A Dissertation
On

" A Fuzzy Logic Based Text Summarization "

Submitted in partial fulfillment of the requirement
for the award of degree of

MASTER OF TECHNOLOGY
Software Engineering
Delhi Technological University, Delhi

SUBMITTED BY

Aditi Sharma
2K15/SWE/02

Under the Guidance of

DR. AKSHI KUMAR

Assistant Professor
Department of Computer Science & Engineering
Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
2017

CERTIFICATE

This is to certify that the dissertation entitled **"A Fuzzy Logic based Text Summarization"** has been submitted by **Aditi Sharma (Roll Number: 2K15/SWE/02),** in partial fulfillment of the requirements for the award of Master of Technology degree in Software Engineering at **DELHI TECHNOLOGICAL UNIVERSITY**. This work is carried out by her under my supervision and has not been submitted earlier for the award of any degree or diploma in any university to the best of my knowledge.

(DR. AKSHI KUMAR)
Project Guide
Assistant Professor
Department of Computer Science & Engineering
Delhi Technological University

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. My greatest thanks are to my parents who bestowed ability and strength in me to complete this work.

I owe a profound gratitude to my project guide **Dr. Akshi Kumar** who has been a constant source of inspiration to me throughout the period of this project. It was her competent guidance, constant encouragement and critical evaluation that helped me to develop a new insight into my project. Her calm, collected and professionally impeccable style of handling situations not only steered me through every problem, but also helped me to grow as a matured person.

I am also thankful to her for trusting my capabilities to develop this project under her guidance.

I would also like to express gratitude to Mrs. Arunima Jaiswal (Research Scholar, Delhi Technological University) for providing me continuous support and guidance during this project.

I would also like to express my gratitude to the university for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

Aditi Sharma

2K15/SWE/02

ABSTRACT

Today's world is all about information, with most of it online which enables anytime, anywhere, easy and unlimited access; participation & publishing of information has consequently escalated the suffering of 'Information Glut'. Assisting users' informational searches with reduced reading or surfing time by extracting and evaluating accurate, authentic & relevant information are the primary concerns in the present milieu. Automatic text summarization is the process of condensing an original document into shorter form to create smaller, compact version from the abundant information that is available, preserving the content & meaning such that it meets the needs of the user. Though many summarization techniques have been proposed but there are no 'silver bullets' to achieve the superlative results as of human generated summaries. Thus, the domain of text summarization is an active and dynamic field of study, practice & research with the continuous need to expound novel techniques for achieving comparable & effectual results.

Fuzzy logic has appeared as a powerful theoretical framework for studying human reasoning and its application has been explored within the domain of text summarization in the past few years. One key aspect of text summarization is accurate identification of keywords from the given textual content. In this project, a new technique based on fuzzy logic has been proposed using two graph based techniques named as TextRank and LexRank and one semantic based technique named as Latent semantic analysis(LSA). In our work, we have also investigated their relative performance with the proposed method. All these methods used in developing hybrid model are of extractive summarization type. The techniques are evaluated on Opinosis data set using 'ROUGE-1' and 'time to extract the keywords'. The proposed technique has outperformed the existing techniques, when compared with the results given by the original studies.

List of Figures & Tables

Contents

CHAPTER 1

INTRODUCTION

This chapter briefly introduces the research work proposed in the thesis. Section 1.1 gives an overview of the research undertaken. Section 1.2 discusses automatic text summarization and then various methods of summarization are briefly described in section 1.3. Section 1.4 explains the motivation behind the proposed method, the scope of the method is discussed in section 1.5. Section 1.6 enlightens the research objectives. Section 1.7 presents an outline of this thesis and labeling the remaining chapters. Finally, Section 1.8 gives the summary of the chapter.

1.1. OVERVIEW

According to www.worldwidewebsize.com, the indexed Web contains at least 4.5 billion pages (Monday, 20 March, 2017). With the massive proliferation in the velocity, volume and variety of information accessible online and the consequent need to develop viable paradigms which facilitate better techniques to access this information, there has been a strong resurgence of interest in Web Information Retrieval (Web IR) research in recent years. The ultimate challenge of Web IR research is to provide improved systems that retrieve the most relevant information available on the web to better satisfy a user's information need [1, 2]. Moreover, with the transformation of Web into a customary decision-support and recommendation tool, tackling the challenge of "information overload" on the Web has become increasingly vital. The Web IR research is typically organized in tasks with specific goals to be achieved [2]. With the ease in availability of Internet and increased use of smart devices like mobile phone, laptop, tablets the access and storage of data has escalated rapidly, resulting in need of tools which can enhance the user's productivity and experience. A simple keyword search on the Internet results in hundreds of result in less than a second, some of which are not even relevant to the users' query and finding the pertinent information is a difficult and time consuming task.

Summarization has been identified as an effective Web IR task, which helps users to locate the right information at the right time thus facilitating timely decisions.Human summarization can be biased, context-dependent and may vary with human cognition. Thus, suitable techniques & tools are needed to extort pertinent and imperative sections such that critical information in the form of summary is

acquired; providing a machine generated summary free from bias. The idea is to create smaller, condensed versions from the abundant information that is available, preserving the content & meaning such that it meets the needs of the user.

The following real-world analogy helps us understand the meaning and need of summary: A person wants to decide his/her visit to an exhibition at a Local Art Gallery; the display list suggests the highlights of the exhibition giving an insight to what is the theme of the exhibition, though nothing really can be said about the quality of art-work presented but the visitation decision can surely be made depending on the person's interest to the summarized highlights. Correspondingly, the articles on the Web such as blogs do not contain a summary or abstract, so to find whether they are valuable, relevant and of interest would require reading the whole document, which is cumbersome, lengthy and annoying especially if after reading few pages it's found irrelevant. Text Summarization tools come to rescue by reducing the size of original textual document to create a summary with just the non-trivial content retained thus assisting users.

Automatic text summarization techniques can be used for only extracting the keyword too. As, no matter what the intent, type or context of summary generated is, the primary objective is to assist users' informational searches with reduced reading/surfing time and also improve the document indexing efficiency at the same time. This will fasten the search process, as the relevance of an article to our interested topic can be deducted by the important keywords of the article. Extracting keywords using text summarization algorithm can optimize the search process.

A text summarization method has been proposed which is a hybrid of four techniques namely TextRank, LexRank, Latent Semantic Analysis and Fuzzy logic each having their own pros and cons . Unitedly, resulting in more confident results. The details of the technique proposed is discussed in Chapter 2.

1.2 Automatic Text Summarization

Automatic Text Summarization is the process of finding a summary of a document available on the Web by a computer without changing the meaning of the original document for retrieving the useful information like the structure of document [8].Text summarizers have demonstrated their use in various application domains that range from stock market prediction to keyword extraction for search optimization. More recently, automatic Email summary generation using cue words has also been suggested by Carenini et al [45]. An automated generated summary can be of various types depending on the purpose and kind of data available for summarization. This categorization of text summarization techniques is depicted in the figure 1 below [9].

Primary categorization of text summarization techniques is on the basis of the type of summary generated. It can either be of extractive or abstractive type. Generating abstractive summary is cumbersome as it gives summary with sentences different from the original document, though the meaning of information is preserved. The content is presented precisely using natural language processing techniques and is a costly, time-consuming process. On the other hand, Extractive text summarization uses sentences from the document to provide condensed form of the document that is in simple terms, it is the subset of the actual document. Most of the studies on text summarization are on extractive techniques [10].

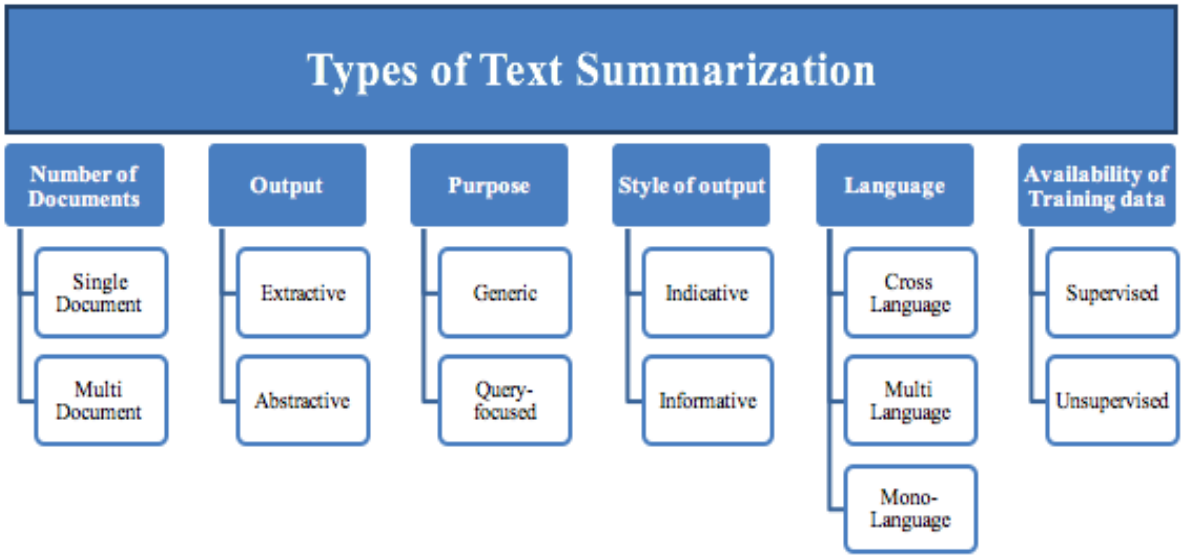

Figure 1. Categories of Text Summarization based on different characteristics.

Other categorizations can be on the basis of the purpose, that is, the summary generated can be either generic, for everyone like summary of a news article, or it can be query-specific/ topic specific, only for a particular user or group of users, generated on the request, e.g. if a user inputs a query regarding weather, then only the important sentences related to weather will be extracted from the document. Moreover, the document can be a single document or a set of documents and the techniques which are applicable to single document may not necessarily be applicable on multi-documents. Further, on the basis of the style of the output, the summaries can be indicative or informative. The former tells what the document is about, and the later gives information on the topic of the document [9].

Summarization techniques can also be categorized on the basis of whether the training data is available or not, that is if it's available they are referred to as supervised in which training data is used to learn about what type of summaries is to be created. If training data is not provided the techniques are known as unsupervised learning, they are preferable for the newly observed data regarding whom

no prior information is available. Another categorization of summarization is on the basis of language. If the language of the document and the summary is same, it is known as monolingual summary. If the document to be summarized is available in various languages, it is called as multi-lingual. If the language of summary and the document is different, it is referred to as cross-lingual summarization [9].

1.3 Summarization Methods

Due to the inherent complexity of generating abstractive summary, extractive summaries have been more frequently generated and used in practical applications [9]. Significant literature studies establish that various types of methods can be used to generate extractive summaries. These include statistical based methods, graph-based methods, discourse based methods, topic based methods, machine learning based methods and swarm based or optimization based techniques [10]. We briefly explicate these methods.

- ***Statistical Based Methods:*** The methods generate summaries using statistical features of the document like sentence position, centrality of the sentence, sentence length, numeric data in sentence, title similarity etc.[8]. These techniques are language independent and do not require much storage or fast processors.

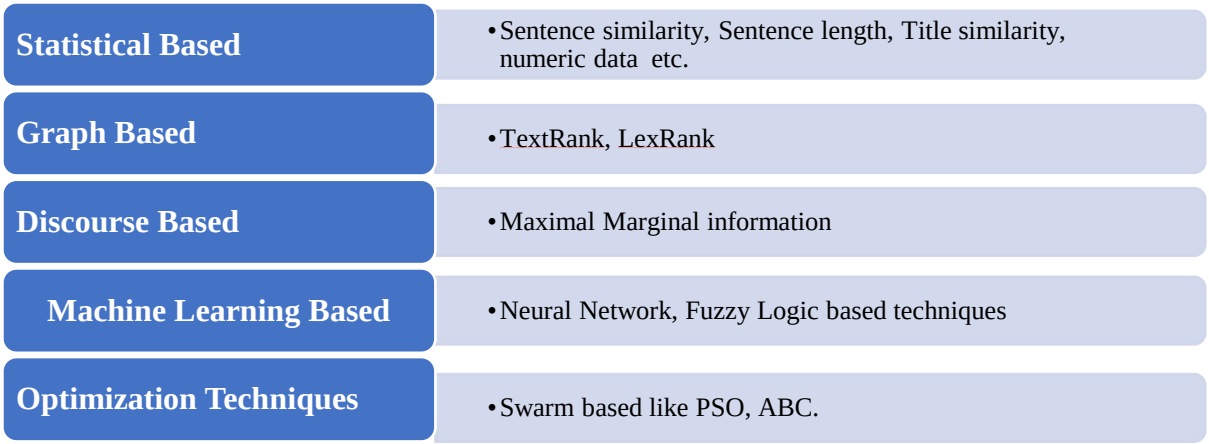

Fig 2. Summarization Methods

- ***Graph Based Methods*:** In this method, the words or sentences are represented by nodes of the graph and edges between these node represents the similarity value between these nodes. The sentences to be taken in extractive summary are found by traversing the graph

and selecting the sentences which have similarity index above the defined threshold. Some of the recognised graph based methods are Text Rank and LexRank[9].

- ***Discourse Based Methods:*** These methods require understanding the textual structure and are complex to use as they take into account the connections between sentences and parts in a text. The Inter-paragraph & Intra-paragraph analysis is also done. Three levels of discourse structure may be identified based on cohesion (relations between textual elements), coherence (relations between ideas expressed in the text realization), and cross-document relations [9].
- ***Topic Based Methods:*** In this method the summary is generated by firstly identifying the subject or theme of the document. Then this is used to extract the sentences which are related to the subject [9].
- ***Machine Learning Based Methods:*** These include approaches which learn from the data provided to the machine for summarization. They can either be supervised where the training data is provided with the summaries of the document such that the machine can learn how to summarize the data or unsupervised where only the documents are provided and the machine learns by analyzing the documents. Unsupervised methods are suitable for new data where similar content is unavailable to us. Some of the machine learning techniques are neural network, SVM, Genetic algorithm and fuzzy logic[8].
- ***Optimization techniques:*** The techniques use nature inspired or swarm algorithms for finding summaries or features for summaries employing optimization algorithms like particle swarm optimization, artificial bee colony. These techniques are usually used in combination with other techniques [10].

