro-blogging websites where people share and expresses their views on various topics, things, product, services and happenings [28]. The people's views servers as important data information for the purpose of evaluating different products, services and events. There are various classification techniques that are available to classify the tweets into different classes like Positive, Negative and Neutral based on there sentiments. This paper proposed the to improve the classification. The paper shows that uses of SentiWordNet along with Naïve Bayse can improve the accuracy of the tweets classification. The implementation is

done in Python with NLTK and the python twitter APIs are used. the final experiment shows the classification accuracy improved to a considerable extent.

Shweta Rana et el. ***Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques.*****2016**

With the expanding web and social networking people starts to share data and information online. This social media data can be used for Sentiment analysis [29]. In this research paper the author analyzed the sentiment of movies reviews. Three different algorithms Naïve Bayes, Synthetic words and Linear SVM have used and compared. The results generated by these algorithms indicates that Linear SVM algorithm provides the best accuracy. The author suggests for future to identify accuracy rate of different products.

Huma Parveen et el. ***Sentiment analysis on Twitter Data-set using Naive Bayes algorithm.*** **2016**

As from last few years the use of Social Media websites has increased extremely. Various kind of user data has been generated on these websites like the chat data, comments, reviews of products [3. This data plays an important role in determining the satisfaction level of users regarding different products. In this research paper the author discusses the techniques of sentiment analysis in order to extract the sentiments of twitter posts. This helps in the business predictions. The Hadoop Framework is used in the research paper in order to process the data sets of movies. The data set that is used is in the forms of tweets. The twitter API is used to extract the tweets. The Naïve Bayes algorithm is used to train the SentiWordNet dictionary. The n uses to implement for sentiment analysis. The results shown on different categories of sentiments like positive, negative and neutral. For the future scope author offer to also includes data from Google and Facebook.

Wiraj Udara Wickramaarachchi, et.al, ***An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis.*****2017**

One significant method of expressing the opinion of the users of social network is expressing genuine feelings or emotions through chats and comments for images, status or recordings that has been uploaded to social network. This will increases the effectiveness of the communication among users since they have no face-to-face cooperation [31]. Content processing strategies are utilized to distinguish emotion which communicated through content. The research proposes a way to deal with get general emotion from comment content towards a picture, which uploaded to social network. The new methodology was developed as an upgraded extension of the previous works and utilizing fitting change and extension to them with Latent Semantic Analysis (LSA) on the grounds that previous researches have proven that LSA is a light weight approach. What's more, it is trusted that, online emotion foreseeing frameworks must be light weight. The research accomplished more efficiency than previous works because of light weight of the methodology, additionally presented a prototype of GUI. Likewise the topic is open for future enhancement.

CHAPTER 3

PRESENT WORK

3.1 PROBLEM FORMULATION

When academic and business ventures are discussed, electronic documents forms the crucial part of receiving and transferring information. There is no use of online information if we cannot extract it and use it to cater our ventures. Text summarization is a great technique that serves our purpose. In order to frame up summary; it is required to find the relevant text with complete omission of unnecessary information while keeping the focus on details and compile them into a document. This is not as easy as it seems to be as the common constrains of natural language processing are commonly encountered. It is highly reliable to depend on a solution which can automatically frame a summary of two or more texts and that is called as text summarization. Summary is categorized into two parts: extractive which means complete explanation of sentences, phrases etc and abstractive means short summary of a particular subject. The sentiment analysis is the technique which can analyze the behavior of the user. It depicts whether the chat is positive, negative or neutral. The research focuses on chat summarization and sentiment analysis of chat.

3.2 OBJECTIVESOF THE STUDY

Following are the various objectives of this research

1. To study and improve lexical analysis technique using weight based text summarization technique

2. To propose technique for the sentiment analysis using technique of classification

3. Implement proposed technique and compare with existing technique in terms of various parameters.

3.3 RESEARCH METHODOLOGY

The summarization technique includes following steps:

1. Dataset inherited :- The data which is given as input will be taken from the twitter. The data will be downloaded using the twitter API
2. Data Pre-Processing :- In the second phase, the data which is taken as input will be pre-processed means the un-wanted data will be removed.
3. Analyzing features of the Dataset :- The dataset which is pre-processed and on that data algorithm of n-gram is applied for the feature extraction
4. Chat Summarization:- In the last step, the rating to each word is given on the basis of their occurrences and the words with maximum rating is considered as most important words that are included in the final chat summary and others are removed.

3.3.1 CHAT SUMMARIZATION

The pattern based algorithm is the algorithm which generates patterns of the input data. The input data will be divided into certain phases and these phases are generated using the N-gram algorithm which will make combinations with the others words in the dataset. The weight is assigned to each word, character in the chat for generation of final chat summary. The patters are generated using N-Gram algorithm.