A great deal of work has been done in these areas and the recent techniques proposed have mostly been machine learning based, swarm intelligence based or the hybrid of two or more types of summarization techniques. R. Abbasi-ghalehtaki et al [15] proposed a hybrid technique for summarization using machine learning based method, optimization method, and a statistical method.

1.4 Motivation

Due to the increase in Web 2.0 services, the amount of data available online has changed drastically in terms of volume as it has become a global source of useful information. But to get the useful information from all the data available over the web, we have to go through a lot of data. To get the

relevant data quickly, the important keywords are extracted from the articles, for getting the overview of the article. These keywords make it easy for search engine and the user for identifying the relevant documents.

When a keyword is searched over an internet, if the search engine has to search through only the important keywords of the article rather than whole of the article, more relevant results could be fetched. For example, I want to read a paper having comparison of TextRank, LexRank and LSA, If I search these terms on a search engine, say Google, it will fetch all the articles containing these three words in it, even if it is a survey paper, just containing the name of these three techniques just once in the article. Whereas, if we do the same search on some digital library, which has options for advance search, we can select that the search should be conducted only on Keywords. The first search will result in thousands of articles, while the search only in keywords can restrict the articles even to less than hundred. So, to find the more suitable document, the important keywords of the article plays a major role.

In the example mentioned above, the digital libraries like IEEE, ACM contain keywords of the articles only because they are provided by the authors of the research papers. But not every article available over the web contains the list of keywords representing the central idea of the article. Therefore, to automatically find the keywords representing main idea of the article we have proposed a keyword extraction algorithm utilizing the benefits of four techniques.

The use of fuzzy logic in ATS since its first occurrence till date (2003 to 2017) has never been combined with graph based methods except for with Bushy path in 2016 by jyoti yadav [7]. Although the fuzzy has been used in combination with LSA a few times, but its effect with graph based methods has never been explored. So, we proposed the hybrid model of all the four techniques with different weights assigned to each method.

1.5 Scope

Fuzzy logic has appeared as a powerful theoretical framework for studying human reasoning. It formalizes human reasoning by setting rules in natural language used to explain decisions from human reasoning [6]. It reinforces flexibility for reasoning, which makes it possible to take into account inaccuracies and uncertainties. Use of fuzzy logic in WebIR has been amply investigated with valuable findings to ease the process of Information Retrieval, for example fuzzy logic has been used to extract key phrases from news articles [7] and from other web articles. Subsequently the capabilities of fuzzy logic have been extended to the classical text summarization models which were based on pure statistics. Fuzzy logic based text summarization has been identified as a novel and a strategic paradigm

which combines fuzzy logic to the statistics based learning algorithms to improve the quality of summaries.

TextRank and LexRank are two popular graph based algorithm for unsupervised text summarization. TextRank is a simple application of the PageRank algorithm [5] which uses voting-based weighting to determine the significance score of a sentence (nodes in the graph) on the basis of incoming and outgoing edges, where weight of each edge is predetermined on the basis of similarity score between the sentences. LexRank on the other hand differs from TextRank as it is a cosine transform based weighting algorithm. Both split the original text into sentences, building graphs using sentences as nodes and then applying PageRank algorithm to score the sentences and sorting them according to significance scores thus summarizing the important information in the text after selecting sentences which are more significant than others.

LSA is another popular method in NLP that attempts to identify and summarize on the basis of semantics of the text. The algorithm tries to identify the underlying concept in a document, and then can extract keywords on the basis of conceptual similarity. The algorithms was introduced in 2002 by Yihong Gong and Xin Liu. In our evaluation project, we chose this algorithm as a candidate due its dissimilarity with the above three algorithms and due to its intuitiveness to perform well.

Above stated algorithms were chosen because they belong conceptually distinct approaches to text summarization. While TextRank and LexRank belong to the category of Structure-based approach, where TextRank extracts keyword by finding common frequent occurring keywords while LexRank is more of a diversity-based technique, LSA belongs to Semantic-based summarization approach [6] and fuzzy logic is a machine learning approach which helps in incorporating a human like thinking in the process.

LexRank and TextRank are two pioneering graph based algorithms which are widely cited in research in the field of ATS. The use of graphs to determine the key textual content by understanding the underlying structure of the language makes these algorithms appealing candidates. Both of the algorithms were chosen in order to compare the effectiveness of their similarity computation method. LSA is semantical analysis and tries to understand meaning instead of just structures of the sentences hence focusing on the central tendency or abstract of the text rather than structural quality, thus providing better human like summaries as compared to the other techniques stated above. Fuzzy Logic formalizes human reasoning by setting rules in natural language used to explain decisions from human reasoning [6]. Different weights can be assigned to features using fuzzy logic like done by humans on the basis of effect of each feature on the keyword selection process. The hybrid of all these four

techniques will result in better results as they consider different aspects of the article, bringing the most important keywords of the article in the lame light.

The use of fuzzy logic in text summarization was first testified in 2003 by Witte & Bergler[44]. From them onwards till date, a lot of fuzzy logic based methods has been proposed some of them being hybrid with other summarization techniques. But none of the studies have combined the result of fuzzy with graph based and meaning based technique. Although in [6] a hybrid model has been proposed utilizing fuzzy logic based text summarization's results with LSA providing the enriched results. TextRank and LexRank are two of the most popular graph based algorithms for ATS. So we have chosen these four techniques to generate a hybrid model by providing different weights to each technique to get the better results. We also identified and compared various automatic extractive text summarization techniques which use fuzzy logic on the basis of the results given in the original studies.

1.6 Research Objectives

The main research objectives of the work done in this thesis are:

Research objective 1 – To study the different Fuzzy logic based Automatic text summarization methods.

Research objective 2 - To propose an optimized Fuzzy logic based hybrid model for automatic keyword extraction which results in more like human-generated keywords from an article to enhance the search process.

Research objective 3 – To identify the scope of the fuzzy logic based methods in ATS.

The objective of this thesis is to find an algorithm which can be a hybrid approach to extract keywords of the articles with an improve accuracy.

1.7 Organization of Report

This thesis is structured into 6 Chapters followed by references and three appendix.

Chapter 1 presents the overview, research objectives scope and motivation of the project. Finally, analyzing the need for solution for which research is done.

Chapter 2 provides the essential background and context for this thesis and provides a complete justification for the research undertaken in this thesis.

Chapter 3 gives the details of the methodology employed and outlines the use of Fuzzy logic in ATS which is proposed approach.

Chapter 4 describes the implementation of algorithm. It discusses all the input sets, platform and tool used to implement result and to compare them.

Chapter 5 describes the experimental results obtained from the given datasets. It presents the analysis of tests performed.

Chapter 6 presents future scope and conclusions based on the contribution made by this thesis.

Appendix A contains the code snippets and **Appendix B** contains the snapshots of the system and **Appendix C** contains the list of publications.

1.8 Chapter Summary

This chapter presents the idea used in this thesis. It discusses research problem, objectives, goals and motivation for the research. Justification for the research problem is outlined, together with an explanation of the research methodology used. The next chapter describes the literature survey and relevant background work done till date in context of this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1. ATS using Fuzzy Logic

Fuzzy logic model intends to offer linguistic representation for handling uncertainties. It is an approach based on the concept of evaluating the degree of truth refrerred to as truthness rather than the simple 0 or 1, or true or false logic. Research on fuzzy logic started in 1965 by Zadeh[11]and since the inception it has been widely accepted and used in various application domains owing to the underlying primary notion which replicates a typical human inference process. All computational data cannot necessarily be expressed in the terms of binary values, for example, a student's Intelligence Quotient (IQ) cannot be defined in the terms of 0 or 1, that is if we are measuring the IQ on basis of aggregate scored in a course, we cannot label all students with aggregate 70-95% into a single class. The use of boolean logic in such situations will not suffice and hence Fuzzy logic models which resemble the human reasoning system were proposed. These allowed a linguistic representation of data with values assigned not only just 0 or 1 but also within 0 and 1, that is, permitting fuzzy values having partial set relationship rather than crisp values. The linguistic representation included using IF-THEN rules and plotting the input/output membership functions for processing data with multiple input values and provide output as a single variable. Conventional machine learning techniques have been used for text summarization but using fuzzy logic handles uncertainties in the input better than other models, and no another method performs better in computing with words [11] and thus is preferably used for linguistic summarization [12].

Fuzzy logic based model for automatic text summarization has been studied across literature [22,25, 30, 32].A typical fuzzy logic based model for ATS takes as input eight features for each sentence (Title word, Sentence length, Sentence position, Numerical data, Thematic words, Sentence to sentence similarity, Term weight, Proper noun) to calculate its importance. Once the value of these eight features has been extracted, it is passed to a Fuzzy Inference System(FIS). FIS basically comprises of three steps, namely, Fuzzification, Inference Logic, Defuzzification. The output of FIS is an importance

score for each sentence and the sentences are arranged in descending order. Also, research has substantiated that a summary length is nearly 10% of the actual document length and the resultant summary consists of sentences extracted with the original order maintained. Figure 3 depicts this architecture.

The steps in the Fuzzy Inference Systems are explained in the following sub-section:

- **Fuzzification :-** In this step the crisp values are converted into fuzzy value using membership function. Various types of membership functions like triangular, trapezoidal, ball, Gaussian distribution function, are available for mapping. For example, if a trapezoidal function is used, then each inputs' membership degree into a fuzzy set usually having three values low, medium or high. The generalized trapezoidal membership function depends upon four parameters p, q, r and s as given by the following equation [5].

$$f(x,p,q,r,s) = \max(\min\left(\frac{\mathrm{x}-\mathrm{p}}{q-p}, 1, \frac{\mathrm{s}-\mathrm{x}}{s-r}\right), 0)$$

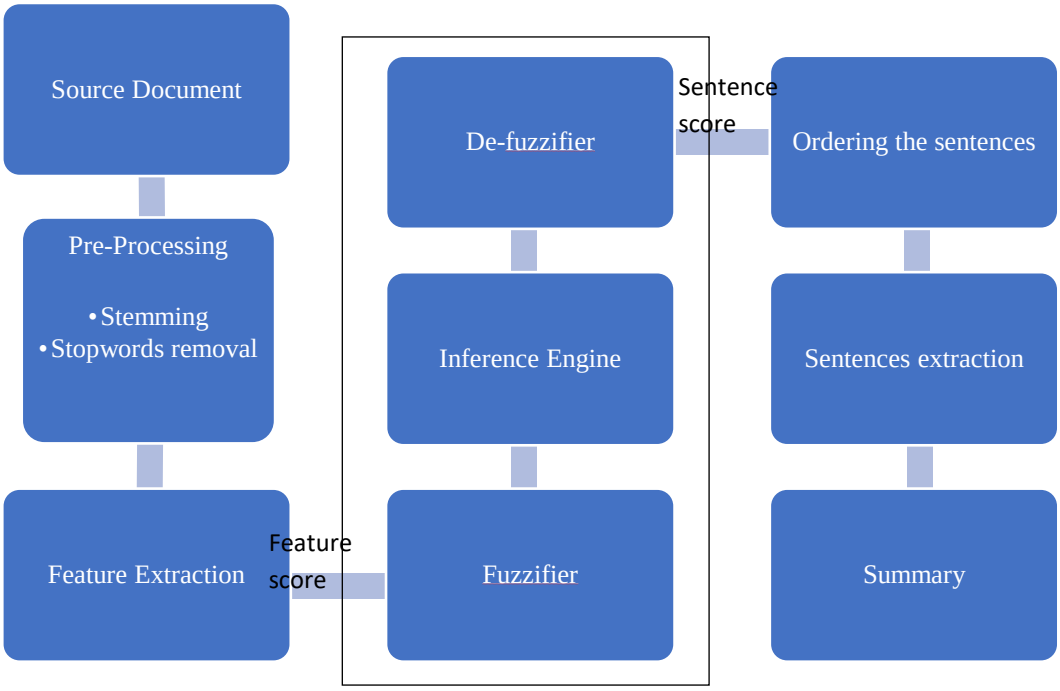

Fig 3 Text Summarization using Fuzzy Logic.

Where p and s represent the "feet" of the trapezoid and q and r represent the "shoulders" [11]. Membership function of one of the input variable title similarity is shown in figure 4.

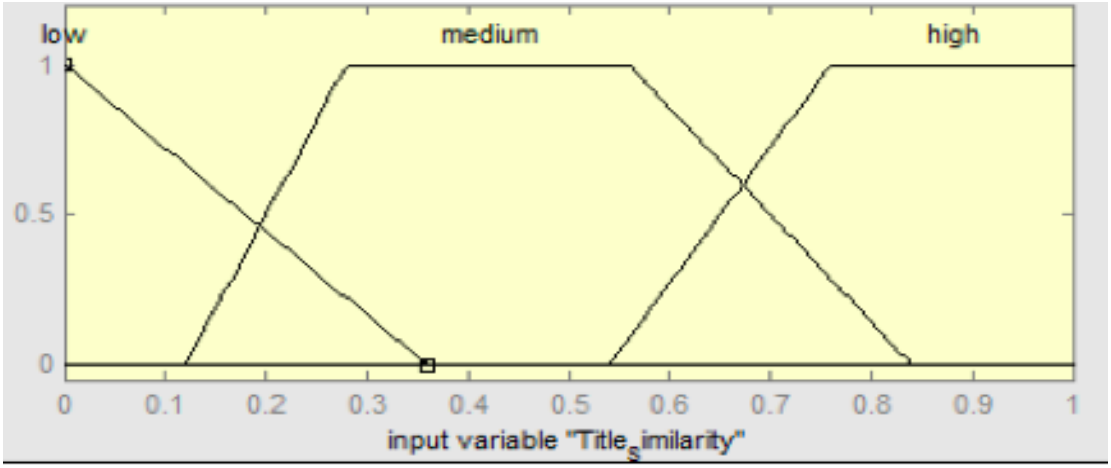

Fig 4 Membership function of an input variable

- **Inference Logic :-** A knowledge base is created with IF-THEN rules and the inference engine derives the output based on these rules taking the input value generated in the first step. IF-THEN rules are used to balance the weights of key and non-key factors.An IF-THEN rule is stated in following format:

If (title similarity is medium) and (sentence length is medium) and (sentence location is medium) and (numerical data is low or medium or high) and (sentence centrality is low) **then** (output is key)

- **Defuzzification : -** In this final step, the results generated in second step are mapped to crisp values using membership function, i.e. it converts the linguistic result from inference engine into a numeric value. The output membership function could be taken same trapezoidal or any other depending on the situation, in this case, taking the trapezoidal method the output membership function is divided into three fuzzy sets: key, partial-key and non-key, in some cases it is taken as 5 value set too like very low, low, normal, high, very high. The centroid

method is used to find the crisp value. The output trapezoidal membership function of a three value fuzzy set is shown in fig.5.

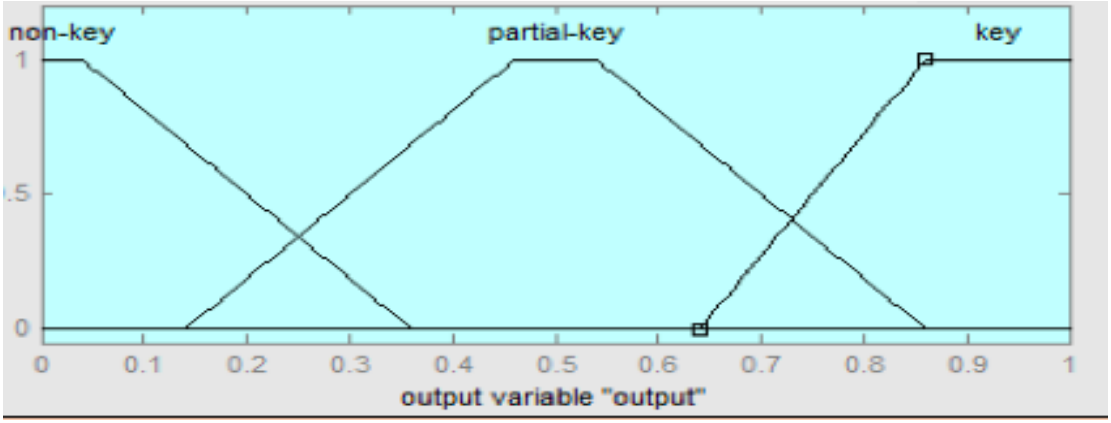

Fig. 5 Membership function of an output variable

2.2. ATS using TextRank

In their paper "TextRank: Bringing Order into Texts", Mihalcea and Tarau [54] introduced TextRank as the first graph-based automated text summarization algorithm.. It belongs to the category of extractive summarization techniques which constructs the summary by extracting the most important sentences from the original text. TextRank algorithm is a simple application of PageRank algorithm. The graph is built using natural language processing which establishes the relationship between the entities of the text. A vertex gains more importance if it forms higher number of links with other vertices because when one vertex links with another, it casts a vote in favour of that vertex. Thus the score of the vertex is calculated by considering the inbounds and outbounds of that vertex. The graph based ranking of the vertices in the graph can be determined by evaluating the associated score of each vertex.

The score of a vertex $\boldsymbol{V}_i$ is defined by the following equation:

$$S(V_i) = (1-d) + d \cdot \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \qquad \ldots(1)$$

Where, G = (V,E) represents a directed graph with the set of vertices V and set of edges E. ***In*** ($\boldsymbol{V}_i$) and ***Out*** ($\boldsymbol{V}_i$) represents the in-bounds and out-bounds for a vertex $\boldsymbol{V}_i$ and $\boldsymbol{d} \in [0, 1]$ is a damping factor in which ***d*** is a probability the node visits neighboring node and (1 − ***d***) is the probability of jumping from one vertex to some random vertex. It is generally assigned as 0.85.

However, we can't measure the strength of the connection between two vertices. Thus, a new formula is introduced which incorporates the weight of the edges while computing the score of vertex and is given as:

$$WS(V_i) = (1-d) + d \cdot \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad ..(2)$$

where, $WS(V_i)$ is the weighted score for vertex V_i and w_{ij} is the weight which represents the strength of the connection between two vertices V_i and V_j .

Figure. 6 below shows a sample graph for an abstract from our test set. The average size of the abstract is about 120 words. Figure represents a small abstract for the purpose of illustration. For example, the lexical units that are found to be of higher importance along with the TextRank score can be seen as: numbers (1.46), inequations(1.45), linear (1.29), diophantine (1.28), upper (0.99), bounds (0.99), strict (0.77). Ranking is different than the one obtained by simple word frequencies. For a same text, frequency method gives following top-ranked lexical units: system (4), types (3), solutions (3), minimal (3), linear (2).

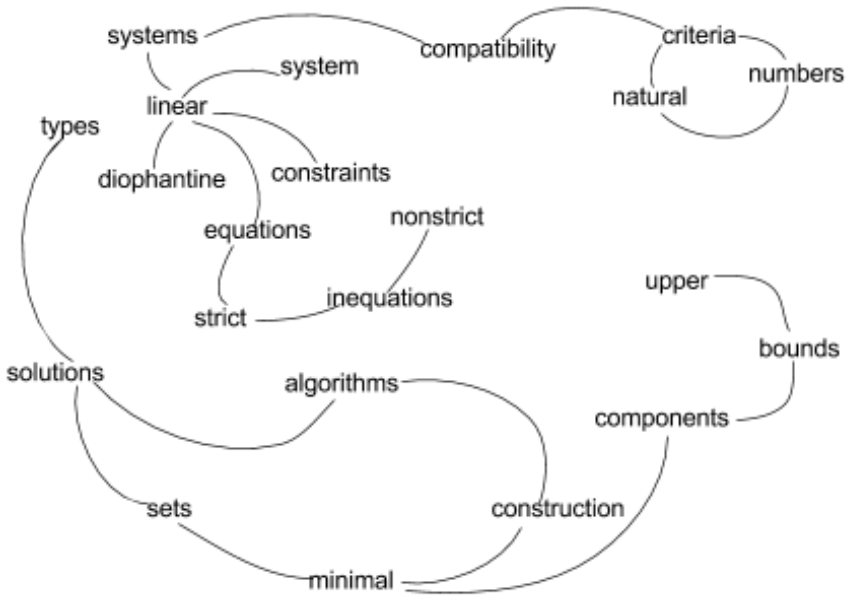

Fig. 6 Small abstract for keyphrase extraction for the purpose of illustration

To enable the application of graph-based ranking algorithms to natural language texts, we have to build a graph that represents the text, and interconnects words or other text entities with meaningful relations. Depending on the application at hand, text units of various sizes and characteristics have to be added as vertices in the graph, e.g. words, collocations, entire sentences, or others. Similarly , it is

the application that dictates the type of relations that are used to draw connections between any two such vertices, e.g. lexical or semantic relations, contextual overlap, etc.

In case of sentence keyword extraction, the very first step is to extract all the sentences from the text. Once we have all the sentences extracted, we build a graph where the sentences are represented by nodes and the edges denote the similarity between the sentences. The number of common tokens or the lexical density such as nouns and verbs which are keywords to the text can be used to determine the overlapping or similarity between the sentences. For two given sentences Si and Sj, where each sentence Si can be represented by the words, the similarity function is given as follows:

$$\text{Similarity (Si, Sj)} = \frac{|\{w_k \mid w_k \in S_i \,\&\, w_k \in S_j\}|}{\log(|S_i|)+\log(|S_j|)} \qquad \ldots(3)$$

We then take the **n** sentences having highest scores in the order as they appear in the text based on their similarity score to generate the extractive summary. TextRank algorithm is fully unsupervised algorithm which didn't attempt to learn by training on the set of summaries but rather relies on the text to derive extractive summary. The ranking is different than the one obtained by simple word frequencies as we first compute the similarity score for each vertex. Also, this algorithm is purely dependent on the word concurrence thus the unreliability on the knowledge of grammar makes TextRank the language independent algorithm.

2.3. ATS using LexRank

LexRank is another graph-based algorithm for automated text summarization, introduced by Güneş Erkan and Dragomir R. Radev [55]. LexRank is a cosine transform based weighting algorithm. This approach is different from TextRank as it determines the significance based on the centrality of a sentence in a graph based representation of the document. A cluster of documents is viewed as a network of sentences that are related to each other. The sentences that are similar to many of the other sentences in a cluster are considered more central or salient to the topic. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences [56]. This give the important information related to the central idea of the article. Thus, the sentences are ranked based on their overall centrality. Firstly, a graph representation composed of all the sentences where each sentence represents a node and edges represent the similarity connection between these sentences is constructed. We measure the similarity between sentences using bag-of-words model by computing the frequency of word occurrence in a sentence. Following algorithm is used for computing the centroid scores.

```
   input  : An array S of n sentences, cosine threshold t
   output: An array C of Centroid scores
 1 Hash WordHash;
 2 Array C;
 3 /* compute tf×idf scores for each word */
 4 for i ← 1 to n do
 5     foreach word w of S[i] do
 6         WordHash{w}{"tfidf"} = WordHash{w}{"tfidf"} + idf{w};
 7     end
 8 end
 9 /* construct the centroid of the cluster */
10 /* by taking the words that are above the threshold*/
11 foreach word w of WordHash do
12     if WordHash{w}{"tfidf"} > t then
13         WordHash{w}{"centroid"} = WordHash{w}{"tfidf"};
14     end
15     else
16         WordHash{w}{"centroid"} = 0;
17     end
18 end
19 /* compute the score for each sentence */
20 for i ← 1 to n do
21     C[i] = 0;
22     foreach word w of S[i] do
23         C[i] = C[i] + WordHash{w}{"centroid"};
24     end
25 end
26 return C;
```

The basic evaluation is done using TF-IDF formulation where TF is the term frequency which computes the similarity strength as the number of occurrences of the words and IDF is the inverse document frequency in which low frequency words inversely contribute to the similarity factor. This TF-IDF formulation is used to evaluate the similarity between the sentences using idf-modified-cosine formula as shown in equation 4.

Then we composed a similarity matrix using similarity measure evaluated before. The importance of sentences also incorporates its relative importance to the neighbouring sentences.

$$\textit{idf-cosine (x,y)} = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{(x_i \in x)} \left(tf_{x_i,x} idf_{x_i}\right)^2} \sqrt{\sum_{(y_i \in y)} (tf_{y_i,y} idf_{y_i})^2}} \quad ..(4)$$

Positive contribution will raise the significance of neighbouring sentence whereas negative contribution will lower its significance. The centrality of any node *u* is given by:

$$p(u) = \sum_{v \in adj[u]} \frac{p(v)}{\deg(v)} \quad \text{...(5)}$$

where, $p(u)$ is the centrality of node u, $adj[u]$ is the set of nodes that are adjacent to u, and $deg[v]$ is the degree of the node v.

Having computed the centrality and the degrees of similarities, the LexRank algorithm extracts key nodes from the graphs based on their weights, and thus prepares an extractive keyword summary of the given text(s).

2.4. ATS using Latent Semantic Analysis (LSA)

Liu et al introduced the idea of using Latent Semantic Analysis in ATS in 2002 [56]. Taking the ideas from the latent semantic indexing, they used the singular value decomposition (SVD) to text summarization domain. SVD is a very powerful mathematical tool that can find principal orthogonal dimensions of multidimensional data. It has applications in many areas and is known by different names: Karhunen-Loeve Transform in image processing, Principal Component Analysis (PCA) in signal processes and Latent Semantic Analysis (LSA) in text processing.

As described by Liu et al in their paper, the process starts with the generation of a terms by sentences matrix A = [A_1, A_2, A_3, … A_n], where column vector A_i represents the weighted term-frequency vector of sentence i in the document [56]. In total of ***n*** sentences having ***m*** terms in the document, a ***m x n*** matrix ***A*** will be generated***.*** As not every single word appears in each sentence so, the matrix generated will not be dense.

The SVD of an ***m x n*** matrix A, where ***m*** ≥ ***n*** is defined as:

$$A = U\Sigma V^{\mathrm{T}} \quad (6)$$

Where **U** = [u_{ij}] is an orthonormal matrix around columns which are also called as left singular vector Σ = diag (σ_1, σ_2, …….. σ_n) is n x n diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order **V** = [v_{ij}] is ***n x n*** orthonormal matrix, whose columns are called right singular vectors. If rank(**A**) = r, then Σ satisfies:

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma r > \sigma_{r+1} = \ldots = \sigma n = 0 \quad (7)$$

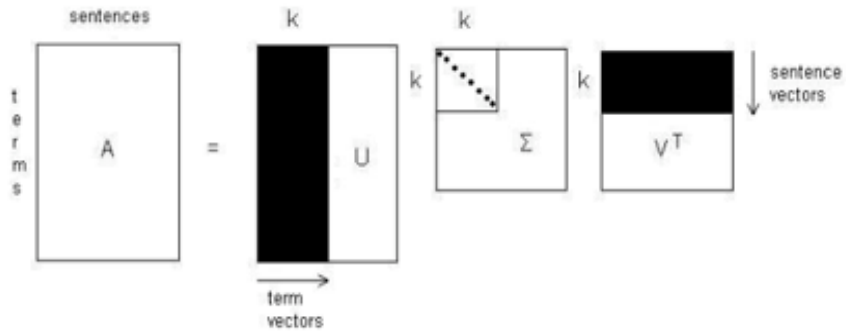

Fig.7. Singular Vector Decomposition

The SVD leads to a mapping between weighted term-frequency vectors and singular vectors on a ***m x n*** metrix, where all of its axes are linearly-independent. This mapping essentially projects each column vector ***i*** in matrix **A** to column vector $\psi_i = [\upsilon_{i1}\upsilon_{i2} \ldots \upsilon_{ir}]$ of matrix $\boldsymbol{V^T}$ and maps each row vector ***j*** in matrix **A**. Each column vector ***i*** in matrix **A** represents the weighted term-frequency vector of ***sentence i*** and row vector ***j*** in matrix **A** tells the occurrence count of the term ***j*** in each of the documents, to row vector $\phi_j = [\boldsymbol{u_{j1}u_{j2}} \ldots \boldsymbol{u_{jr}}]$ of matrix **U**. The SVD-based document summarization algorithm is as follows:

Step 1. *Deconstruct the document D into single sentences, and use these sentences to form the candidate sentence set* ***S****, and set* $\boldsymbol{k} = 1$.