Flowchart of Sentiment Analysis : The flow chart of the proposed technique is shown on next page.

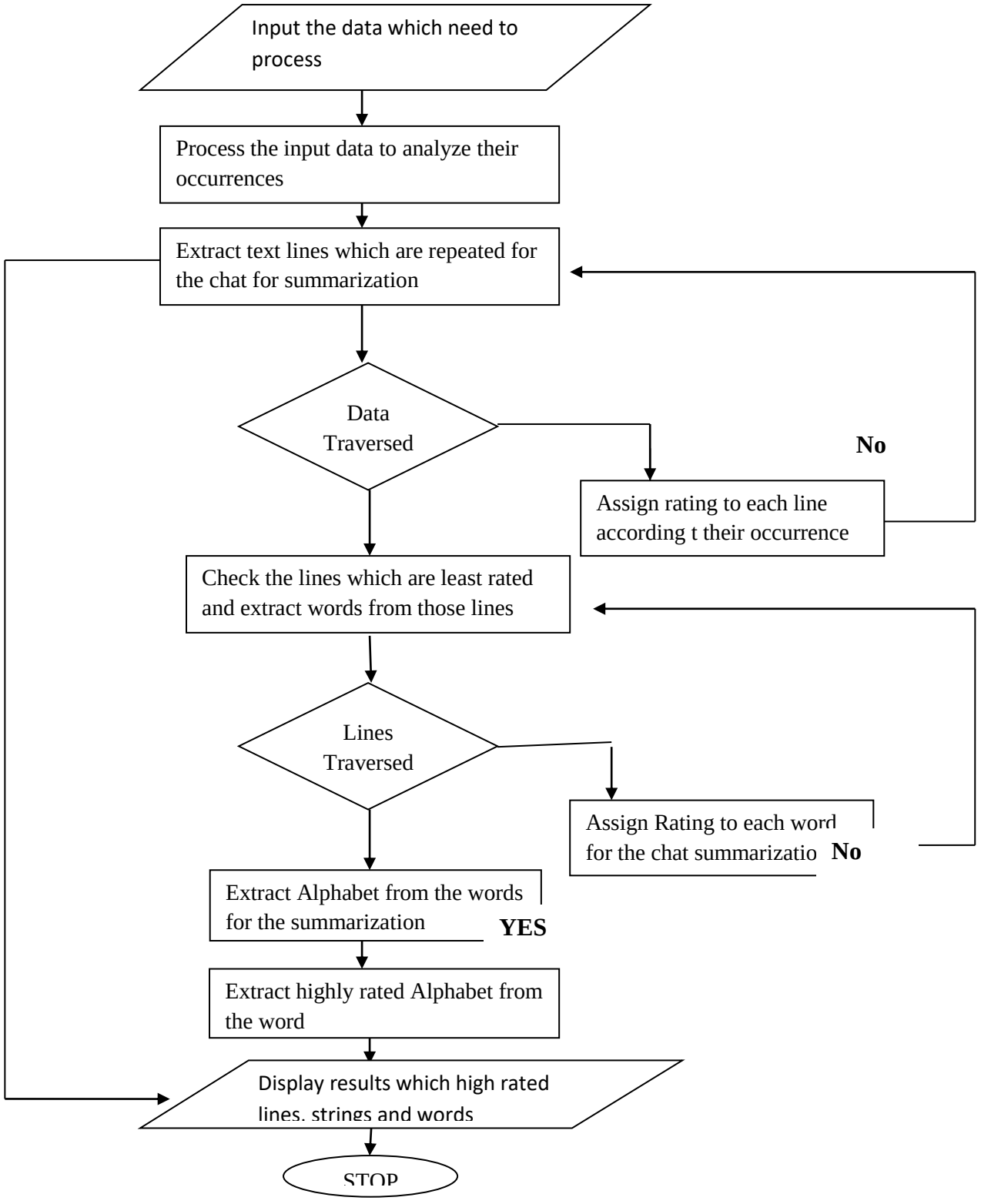

Figure 3.1: Flowchart of Proposed Chat Summarization Technique

3.3.2 SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is a way to evaluate written or spoken language to determine if the expression is favorable, unfavorable or neutral and to what degree. In this technique, the chat is analyzed to depict the behavior of the users.The chat is positive if the score after the analysis is greater than zero or negative otherwise. If the analysis score is equal to zero, the chat is neither positive nor negative. The chat is then categorized as neutral.