Step 2. *Construct the terms by sentences matrix* ***A*** *for the document* ***D***.

Step 3. *Apply the SVD on matrix* ***A*** *to obtain the singular value matrix* $\boldsymbol{\Sigma}$ *and the right singular vector matrix* $\boldsymbol{V^T}$.

Step 4. *Select the* $\boldsymbol{k^{th}}$ *right singular vector from matrix* $\boldsymbol{V^T}$.

Step 5. *Select the sentence which has the largest index value with the* $\boldsymbol{k^{th}}$ *right singular vector, and include it in the summary.*

Step 6. *If* ***k*** *reaches the predefined number, terminate the operation; otherwise, increment* ***k*** *by one, and go to Step 4.*

Since all the singular vectors are independent of each other therefore, the sentences selected by this method contain the minimum redundancy.

2.5. Related Work

Now a brief introduction about all the related studies has been given in reverse chronological order. Table 2 contains the summary about the applicability of various techniques proposed in these studies on varied data sets, motivation for proposing those technique, their limitations and accuracy of the proposed method.

The most recent model for generating text summarization using fuzzy logic (FL) is proposed by Kumar et al [14] in 2017, where they use an Adaptive Neuro-Fuzzy Inference System (ANFIS) to classify sentences for generating high quality. ANFIS is a hybrid model that relates the knowledge reasoning of fuzzy logic and the learning nature of neural network (NN) resulting in improved summary. Use of NN with FL eliminated the need of human expert in defining fuzzy rules and shaping of membership function. This model showed the remarkable results, but the results can't be compared with other methods as they didn't evaluate it using the standard evaluation metric i.e. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [46].

In 2016 three models based on FL were proposed which majorly focussed on generating summary containing only prime sentences. First was given by Razieh et al [15]. They had developed a hybrid model taking the merits of FL, Particle Swarm optimization (PSO), Genetic Algorithm (GA), Artificial Bee Colony (ABC) and Cellular Learning Automata (CLA), known as FPGAC for generating the sentences with high diversity. In FPGAC, CLA n-grams joints were utilized for extracting sentences whose neighbors were extracted using ABC. FPGAC outperformed other methods [15] as it also optimizes the similarity measure for extraction process. Another model was put forward by Yadav and Meena [17] in 2016. They had used FL to handle the issues of ambiguity and imprecise values with the traditional models, WordNet to reflect the semantics of the text in the method, and lastly the Bushy Path, graph based text summarization approach is used to identify the similar sentences. The sentences that were common in all the produced summaries were found to be more influential and hence they ought to be included in the final summary. Third model proposed by Jafari et al in 2016 where they considered both the syntactic parameter and semantic relations of the expressions to achieve the finest summary, with the belief that attention to concept and meaning of the words could result in better summary [18].

Farshad Kiyoumarsi [19] (2015), had compared human-generated summaries with automated summary by use of FL. The three summaries were evaluated using both the ROUGE evaluation system, by human judges and few of the English professors. By all means, summaries produced by FL had a much closer score for human-generated summaries in comparison to vector method, which shows that

automated summary using FL is a good replacement of human generated summary, thus, making the process of summarization easy and quick, especially for large documents.

Another conceptual model was put forward by Babar and Patil [20] where a hybrid of FL and Latent Semantic Analysis (LSA) was proposed to achieve intelligent summary. LSA was used to correlate the semantic relations of different contexts of the text and to capture more meaningful sentences in the summary excluding the words with multiple meanings [21]. To balance out the limitations of LSA, FL was used, finally providing understandable summaries. Patil, Mane [21] had also proposed a similar hybrid of LSA and FL for multi-document summarization, using Agglomerative K-means algorithm with concept analysis for handling multi-documents. In this model, clusters are formed with K-means, and each cluster is named by calculating their Term Frequency (TF) weight using word cloud. This TF value of each cluster is given as input to LSA.

Razieh et al in 2014 had proposed two methods, one using Cellular Learning Automata (CLA) and the other using CLA with PSO and FL. The former was used to calculate similarity of the sentences [22]. They concentrated on reducing the redundancy of the summary for yielding improved summaries. Later method was used for calculating the sentence scoring [23].

Pallavi and Kulkarni [24] proposed a technique utilizing FL on eight features using triangular membership function to calculate the scores of the sentence. A similar model using the same eight features was proffered by Shinde et al [25], they had tested the model on few random datasets and compared it with Ms Word on the basis of number of lines in the summary generated.

Y. J. Kumar et al [26] had discussed about the applicability of two techniques for generating efficient summary of news articles belonging to similar domain. They had presented a technique for identifying the cross-document relations for classifying the important sentences in multiple documents regarding same news using case based reasoning (CBR). This technique was termed as Generic-CBR and it yielded better results when tested on CSTBank dataset which contained English news articles, as compared to other classifier models like SVM, NN and simple CBR.

In 2012 Ladekar et al [28] had proposed a novel ATS technique, which used evolutionary algorithms like GA and genetic programming (GP) to optimize the result of fuzzy inference system (FIS). FIS was given as input the features of a sentence stating the number of title words, thematic words emphasize words, and total words in the sentence, the position of sentence, whether first or last in the paragraph. Results depicted that GA had improved the membership function value.

Apte and Dixit [29] also proposed a similar FL based text summarization in 2012, taking the eight feature extraction values of each sentence as the input to FIS. Triangular membership function with five fuzzy sets was used. And for calculating the output value, simple fuzzy centroid method was used.

In 2011, a hybrid technique using GA, FL and Semantic Role Labelling (SRL) was proposed by Suanmali et al [30] for generating automated summary. GA optimized the process of feature selection during feature extraction step and also assigned the weight to each feature during training phase. FL was used to calculate the scores of sentences on the basis of the feature values and their weights were obtained from GA. For embracing the semantic relations, semantic role labelling was used along with GA and FL. It calculated the relevance score of sentences using similarity measure. The score of each sentence was obtained from both FL and SRL and was summed up to find the final score for each sentence. The sentences with the highest scores were selected to be in the part of the summary.

In 2010, M. S. Binwahlan et al [31] presented two hybrid techniques. They focused on removing the redundancy by using a diversity based Maximal Marginal Intelligence (MMI) technique to select the unique sentences only. Not every attribute had the same importance in a selection procedure, so to handle this issue they had used a swarm based Particle Swarm Optimization (PSO) technique to assign the appropriate weights to each feature according to their relevance in the selection procedure. When human generate summary from a document, it may not completely match with the other human generated summary of the same document, as human think fuzzy. So to handle such imprecision in the training data they chose FL for text summarization. Each technique had been assigned different weight according to their key features. Both the methods are hybrid of three techniques, and are generating the summary from the highest of score from all the three techniques. The first model proposed dominates the diversity technique, second takes all the three techniques equally. The first model uses MMI, Swarm-MMI, and the Fuzzy Swarm method. The second model uses Fuzzy Swarm in place of swarm MMI, and Swarm based instead of Fuzzy swarm based method. When both the models were tested over DUC 2002, later performed better than the former, but still the author favours the first model as it handles the redundant data more precisely [31,35,37].

In 2009 Hsun-Hui Hunag et al [33] had proposed a combination of FL and rough sets to generate effective summaries for multi-document. They had used some sentence level and concept level features. Fuzzy set, and Rough seta were used to balance the weights of these features. They had tested the method on DUC 2006 along with 35 other models.

Suanmali et al (2009) had presented four studies on the use of FL in text summarization [34, 35, 36, 37]. In their first study [34] they had used the standard eight features, passed their value to FIS and

had tested it on a 30 document dataset. In [35] the similar technique were studied and tested on DUC 2002. In the third study [36], for improving the efficiency of the model they had considered one extra feature and had also replaced the triangular membership function with the Gaussian membership function. The results obtained were almost similar. The third study [37], was based on the fact that not every feature is of same importance, thus, they had used PSO, a nature inspired optimization technique for determining the importance of each feature. A similar study was also given by them in 2008 in which they used the bell membership function [38].

In 2008, Kyoomarsi et al [39] had proposed a model based on fuzzy oriented approach where they had passed the statistically calculated values to the Fuzzy toolkit of MATLAB. They had also compared the results with vector approach. They had tested the tool on 10 documents and gave the generated summaries to 5 humans with English major to evaluate the accuracy of the summary. The generated summaries got an average accuracy of 77%.

One of the earliest methods on FL based text summarization was put forward by Arman Kiani-B et al in 2006 [40]. They had used six basic statistics values as input features for fuzzy. They had used these features in triangular membership function of FIS. The result was then tested only on three documents.

In 2006, Huang et al [41] had discussed about the usage of semantic logic based technique for reducing the redundant sentences in the document. Statistical techniques can efficiently stem the word during pre-processing, but cannot find the similarity in case where synonyms are used. So they had used clustering techniques where they had clubbed the similar sentences, and from the dissimilar sentences, features values are calculated and adjusted using fuzzy sets and rough sets, and the resulted values are passed on to the FIS with triangular membership function to calculate the importance score of each sentence.

In [42], a novel approach for generating intelligent summaries was put forward by Arman- Kiani –B and M.R. Akbarzadeh-T which has been further used by [27]. A similar technique using FL was also proposed by them in 2006 [40]. In [42], the authors worked up for improving the results by focussing on the two main units which had the highest influence on calculating the final score. They had emphasized on increasing the efficiency of the machine, by optimizing the rule set and the membership function of FIS. GA was used for improving the membership function and GP was used for improving the rule set. Although they had tested it only on 3 articles, but their results looked promising.

In [43], author had presented a fuzzy based method for text summarization especially for news summarization. Their main focus was on news articles, so they had tried to reduce the length of the summary approximately to 9% compression rate. To improve the results they had proposed a new Fuzzy inference system with seven layers. This method seemed to produce good results, but still no further research has been conducted with this type of fuzzy machine.

First study showing the use of FL in text summarization was given by Witte and Bergler [44]. FL was used for solving the uncertainty issue in co-reference based text. Although it is not a good measure to be used for text summarization, but if used with other methods it could have given better results.

The following table 1 contains the state-of-art of this fuzzy logic based ATS where all the primary studies are briefly reviewed based on year, author, dataset used and evaluation metrics and motivation of their work, publication year, details and scope of the work. Motivation discusses the reason why the authors have proposed this method, and what were the limitation of the existing techniques that lead to this work. Details and scope section contains the basic overview of the technique and its pros and cons of the proposed technique.

Sr. No.	Year	Author	Technique	Data set	Accuracy	Motivation	Details & Scope
S-1	2017	Y. J. Kumar, F. J. Kang, O. S. Goh, A. Khan	Adaptive Neuro-Fuzzy Inference System	DUC 2002	Precision :- 71.28 Recall :- 69.82 F-measure :- 70.54	•Utilizing the benefits of both fuzzy logic and neural network, as fuzzy logic is knowledge-driven, whereas neural network is data-driven. •Need of human experts for implementing fuzzy logic techniques for determination of rules and tuning of membership functions. •Enhanced fuzzy system can be generated by incorporating the learning and adaptive capabilities of neural network.	• Performs better than Fuzzy logic and neural network. • Evaluation should be done using ROUGE. • LM backpropagation method is used in combination with Least-square Estimate model to estimate the parameters of membership function.
S-2	2016	R. Abbasi-ghalehtaki, H. Khotanlou, M. Esmaeilpour	Artificial Bee Colony (ABC), Cellular Automata (CA), Particle Swarm Optimization (PSO), Genetic	DUC 2002	ROUGE-1: 0.48685, ROUGE-2 : 0.22910	•Better Similarity measure was required. •A hybrid technique to perform best at every situation to get accurate summary.	• A hybrid technique which outperform other similar summarization techniques. • Uses combination of CA and ABC for finding best diversity sentences. • PSO and GA are used for assigning weights to features extracted.

			Algorithm (GA), Fuzzy Logic (FL).				• FL is used to score the Sentences.
S-3	2016	Jyoti Yadav, Dr. Yogesh Kumar Meena	Fuzzy Logic, Bushy Path, WordNet Synonyms	DUC 2002	ROUGE-1 (Recall) : 0.46824, ROUGE-1 (precision): 0.43411 ROUGE-1 (F-measure): 0.44829 ROUGE-2 (Recall) : 0.24494, ROUGE-2 (precision): 0.22553 ROUGE-2 (F-measure): 0.23365	• To handle the ambiguity and imprecise values using Fuzzy Logic. • To consider the semantics of the text using WordNet.	• Summary from three techniques are generated simultaneously. • The sentences which come in all three are considered most important and then put into the summary, after that according to the length of the summary defined, the top scorer from all the three summaries are selected. • provides better result than individually used • Should be tested for long data set and for multi-document.
S-4	2016	Mehdi Jafari, A. M. Shabhab	Fuzzy Logic using both syntactic and	50 random articles (average number	Fitness :- 0.59, Precision :- 0.6,	• Only syntactic and semantic parameters have been used.	• A technique taking into consideration both the syntactic and semantic parameters

		i, J. Wang, Y. Qin, X. Tao, M. Gheisari	semantic parameters	of words per document 436)	Recall :- 0.58	• Semantic relation between the words has not been used.	for achieving high quality summary. • Compared with MS word, Copernic, and Huang, the purposed method performed better than others. • Not tested on the standard dataset, so cannot generalize the result.
S-5	2015	Farshad Kiyoumarsi	Fuzzy Logic, Vector approach	DUC 2004 (100 Documents)	ROUGE 1 :-29.6, ROUGE 2:- 7.8, ROUGE 3:- 2.6, ROUGE 4 :- 0.9, ROUGE L:- 25.3, ROUGE W-1/2 :- 18.9 (10% four human written summaries)	• To compare human generated summaries with automatic summary to compare the result. • Taking rhetorical features into account while using fuzzy logic	• Analysed human summaries to understand why they perform better than automatic summaries. • The cue features are taken in account also at paragraph and essay level not just on sentence level in fuzzy method. • Shown that human summaries are more accurate than automated summaries. • Fuzzy performs better than vector method.
S-6	2015	S. A. Babar, Pallavi D. Patil	Fuzzy Logic, Latent Semantic Analysis (LSA)	10 datasets (small)	Precision :- 90.77572, Recall :- 44.36375, Fitness :- 67.56974	•Fuzzy do not take into account the semantic relations between concepts. •LSA does not give same importance to every attribute. •Drawbacks of LSA and Fuzzy overcome with their combination.	•A hybrid technique of LSA and Fuzzy logic. •LSA is used for handling semantic relations of the text. •Fuzzy logic with eight features is used for improving the summary. •Result of both the techniques is combined to achieve more accurate

							summary using set operations. • To be certain of the accuracy, system should be tested on large dataset.
S-7	2015	Pallavi D. Patil, P. M. Mane	Fuzzy Logic, Latent Semantic Analysis (LSA), Agglomerative K-means	Random Dataset	Precision :- 89, Recall :- 43.6, Fitness :- 66.3	• To be used for multi-documents.	• Uses LSA and Fuzzy logic for single document summarization. • For multi-document, agglomerative K-means algorithm is used. • The term frequency of each cluster is assigned as the name of the cluster an given as input to LSA. • Overall Fitness measure of the proposed system results in slightly lesser value than LSA summary.
S-8	2014	S. A. Babar, S. A. Thorat	Fuzzy Logic, Latent Semantic Analysis (LSA)	5 datasets with different length of summary.	Average Accuracy :- 85.332, Time complexity ranges from 80msec to 94msec for summary of different length.	• To improve accuracy of summary. • To take semantic parameters into consideration.	• Hybrid of Fuzzy logic and LSA. • Tested on 5 different datasets and compared with gold standard human generated summary. • Proposed method performs better than only fuzzy-based summarization.

S-9	2014	R. Abbasi-ghalehtaki, H. Khotanlou, M. Esmaeilpour	Fuzzy Logic, Cellular Learning Automata (CLA), Particle Swarm Optimization (PSO)	DUC 2002 (100 documents)	ROUGE 1(avg-F) :- 0.46622, ROUGE 2(avg F) :- 0.2075, ROUGE L (avg F) :- 0.43001	• Similarity measure needed to be improved. • Some features are more important than other, so to assign appropriate weights.	• Two techniques were proposed, one using CLA for calculating similarity of sentences and the calculating the score of the sentences on the basis of statistical features. • Second is a hybrid technique of CLA, PSO and fuzzy. • CLA is used in feature extraction for calculating similarity. PSO is used for assigning weights to the features, and fuzzy is used for calculating the sentence score. • Second method performs better than other compared methods except for H2-H1 method.
S-10	2014	Pallavi D. Patil, N. J. Kulkarni	Fuzzy Logic		----------	• To develop a computationally efficient tool.	• Uses Fuzzy method to generate Scores of sentences. • Eight types of features are extracted. • Triangular membership function is used. • Method was not tested on any dataset.
S-11	2014	R. J. Shinde, S. H. Routela, S. S. Jadhav,	Fuzzy Logic	Random dataset	Number of sentences in the summary (compare	• To use fuzzy logic in ATS.	• Similar to the method proposed in [9]. • Shows better result than Ms Word. • Method was not tested compared to

		S. R. Sagare			d with Ms Word)		human generated summaries.
S-12	2014	Y. J. Kumar, N. Salim, A. Abuobieda, A. T. Albaham	Generic-Case Based Reasoning (CBR), Fuzzy Logic	DUC 2002	Classifier accuracy F-measure:-84.47%, ROUGE 1:-0.335, ROUGE 2:- 0.128, ROUGE S:-0.096, ROUGE SU :-0.1009	• To produce high quality multi-document news summary. • To find intelligent summary cross-document relations should be studied.	• Presented two techniques. • Generic CBR, for finding cross-document relations from un-annotated text. • A fuzzy model for ATS using Generic CBR. • Both the methods performed better than other similar techniques when compared.
S-13	2013	A. R. Kulkarni, S. S. Apte	Fuzzy Logic	2 sports news articles	Average fitness=0.71	• To use the idea of both statistical and linguistic methods.	• Uses the same technique as Patil. • Not tested on enough dataset to be said as good results.
S-14	2012	A. Ladekar, A. Mujumdar, P. Nipane, S. Tomar, Kavitha S.	Fuzzy Logic, Genetic Algorithm (GA), Genetic Programming (GP)	---------	----------	• Value of membership function and the rule base of fuzzy system should be updated according to the input text.	• Proposes an optimized membership function and rule based fuzzy system with GA and GP. • Fuzzy logic is given unstructured features of the text as input. • The Membership function is optimized using GA. • Rule sets are optimized using GP. • The method is not evaluated on any dataset.

S-15	2012	R. S. Dixit, S.S. Apte	Fuzzy Logic	30 documents from news based URL	Accuracy :- 81%, Position similarity of sentences in generated and human summary :- 79%	• Fuzzy logic to remove the uncertainty of co-reference of noun-phrases. • To improve accuracy of text summarization.	• Uses Fuzzy logic for calculating relevance score of each sentence on the basis of the values of eight features. • Provides more intelligent summaries as compared to Ms Word and Copernic summarizer. • Need to test on standard dataset of multiple domains.
S-16	2011	L. Suanmali, N. Salim, M. S. Binwahlan	Fuzzy Logic, Genetic Algorithm (GA), Semantic Role Labelling (SLR)	DUC 2002 (100 documents)	ROUGE 1 (Average Precision) :- 49.95%, ROUGE 1 (Recall) :- 45.19%, ROUGE 1 (Fitness) :- 47.04%	• Fuzzy and GA cannot take into account semantic relations between concepts of text. • SLR is good in finding semantic relations of the document, but cannot capture the main text of the document.	• Proposes a hybrid model for ATS using Fuzzy logic, GA, and SLR. • GA is used for optimizing feature selection process and for also calculating the weights of each feature during training time. • Fuzzy logic is used to handle uncertainties in the data and balance thee weight between more relevant and less relevant features. • SLR captures the semantic data in the text and includes those in the summary. • The score of each sentence by both fuzzy logic and SLR is added up to find the final score, on basis of which the sentences are

							extracted for automated summary.
S-17	2010	M. S. Binwahlan, N. Salim, L. Suanmali	Maximal Marginal Importance, Particle Swarm Optimization, Fuzzy Logic	DUC 2002 (100 Documents)	For first model: ROUGE 1 (F) :- 44.94% ROUGE 2 (F) :- 20.07% ROUGE-L (F):- 41.38% For second method ROUGE 1 (F) :- 45.87% ROUGE 2 (F) :- 20.42% ROUGE-L (F):- 42.29%	• Redundancy should be handled more precisely. • Issue of inconsistency of the training data because of fuzzy and low agreement of human should be handled with fuzzy logic. • Every attribute does not have the same importance, so they should be weighted according to their relevance.	• Two models were proposed, one dominating the diversity measure, second not. • In first the scores are calculated by three different methods: diversity, swarm diversity, and fuzzy swarm based. Different weights are assigned to all the three components to find the appropriate sentences from all the components. • Second also takes three components, but in place of swarm diversity, it uses fuzzy swarm base instead of swarm diversity, and third component is replaced by swarm based. • While the first is handling redundancy better than second, but the overall more accurate summary is being generated by the second method.
S-18	2010	S. Alzahrami, N. Salim, C.K. Kent	Fuzzy Swarm based Technique	-------	--------	• To utilize the text summarizations' benefit in plagiarism detection.	• Text summarization solves multiple issues, it can handle cross language documents very easily. • The redundancy in data can also be

							found with the use of TS. • Improves the plagiarism detection process.
S-19	2009	Hsun-Hui Hunag, Horng-Chang Yang, Yau-Hwang kuo	Fuzzy logic, Rough Set	DUC 2006	ROUGE-1 :0.40636, ROUGE 2:- 0.08245	• To face synonymous and ploysemous problems of textual terms.	• Proposed a hybrid approach of rough set and fuzzy logic for multi-document summarization. • Tested with other 35 techniques. • Performance comes in first one third techniques.
S-20	2009	L. Suanmali , N. Salim, M. S. Binwahlan,	Fuzzy Logic	30 documents	F-value:- 0.47873	• To use human reasoning system.	• For both single and multi-document summary generation.
S-21	2009	L. Suanmali , N. Salim, M. S. Binwahlan,	Fuzzy Logic	DUC 2002 (125 Documents)	F-value:- 0.47181	• To standardize the result by testing it on DUC.	• Same technique as above. • Use triangular membership function. • Tested on DUC to make results trustworthy.
S-22	2009	L. Suanmali , N. Salim,	Fuzzy Logic	DUC 2002 (30 Documents)	F-value :- 0.47019	• To improve the above model.	• 9 Features were selected instead of 8 in the above model. • Gaussian Membership function has been used.

		M. S. Binwahlan,					• Results do not improve.
S-23	2009	L. Suanmali , N. Salim, M. S. Binwahlan,	Fuzzy Logic, Particle Swarm Optimization (PSO)	DUC 2002	ROUGE 1(F-value):- 0.45524, ROUGE 2(F-value):- 0.20847	• To use the optimization algorithm to improve the performance.	• Not every feature is of same importance, so to treat them accordingly, the PSO is used. • PSO assigns the weights to each feature and then calculate the score of each token with its feature value and weight and pass it to FIS. • Performs better than other models, but still only fuzzy also performed the same.
S-24	2008	L. Suanmali , N. Salim, M. S. Binwahlan,	Fuzzy logic	DUC 2002 (6 documents)	F- value :- 0.50433	• To use human reasoning system.	• It was the first basic model of the techniques proposed in their rest five papers. • It uses eight features and Bell membership function.
S-25	2008	F. Kyoomarsi, H. Khosravi, E. Eslami,	Fuzzy Logic	Random 10 documents	Used Humans to test the summaries. Average accuracy of summary :77%	• Number of two-valued systems in world is very few.	• They compared vector approach based methods with fuzzy logic. • They first calculate the features scores and then used the MATLAB's Fuzzy tool to calculate the final score on their basis.

		P. K. Dehkordy, A. Tajoddin					
S-26	2006	Arman Kiani-B, M.-R. Akbarzadeh-T., M.H. Moeinzadeh	Fuzzy Logic	3 news articles	Average F- value:-0.752	• A novel technique for summarization to handle uncertainties	• They proposed the use of fuzzy logic with triangular membership function on the six features. • This was one of the basic method, whose limitations has been covered by the above models.
S-27	2006	Hsun-Hui Hunag, Horng-Chang Yang, Yau-Hwang kuo	Fuzzy Logic, Rough sets, Semantic patterns	8 articles from JAIR	Average ROUGE-1(F) :-0.4620391	• To remove the redundancy and uncertainty.	• Semantic of the words are taken to reduce the similar sentences with synonym words. • Fuzzy set and rough set improves the process of removing uncertainty in the feature values extracted. The final score of the redundancy removed articles are calculated by FIS.
S-28	2006	Arman Kiani-B, M.-R. Akbarzadeh-T.	Fuzzy Logic, Genetic Algorithm, Genetic Programming	3 news articles	F1 :-0.728, F value:-0.961	• To improve the results by improving the parameters of the machine (FIS).	• GA improves the membership function of the FIS. • GP improves the rule set according to the training data. • The accuracy of a fuzzy machine depends only on three thing its input, which we are pre-processing to improve better results, rest of the

							two have been improved in this method. • Performs better than other ATS techniques.
S-29	2005	Chang-Shing Lee, ZHi-Wei Jian, Lin-Kai Huang	Fuzzy Logic	News articles	Not accuracy, But the compression ratio was used for evaluation	• To use the ATS for news articles. • To improve the compression ratio of summary.	• Enhanced the FIS. • Instead of the five layer standard fuzzy system, a new seven layer fuzzy system was proposed. • Haven't been used by any researcher again, not tested for accuracy of the summary.
S-30	2003	R. Witte, S. Bergler	Fuzzy Logic, ERSS	DUC 2003	Not evaluated using standard measures.	• To resolve the co-reference resolution of the text words by reducing the uncertainty using Fuzzy logic.	• Used a POS tagger along with fuzzy logic and WordNet to handle the uncertainties in the co-reference text. • Results are not that satisfactory and need complex heuristics.

Table 1. Summary Table

CHAPTER 3

PROPOSED WORK

This chapter illustrates a novel approach that helps in achieving better results (in terms of ROUGE 1) as compared to previously used algorithms. Section 3.1 gives an overview of the research undertaken. Section 3.2 portraits the architectural view of the proposed paradigm. Section 3.3 describes each module of the system and how proposed algorithm works. Lastly, Section 3.4 gives the summary of the chapter.

3.1 Proposed Framework

The use of fuzzy logic in ATS since its first occurrence till date (2003 to 2017) has never been combined with graph based methods except for with Bushy path in 2016 by jyoti yadav [7]. Although the fuzzy has been used in combination with LSA a few times, but its effect with graph based methods has never been explored. This thesis focuses on the Fuzzy logic extraction approach for text summarization and the graph based approaches TextRank and LexRank for extracting the keyword using structure of the article and the semantic approach of text summarization using Latent Semantic Analysis.

Our hybrid model consists of four components: TextRank, LexRank, LSA, Fuzzy Logic. In this model, each text summarization method is used to extract the keywords, each method has ranked the keywords based on their importance (recurrence, centrality, noun etc), These keywords with their scores are then given as input to the final keyword extractor. In this phase, the keywords occurring in the final result of all the four methods are taken into final keyword list, Then the remaining keywords of all the methods are arranged in the descending order of the scores. From this list the top m keywords are selected in the final output. Keyword

extraction can be done to select as many keywords we want, but, usually 1% of the total keywords (except stop words) are enough to represent the central idea of the document.

Initially the document is passed through a pre-processing phase, to get only the required data. Like in the previous line, the keywords like is, the, a, to, only are not going to be the keywords representing the idea of the document. So, we first remove the unwanted stuff from the document and then also add some words which semantically mean the same, to get better semantic understanding. For example, the keywords like improving, enhancing are typically mean the same, so we can replace one with the other, so the actual occurrence of the word could be identified.