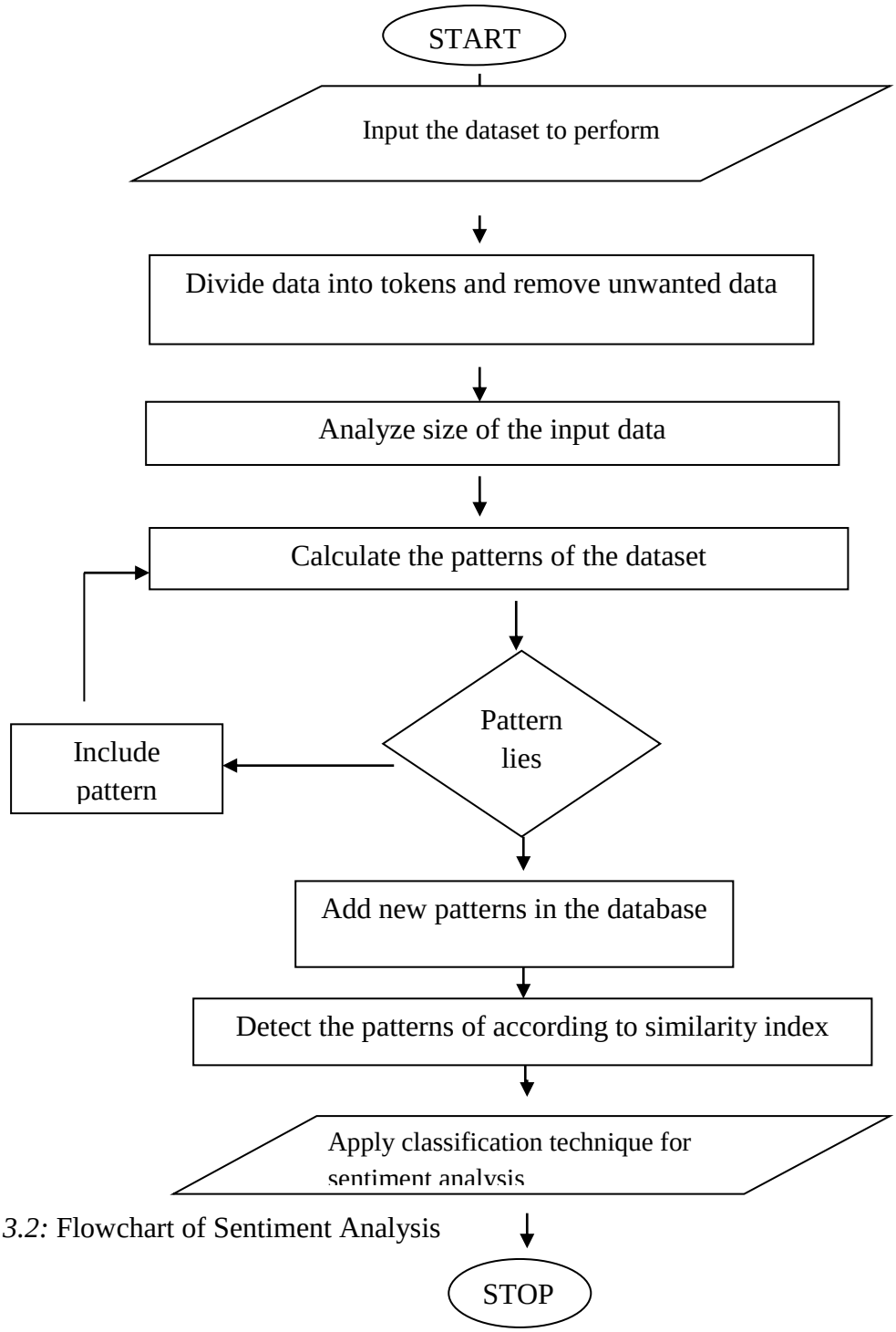

Figure 3.2: Flowchart of Sentiment Analysis

3.4EXPECTED OUTCOMES

Following are the various expected outcomes of this research:

1. The text summarization is the technique which can summarize the data according to their important. In the base paper, ontology-based technique is proposed which is improved in this research. This directly leads to increase accuracy of text summarization.

2. The sentiment analysis is the technique in which the positive, negative and neutral opinions are analyzed. The proposed improvement reduces the execution for sentiment analysis.

CHAPTER 4

CONCLUSION

4.1 CONCLUSION

In this work, it has been concluded that text summarization is the technique which can summarize the data according to their importance. The ontology-based technique is the most efficient technique to generate summarized data. The number of times the word repeats in the text is the key point in ontology-based technique to generate text summary. The higher the number of times a word repeats, higher is its importance as compared to other words. In this research, the ontology-based technique is improved for text summarization which increases its accuracy and sentiment analysis will be done with the classification technique which reduces the execution time.

REFERENCES

[1] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen. *Summarizing email threads.* In Proceedings of HLT-NAACL 2004: Short Papers, pages 105–108, 2004.

[2] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval. Information Processing and Management*, 24:513–523, 1988.