In TextRank the keywords or sentences of the document are represented as the vertices of the graph and the edges exist only between the vertices which has some sort of relationship like in case of sentences the common words, the position of two sentences, for keywords nouns, adjectives etc can be used for defining the relationship between the vertices. Number of common tokens like noun, verbs which are keywords to the text define the similarity of sentences or also called as overlapping nature.

LexRank calculates the importance of a sentence or keyword by finding their eigenvector centrality in the graph representation of the sentences. A connectivity matrix based on intra-sentence cosine similarity is developed to find the relation of a sentence with other sentences in the document.

In LSA Singular value decomposition(SVD) is used to capture and model inter-relationships among terms, so that it can semantically cluster terms that are recurring in the document and are represented as singular vectors. It improves the chances of getting more relevant results, as semantic content of the words is used rather than the direct mapping.

Fuzzy Logic is used as it handles the imprecision and vagueness in the data, but it does not consider the semantics of the text. Feature score of each word are calculated using statistical methods. First each feature is inputted with Trapezoidal membership function, then IF-THEN rules are applied to get the output score of each sentence. These processes has been explained in detail in chapter 2.

Summary generated by each method individually is passed on to summary selector. Which arranges the keywords in descending order of their scores. Top n unique keywords are selected from the list giving the keywords identifying the central idea of the document.

3.2 Architectural View

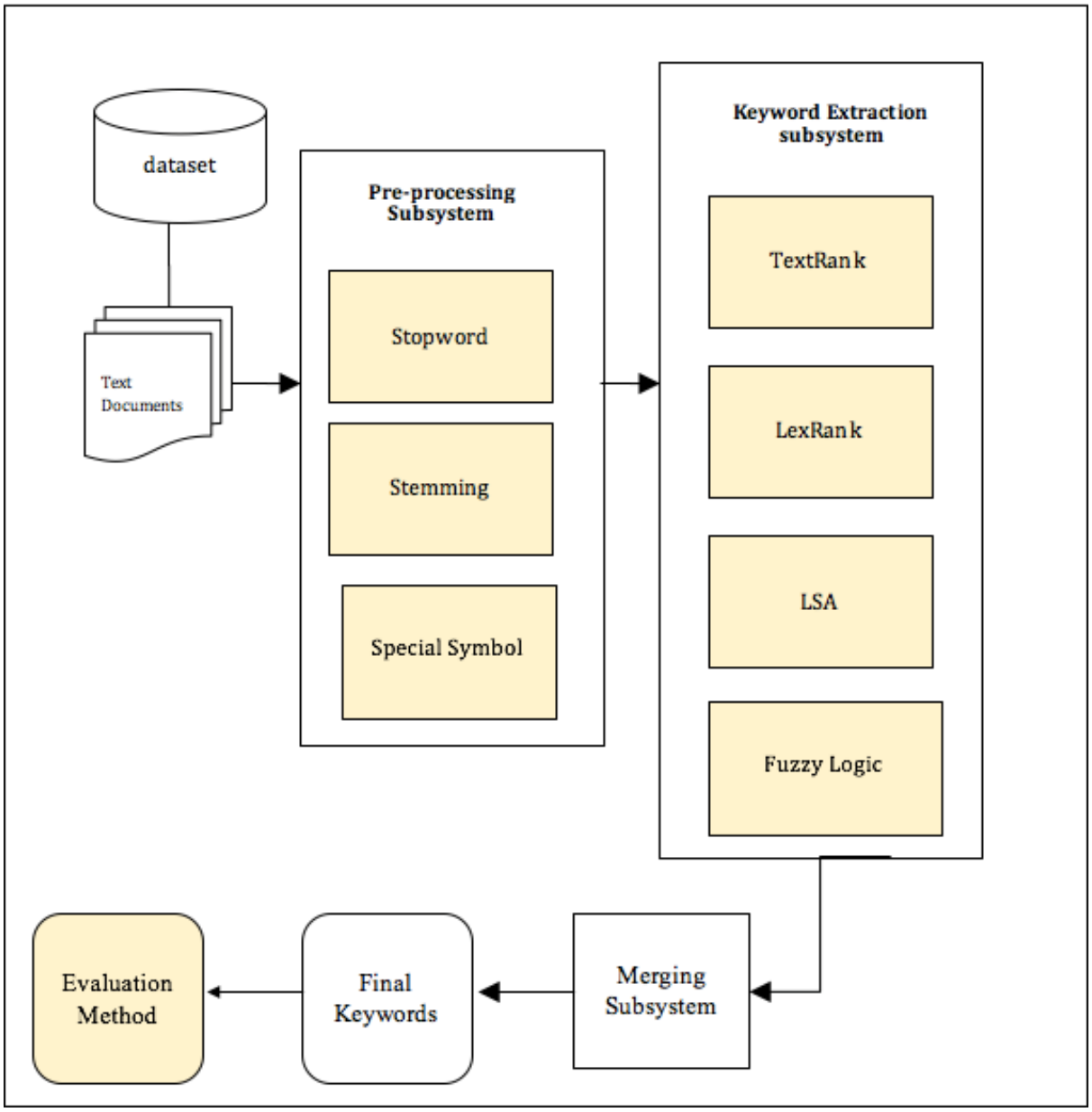

Figure.8. Proposed Architecture

The proposed approach firstly retrieves text documents from Opinosis dataset and then preprocess them by removing the unwanted words and converting rest of the words to its stem(root). For Keyword extraction the four methods processes the document in their own way. Later, the keywords extracted from all the methods are merged and arranged in descending order of their scores. Finally, the top 1% keywords are extracted from the list. Figure shows the overview of the system proposed in this research.

3.3 Evaluation Method

For measuring the extent to which the extracted keywords are accurate, the standard ROUGE-1 metric is used. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

The following five evaluation metrics are available.

- ROUGE-N: N-gram based co-occurrence statistics.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics. Longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.
- ROUGE-W: Weighted LCS-based statistics that favors consecutive LCSes.
- ROUGE-S: Skip-bigram based co-occurrence statistics. Skip-bigram is any pair of words in their sentence order.
- ROUGE-SU: Skip-bigram plus unigram-based co-occurrence statistics.

For our experiments, we use the ROUGE-N metric, with ***N = 1***, since we are only concerned with single-words extracted. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries.

The general formula for calculating the metric is:

$$ROUGE-N = \frac{\sum_{S \in ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} Count(gram_n)}$$

….(6)

Where n stands for the length of the n-gram, gramn, and $Count_{match}$ ($gram_n$) is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. It is clear that ROUGE-N is a recall-related measure because the denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side.

Note that the number of n-grams in the denominator of the ROUGE-N formula increases as we add more references. This is intuitive and reasonable because there might exist multiple good summaries. Also note that the numerator sums over all reference summaries. This effectively gives more weight to matching n-grams occurring in multiple references. Therefore a candidate summary that contains words shared by more references is favoured by the ROUGE-N measure. This is again very intuitive and reasonable because we normally prefer a candidate summary that is more similar to consensus among reference summaries.

3.4 Chapter Summary

This chapter explains the proposed method for extracting keywords in a single text document.

CHAPTER 4

IMPLEMENTATION

In this chapter we will discuss the experimental setup of the research work done. First section will discuss the data set followed by tools used for programming. In the next section, evaluation procedure is discussed. In the last section summary of the chapter is given.

4.1. Data Set

To begin the project, there are certain preparatory steps which were taken for implementing the proposed model. First and foremost, was data collection. We required a moderately large amount of meaningful textual content. The text should not be too specific about any particular genre of writing, nor be too generic to be summarized. The choice of a dataset to use a standard was crucial to the correctness and credibility of this study. The dataset considered standardized are provided by National Institute of Science and Technology (nist) under their information access division containing data from Document Understanding Conference (DUC) from 2001 to 2007 and from Text Analysis Conference (TAC) after that. For our project, we selected the Opinosis dataset. The reasons for selection of this dataset were the variety of non-related articles present in the dataset and the hand-written summaries of these articles provided with the dataset.

This dataset contains sentences extracted from user reviews on a given topic. Example topics are "Performance of Toyota Camry" and "Sound quality of ipod nano", etc. In total there are 51 such topics with each topic having approximately 100 sentences (on an average). The reviews were obtained from various sources - Tripadvisor (hotels), Edmunds.com (cars) and Amazon.com (various electronics). The dataset file also comes with gold standard summaries used for the Opinosis summarization paper. The Opinosis dataset also comes with human composed summaries used for the above topics. Furthermore, there were five distinct summaries for each of the articles, prepared by different people with each summary about 2 sentences (on an average) long, which helped in preparing a more reliable and accurate set of 'gold' keywords.

4.2. Algorithm

Algorithm proposed in chapter 3 was used. Each of the four techniques were first implemented separately, then their output is given as the input to the algorithm to extract more effective keywords, for determining the central idea of the document. Following figure shows the detailed flow chart of the proposed method containing step-wise procedure of all the four methods.

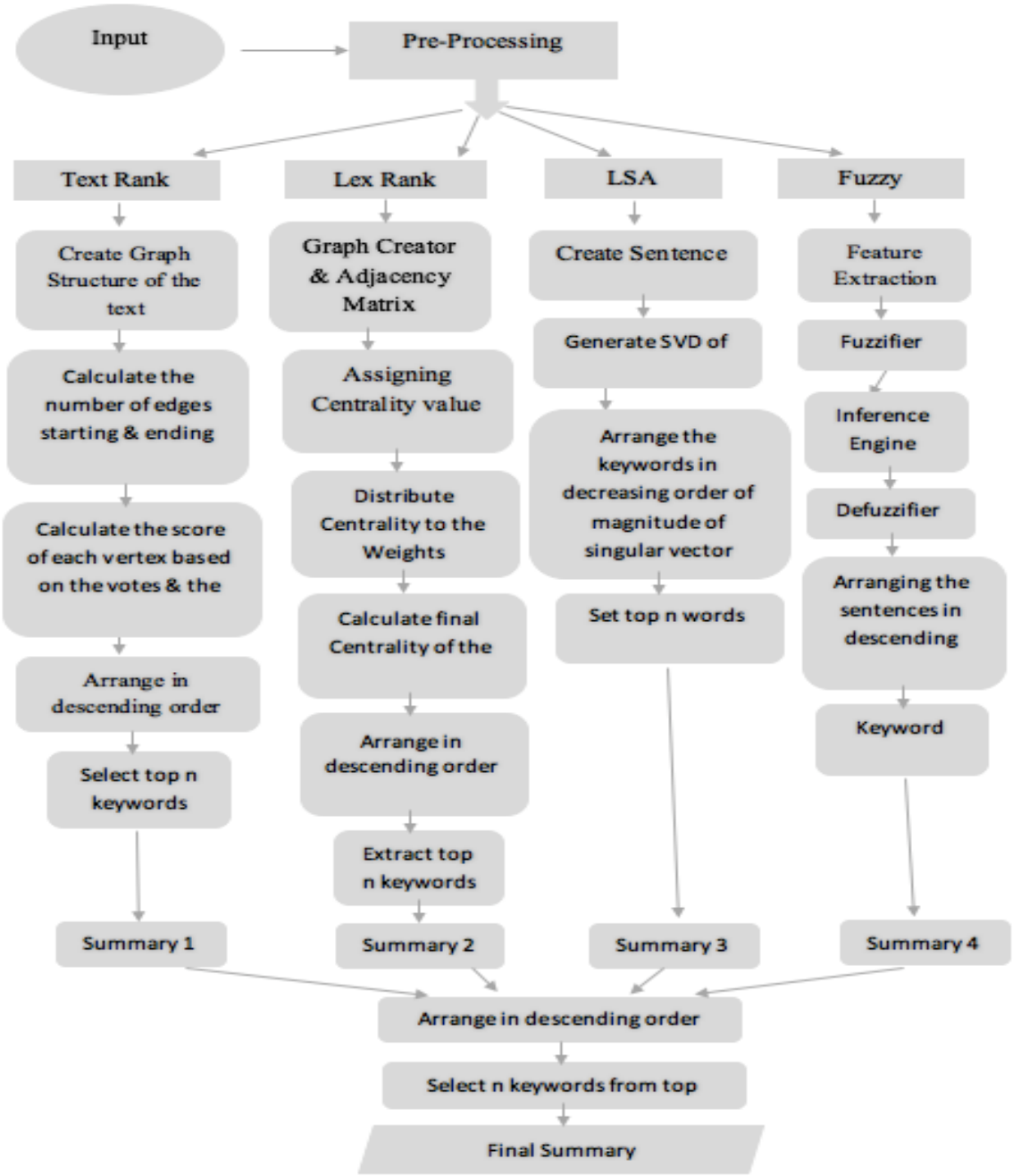

Fig.9. Flow Chart of the proposed model

4.3. Programming Tool

Firstly, features selection process is done, in our approach we have used eight features for determining the prominence score of a sentence These are Title word, Sentence length, sentence position, numerical data, thematic words, sentence to sentence similarity, term weight and proper noun. The feature score are then given as input to fuzzifier, which using trapezoidal membership function represents these scores in fuzzy values. These are then passed on to the inference engine, where based on the fuzzy rule set, the status of the sentence is extracted. The final step is defuzzifier which using the triangular membership function provide the prominence score to each sentence. These sentences with their score are then given as input to the merger subsection of the proposed method.

The rest of the three methods TextRank, LexRank and LSA were implemented in a Linux environment using python and bash. The bash scripts were written to act as drivers for automation of the project. The algorithms were implemented with python. The output file of each of the file is then passed onto the merger subsystem.

To generate the final summary having the keywords of the document the coding is done in Java language in NetBeans IDE. The program took four files generated by each method as input files containing the sentence and their scores determining their importance in the document. Firstly, the content of all four files are merged in one file, then the sentences which were occurring in more than one input file are kept only once to remove the redundancy. The sentences which were coming in more than one algorithm as the important sentence are first taken into the output file and are removed from the merged document. The rest of the sentences in the merged document are then arranged in descending order of their scores. If a document is of 1000 sentences and we want summary of only 20 lines that is only 2% of the actual document, we will select the top 20 sentences from the arranged list, only if none of the sentence was occurring in more than one method. If some sentences were coming as important in more than one method, say 3 sentences were found in summary of more than one method, then from final list only 17 sentences will be selected.

The final output file containing these sentences provide us with more relevant results, as proved by the comparison with human generated summary. The process of evaluating the summary is discussed in next section.

4.4. Evaluation Framework

The evaluation framework used is a composition of python and bash scripts developed for our specific purpose. The scripts perform a series of actions that filter and clean the dataset of unwanted characters, compute ROUGE-1 scores of human generated summaries, compute ROUGE-1 scores of the target algorithms and then compare them with those of human generated summaries. We explain each of these in detail.

First, before operating on the dataset, we must ensure that it is ready to use. It is ensured that the topic articles are titled appropriately, and that the text is free of punctuational or grammatical errors. This is followed by a filtering out the non-Unicode characters from the data, since they merely add to unwanted noise. Having prepared the data for use, we first extract keywords from the human generated summaries. The summaries individually are, on an average, 2 sentences long; with 5 such handwritten summaries per topic. To extract keywords, we use a simple frequency based filtering, wherein we tokenize the summaries individually, filter stop words and then select the most commonly occurring keywords. It was observed that, given how distinct the handwritten summaries were, taking into account the words that occurred twice or more were appropriate to be included in the gold keywords set. With this in mind, summaries of all topics were reduced to a list of prominent keywords of the respective topics. Once the keywords were extracted from the gold summaries, we compute the ROUGE-1 score of these summaries using k-fold validation technique. This is done since there are 5 unique summaries per topic, with each of them contributing to the final list of keywords in part. So, we compute their average ROUGE score by extracting keywords from individual summaries, and then comparing them against the gold keywords list using the ROUGE-1 technique.

We run a python implementation of the three algorithms, i.e. TextRank, LexRank and LSA , and Matlab implementation of Fuzzy logic on the dataset and the keywords were extracted. The ROUGE-1 score is then calculated according to the formula provided in section 3.3. The scores obtained are presented in the next chapter.

Programming Tools and software used

Operating System	: Macintosh, Linux
Language used	: Python, Java, Matlab
Library used	: NLTK, SUMY, NETWORKX
Software	: Matlab, NetBeans

CHAPTER 5

RESULTS & ANALYSIS

In Chapter 5, we discuss and analyse the result and related studies to fulfil our research objectives mentioned in chapter 1.

5.1 Output

The algorithms were run multiple times to and their scores and run-times were averaged for better understanding of the results. The following table summarizes the results obtained.

Sr. No.	Automatic Text Summarization Technique	ROUGE Score N-gram (N = 1)	Time (in sec)
1	LexRank	0.64580	17.52
2	Latent Semantic Analysis (LSA)	0.59937	12.08
3	TextRank	0.76217	13.97
4	Fuzzy Logic	0.54730	19.07
5	Hybrid Model (Proposed Method)	0.87634	16.48 (additional time needed to run all the four methods)

Table 2. Evaluation of Proposed model

As we can see, the ROUGE-1 score of the Proposed method is an upper bound on the results. Out of the five algorithms, the proposed model appears to be the most effective in extracting keywords. Its ROUGE score 0.87634 showed that the model performs far superior than the performance of any technique alone. Model runs in approximately 17 seconds, which is average with respect to the time

performance of the other algorithms. But this time is only of running of the final script which takes the input of the other four methods, so total time taken in the execution of the code is total of running time of all the 5 algorithms. Following graph shows the graphical representation of these results.

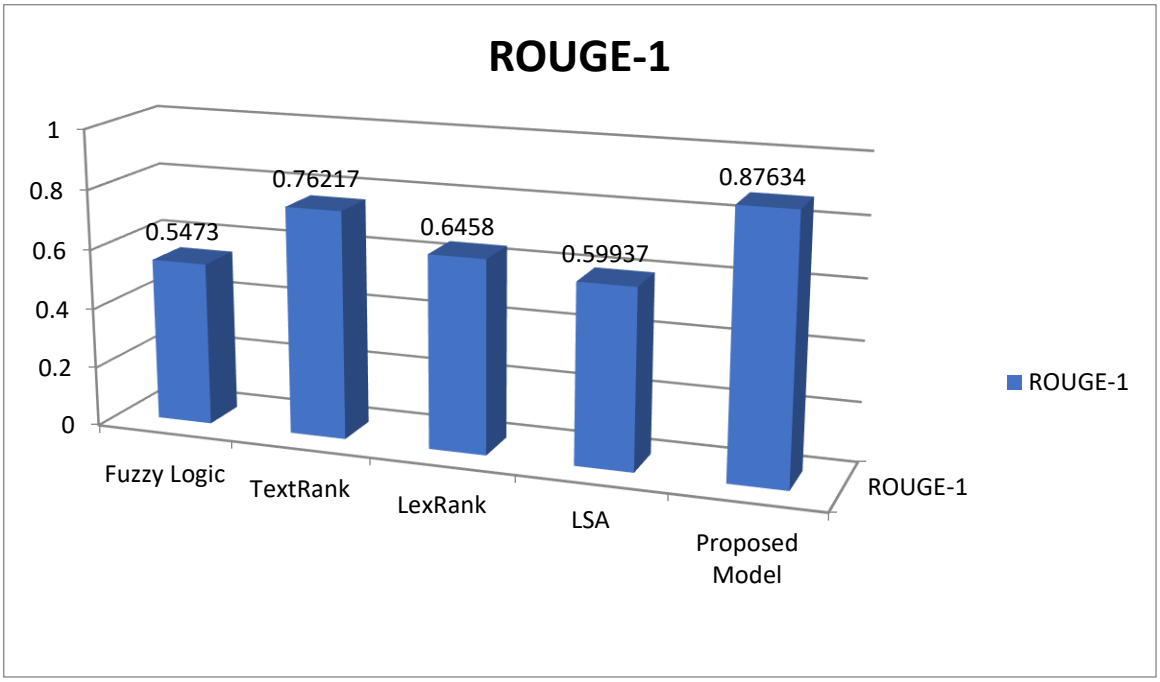

Fig.10 . Comparison of Proposed method using ROUGE-1

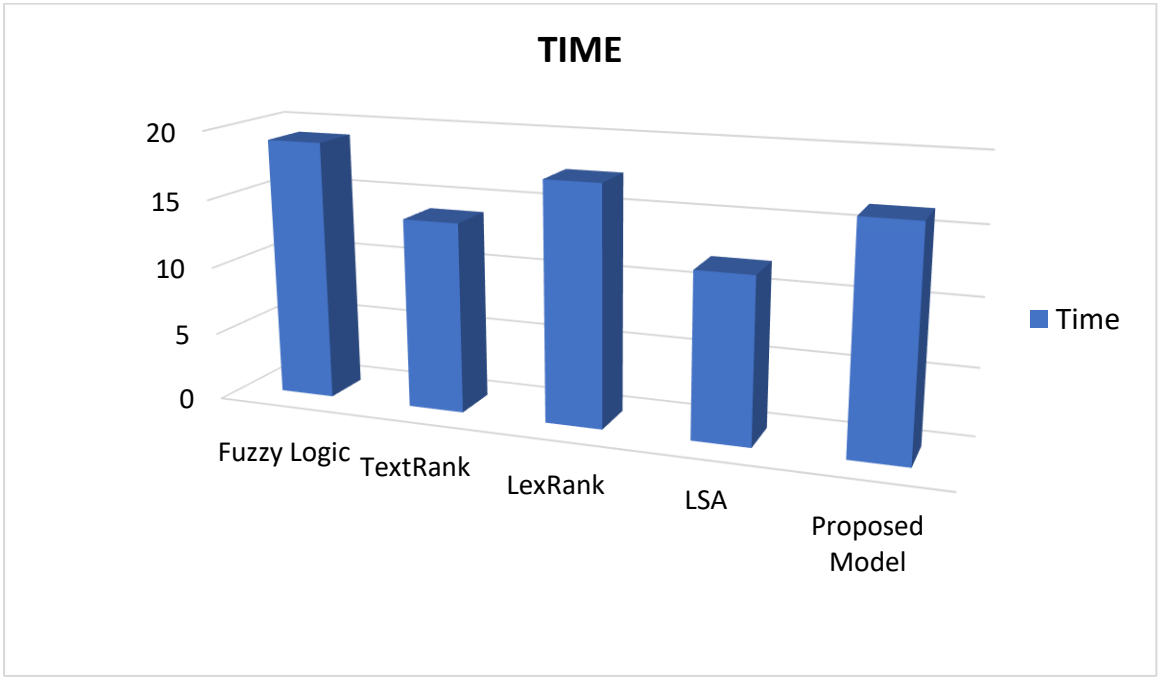

Fig.11 . Comparison of Proposed method using time taken to execute

The fastest of the algorithms was LSA, which extracted keywords in about 12 seconds. However, its ROUGE score was the second lowest compared to others, at 0.59937. LexRank and TextRank performed second best with respect to both time and ROUGE scores, scoring 0.76217 in less than 14 seconds.

5.2 Analysis

To comprehend the status of research in this field, we examined year-wise qualitative studies. Although the research in ATS started in late 1950s, but usage of fuzzy logic in ATS was first reported in twenty-first century. Rather more significant work in this field has been accounted for in the last 15 years. The following chart depicts the numbers of papers that have been published annually in this domain and indicate a positive trend with more research participation and gained momentum due to promising results & increased applications.

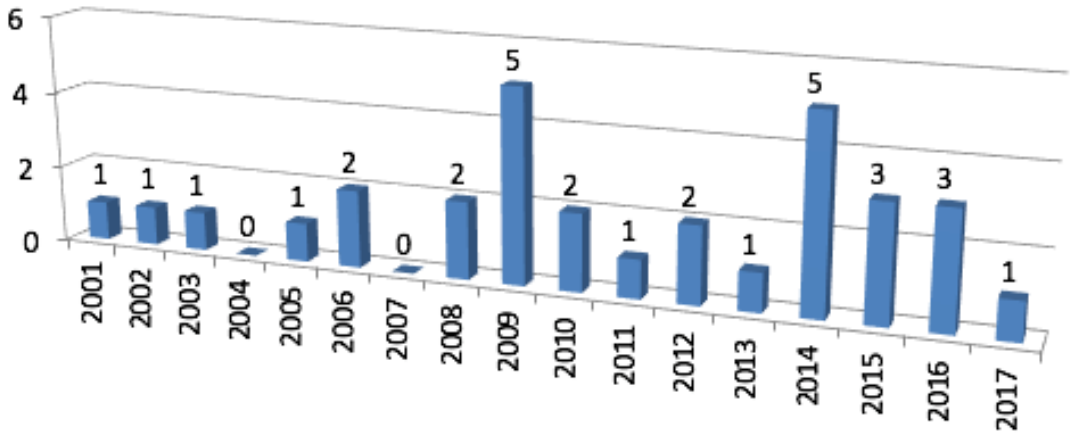

Fig. 12 Year wise distribution of studies

Some of the studies proposing fuzzy logic based text summarization methods are published in high impact journal. Scrutinizing the journals and conferences in which the selected studies have been published, the prominent work was found in publications by Elsevier. The journals which have published the 13 primary studies have been listed below. The rest of the 17 studies shown in related work section of chapter 2 were from conference proceedings (mostly IEEE), book chapters from springer.

Sr. No.	Journals	Number of studies
1.	Swarm and Evolutionary conference (Elsevier)	1
2.	Information Processing and Management (Elsevier)	1
3.	Applied Soft Computing (Elsevier)	1
4.	International Journal of Computer science trends and technology	1
5.	International Journal of innovative research in advance engineering	2
6.	International Journal of Soft Computing and Information Technology	2
7.	International Journal of Computer Engineering	2
8.	IEEE Transactions on systems, MAN & Cybernets	1
9.	International Journal of Computer Science and Technology	1
10.	International Journal of Engineering Research & Applications	1

Table 3 Journals Covering the Studies on ATS using Fuzzy Logic

The advantages of using fuzzy logic in ATS endorsed by the literature can be enlisted as:

- It handles the uncertainties in the input.
- It resembles the human reasoning system.
- Not everything in world can be defined in terms of zero and one, so provides more efficient ways to represent the feature values of sentences.
- It provides better way to calculate the sentence score using various types of membership functions.
- It handles the words better than other statistical techniques.

Some other hybrid techniques using fuzzy logic have been explored and the following table represents the techniques which have been used in combination with fuzzy logic for enhancing the results along with the number of studies conducted.

Sr. No.	Technique	Summarization Method	Number of times
1.	Genetic Algorithm	Optimization based	4
2.	Genetic Programming	Optimization based	3
3.	Particle Swarm Optimization	Optimization based	4
4.	Artificial bee Colony	Optimization based	2
5.	Rough sets	Statistical based	2
6.	Cellular Learning Automata	Statistical based	2
7.	Agglomerative K-means	Statistical based	1
8.	Latent Semantic Analysis	Semantic based	3
9.	WordNet	Semantic based	2
10.	Maximal Marginal Information	Diversity based	1
11.	Case Based Reasoning	Machine Learning based	1
12.	Neural Network	Machine Learning Based	1
13.	Bushy Path	Graph Based	1

Table 4 Techniques used along with Fuzzy logic in ATS

Figure 13 shows that around 50% of the studies about hybrid techniques are using optimization based methods, as they enhance the performance of the summarizer as compared to other techniques. Also, most of the studies combine more than 2 techniques to describe a hybrid model. The idea is to generate finest summaries with a fuzzy based model optimized with a nature inspired technique. Moreover, it was observed that the best results were achieved by combining more than two summarization methods with fuzzy logic [15]. So, we have chosen the hybrid model of four techniques. Two of which were graph based method, as they hadn't been explored with fuzzy logic yet. And the proposed model has performed well relative to the existing techniques. The following pie-chart depicts the number of studies classified on the basis of methods hybridized.