[3] O. Sandu. *Domain Adaptation for Summarizing Conversations.* PhD thesis, Department of Computer Science, The University Of British Columbia, Vancouver, Canada, 2011.

[4] S. Teufel and M. Moens. *Summarizing scientific articles: Experiments with relevance and rhetorical status.* Computational Linguistics, 28:409–445, 2002.

[5] J. Ulrich, G. Murray, and G. Carenini. *A publicly available annotated corpus for supervised email summarization.* In AAAI08 EMAIL Workshop, Chicago, USA, 2008. AAAI.

[6] D. C. Uthus and D. W. Aha. *Plans toward automated chat summarization*. In Meeting of the Association for Computational Linguistics, pages 1–7, 2011.

[7] C. Whitelaw, B. Hutchinson, G. Chung, and G. Ellis. *Using the web for language independent spellchecking and autocorrection.*In Empirical Methods in Natural Language Processing, pages 890–899, 2009

[8] L. Zhou and E. H. Hovy. *Digesting virtual geek culture: The summarization of technical internet relay chats.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 298–305, 2005

[9] Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage. *A method to extract essential keywords from tweet using NLP*. 2016 16th International Conference on Advances in ICT for Emerging Regions(ICTer).

[10] Ibrahim A. Hameed. *Using Natural language processing for designing socially intelligent robots.* 2016 Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob).

[11] L. SUanmali, M. S. Binwahlan, and N. Salim. *Sentence features fusion for text summarization using fuzzy logicin Hybrid Intelligent Systems*. 2009, HIS'09, Ninth International Conference on, vol. 1, IEEE, 2009, pp. 142-146.

[12] L. Suanmali, N. Salim, and M. S. Binwahlan. *Fuzzy logic based method for improving text summarization.* arXiv pre print arXiv:0906.4690, 2009.

[13] X. W. Meng Wang and C. Xu. *An approach to concept oriented text summarization,* Proceedings of ISClTS05, IEEE internationalconference, China,1290-1293" 2005.

[14] M. G. Ozsoy, F. N. Alpaslan, and 1. Cicekli. *Text summarization using latent semantic analysis.* Journal of Information Science, vol. 37, no. 4, pp. 405-417, 2011.

[15] Adyan Marendra Ramadhani, Hong Soon Goo. *Twitter Sentiment Analysis using Deep Learning Methods.* 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.

[16] K. Kaviya, C. Roshini, V. Vaidhehi, J. Dhalia Sweetlin. *Sentiment for Restaurant Rating.* 2017 IEEE International Conference on Smart Technologies and Management for Computing, Controls, Energy and Material (ICSTM).

[17] Dan Cao, Liutong Xu. *Analysis of Complex Network Methods for Extractive Automatic Text Summarization.*2016 2nd IEEE International Conference on Computer and Communications.

[18] Rasim Alguliyev, Ramiz Aliguliyev, Nijat Isazade. *A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization.* 2016, IEEE.

[19] Narendra Andhale, L.A. Bewoor. *An Overview of Text Summarization Techniques.* 2016, IEEE.

[20] Rupal Bhargavaand Yashvardhan Sharma. *MSATS: Multilingual Sentiment Analysis via Text Summarization.*2017, IEEE.

[21] Archana N.Gulati, Dr.S.D.Sawarkar. *A novel technique for multi-document Hindi text summarization.* 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017).

[22] Manisha Gupta, Dr.Naresh Kumar Garg. *Text Summarization of Hindi Documents using Rule Based Approach.*2016 International Conference on Micro-Electronics and Telecommunication Engineering.

[23] Akshi Kumar, Aditi Sharma, Sidhant Sharma,Shashwat Kashyap. *Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization.* International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

[24] N. Moratanch, S. Chitrakala. *A Survey on Extractive Text Summarization.*IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017.

[25] Mihai Dascălu, Ciprian Dobre, Ştefan Trăuşan-Matu, Valentin Cristea. *Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing.*2011 10th International Symposium on Parallel and Distributed Computing.

[26] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. "*Understand Short Texts by Harvesting and Analyzing Semantic Knowledge.* 2016, IEEE.

[27] Pierre Ficamos;Yan Liu, Weiyi ChenA. *Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo.* 2017 IEEE International Conference on Big Data and Smart Computing (BigComp).

[28] Ankur Goel, Jyoti Gautam, Sitesh Kumar.*Real time sentiment analysis of tweets using Naive Bayes.* 2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[29] Shweta Rana, Archana Singh. *Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques.*2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[30] Huma Parveen, Shikha Pandey. *Sentiment analysis on Twitter Data-set using Naive Bayes algorithm.*2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

[31] Wiraj Udara Wickramaarachchi, R. K. A. R. Kariapper. *An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis*.2017 2nd International Conference on Image, Vision and Computing.