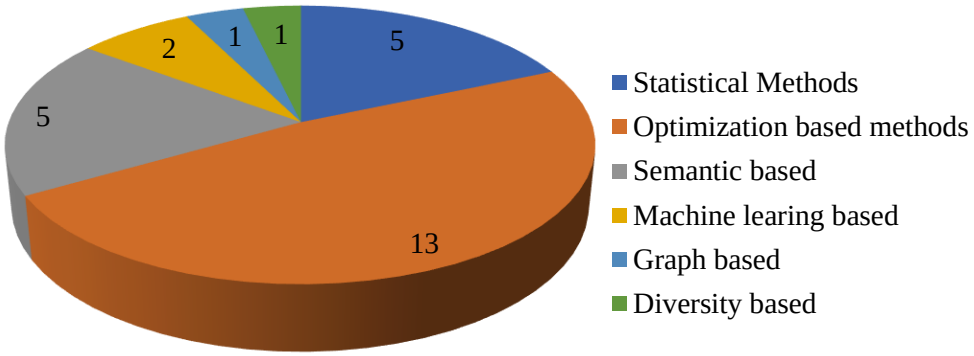

Fig 13. Classification of Hybrid models

For comparing the accuracy of different Fuzzy Logic Techniques used in Text Summarization we need a common dataset. But, not all the studies have been tested on the benchmarked dataset, rather some have been evalutated on small random data set.The distribution of these techniques according to the dataset used is shown in the following figure (Fig 14). Moreover, the techniques have not been evaluated using the same criteria and the standard evaluation tool for text summarization, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), it contains different matrix for automatically evaluate the worth of a summary generated by comparing it with a human generated summary [46]. Some of the studies have been tested on other basis too, such as testing based on humans understanding, compression ratio, fitness value etc.

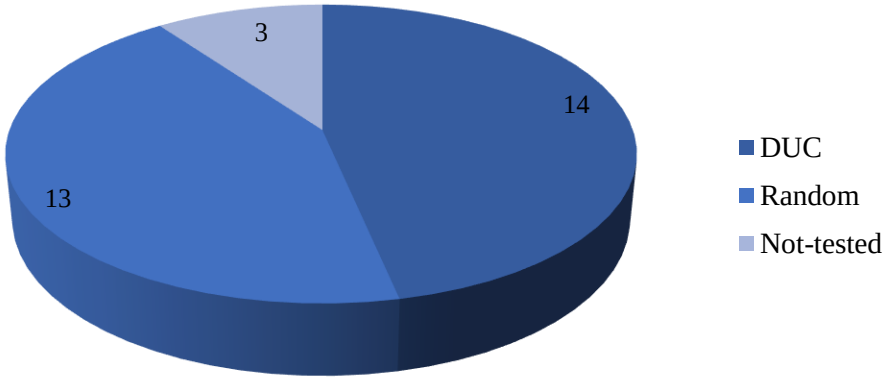

Fig 14 Studies distributed according to dataset used

5.3 Comparison

The *DUC* (*Document Understanding Conference*) *datasets* [9] are the de-facto *standard data sets* that the NLP community uses for evaluating summarization systems. From the 14 studies on DUC, 11 of them were using DUC 2002. We tried to compared all these techniques on the basis of evaluation results they have given but not every author had evaluated their generated summaries using standard matrices. Some of the studies had used other methods like precision, recall, and F-measue for determining the quality of the summary produced by the model, So the comparision of all the models are not possible, although the comparision between these different matrices provided by Ravindra et al shows that ROUGE values except ROUGE-WLCS are equivalent to F-measure [47], but no other study has still indorsed this result, that's why we are compaing the results of only the studies evaluated using ROUGE score. Though the techniques were tested on DUC but not all of them used the same set of documents and moreover the number of documents were also different. Thus, the comparison is not a benchmark result but still was adequate to show that proposed method performs well with respect to other existing models.

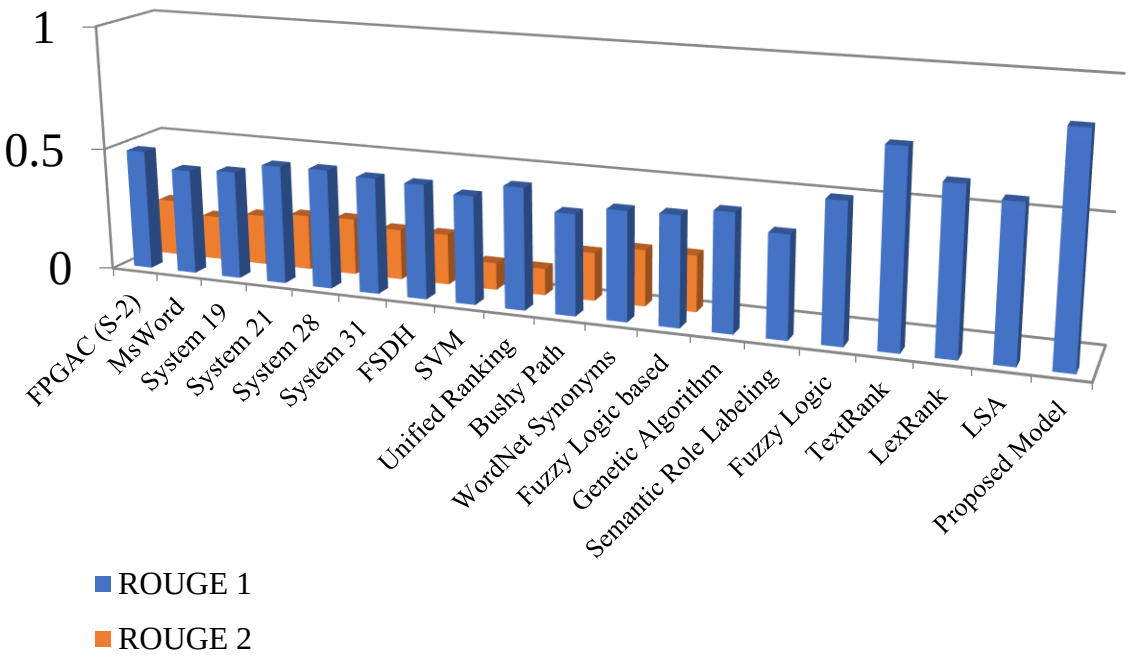

Fig 15. Comparison of Fuzzy based model with other text summarizers

Figure 15 shows the comparison of the best fuzzy based model FPGAC of all the fuzzy based model taken in related work, with other models using their ROUGE values on DUC 2002 dataset provided by [8,9,14,16,19,23,24,29,30]. The latest method ANFIS is also showing superior results, but as it was evaluated using F-measure, we cannot compare it with the FPGAC which surpassed all the other techniques when tested on the standard dataset DUC 2002. Our proposed method was evaluated only using ROUGE-1, So, even if it is showing the far better results than other methods, we cannot generalize the result until it is also tested on DUC and evaluation is done using ROUGE-2 too. The method can't be said to be the best one.

CHAPTER 6

CONCLUSION

This chapter concludes the contributions made by this thesis. Also figure out the limitation of the work done and briefly discuss the future scope of the research.

6.1 Research Summary

The evolution of Web from the 'Web of Documents' to the 'Web of People' (Social Web) to the 'Web of Data' (Semantic Web) has multiplied the volume, velocity and variety of information available online. Finding useful, relevant, trending, interesting information from this 'garbage in-garbage out' puddle has been a major focus of current research. Text summarization has emerged as one of the vital technique to handle this problem. The incredible ability of fuzzy inference systems to make logical assessment in an ambiguous and uncertain environment has made is a trending choice for practical applications such as text summarization, that involve imprecision and uncertainty. In our work, we have reviewed various studies on fuzzy-logic based text summarization from 2003 to 2017, published in conference proceedings and journals of high repute. The purpose was to evaluate the progress made so far, identify the trends and gaps in studies to ascertain the future scope of research within the domain. The following key-points were observed:

- Fuzzy logic is an attractive choice to achieve optimal summaries due to its resemblance to human reasoning. The goal of text summarization is to achieve superlative results comparable to human generated summaries and mapping the fuzzy logic inference mechanism gives the desired brainpower.
- It is promising to see the paradigm shift from conventional summarization methods to contemporary, novel intelligence based methods. Nearly 50% studies have been done with hybrid models of fuzzy logic with optimization methods, followed by equal

number of studies done on hybrids of fuzzy-logic with statistical and semantic techniques respectively. Combinational hybrids using two or more conventional techniques with fuzzy-logic though have been reported but the sphere is an open research problem to achieve enhanced summarization results.

- Through various hybrid techniques have been proposed, but none has taken the fusion of graph based techniques with fuzzy model.
- Hybrid models taking semantic into consideration are generating better results than the other methods
- Though the *DUC* (*Document Understanding Conference*) *datasets* are the de-facto *standard data sets* used for evaluating summarization systems but less than 50% studies have reportedly used it. This makes comparing the empirical results vague and non-uniform, thus identifying the need of more benchmarked studies and subsequent evaluation.

The research in this work introduces a fuzzy logic based hybrid model taking 2 graph based and one semantic based technique together to generate an upgraded summary, which performs better than all the fuzzy logic based models till date. From the systematic review of the existing techniques, we have seen that the proposed model is giving better-quality results in terms of ROUGE-1.

We identified and compared various automatic extractive text summarization techniques which use fuzzy logic. The limitation of this work is that the techniques obtained across studies are not compared on a common dataset, though most of them have used the ROUGE-N metric for evaluation. The need of more standard studies on benchmark dataset thus exists. Moreover, no studies have been done to classify the use of type-1 and type-2 fuzzy within the domain, calling for an investigation in this direction.

6.2 Limitation

The algorithms are compared on time scale and on their effectiveness in extracting keywords, which is measured by the ROUGE-N metric. As the techniques are not evaluated using the standard dataset provided by DUC, so the results cannot be generalized. We need to test the proposed method on benchmark dataset, and not just for keyword extraction but also for automated generated summary and the result should be evaluated at least by ROUGE-1 & ROUGE-2.

6.3 Future Scope

This study is empirical in nature and can be improved by giving different weights to the four techniques like in [17], the weightage given to the methods can affect the output of the proposed method drastically. Features taken in the feature extraction step can also be given different weightage to make the summary creation process like humans, as we human don't take each attribute as same, like the title keywords are given more importance than the noun. So, the different weights assigned to each feature can improvise the results.

The results demonstrate that the discussed algorithm has the characteristics of extracting the keywords which depicts the focus of the document, making the search process easy to find the relevant documents. The algorithm has only been investigated on Opinosis Dataset. The study can be carried further on benchmarked dataset. We can further extend this work by:

- Assigning different weights to each algorithm used.
- Assigning different weights to the features extracted in the process of summary generated using Fuzzy logic.
- Test the model on benchmark dataset DUC.
- Use the same method not just for keyword extraction but also for generating summary.
- Incorporate some more Nature Language based (NLP) techniques to generate abstractive summary.

REFERENCES

1. Bhatia, MPS. & Kumar, A. (2008). Information retrieval and machine learning: supporting technologies for web mining research and practice. Webology, 5(2), article 55.
2. Bhatia, MPS. & Kumar, A (2008). A primer on the Web information retrieval paradigm. *Journal of Theoretical and Applied Information Technology,* 4(7), 657-662
3. H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBM Journal of Research and Development, vol. 2, pp.159-165. 1958.
4. Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization MIT Press. 1999.
5. Paice, C.D. 1990. Constructing literature abstracts by computer: Techniques and prospects. Inf.Proc. & Management 261,171-186.
6. C. Freksa, "Fuzzy Logic: An Interface Between Logic and Human Reasoning," IEEE Expert 9, pp. 20–21, 1994.
7. P. Modaresi, S. Conrad, "From Phrases to keyphrases: An Unsupervised Fuzzy Set Approach to Summarize News Articles", ACM, 10.1145/2684103.2684117, Dec, 2014.
8. R. Sharma, P. Sharma, "A survey on extractive text summarization", IJARCSSE, vol.6, issue 4, 2016.
9. M. Gambhir, V. Gupta, "Recent automatic text summarization techniques: a survey", Springer, 10.1007/s10462-016-9475-9, 2016.
10. Raj kumar V. S., Chandrakala D., "A survey on text summarization using optimization algorithm", ELK Asia Pacific Journals, 2016.
11. Lotfi A. Zadeh, "Fuzzy Logic = Computing with Words", IEEE Transactions on Fuzzy Systems, VOL. 4, NO. 2, MAY 1996.
12. F.E. Boran, D. Akay, R. R. Yager, "An overview of methods for linguistic summarization with fuzzy sets", Expert Systems with Applications, Elsevier, Vol.61:356-377, 10.1016/j.eswa.2016.05,044, 2016.
13. B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, "Systematic Literature Reviews in Software Engineering- A systematic literature

review", Information and Software Technology, Elsevier, Vol.51(1):7-15, 10.1016/j.infsof.2008.09.009.

14. Y. J. Kumar, F. J. Kang, O. S. Goh, A. Khan, "Text Summarization based on Classification using ANFIS", Advanced topics in Intelligent Information and database systems, Vol. 710, pp : 405-417, 10.1007/978-3-319-56660-3_35.
15. R. Abbasi-ghalehtaki,H. Khotanlou, M. Esmaeilpour, "Fuzzy Evolutionary cellular learning automata model for text summarization", Swarm and Evolutionary Computing, Elsevier, vol.30:11-26,2210-6502, 2016.
16. Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. British MedicalJournal, 323(7303), 42-46.
17. Jyoti Yadav,Dr. Yogesh Kumar Meena, "Use of Fuzzy logic and WordNet for improving the performance of extractive automatic text summarization ", IEEE, 978-1-5090-2029-4/16, *2016.*
18. Mehdi Jafari,A. M. Shabhabi, J. Wang, Y. Qin, X. Tao, M. Gheisari, "Automatic text summarization using fuzzy inference", IEEE, 10.1109/IConAC.2016.7604928, 2016.
19. FarshadKiyoumarsi, "Evaluation of automatic text summarizations based on human summaries",Procedia – Social and Behavioural Sciences, Elsevier, Vol. 192:83-91, 10.1016/j.sbspro.2015.06.013, 2015.
20. S. A. Babar, Pallavi D. Patil, "Improving performance of text summarization", Procedia Computer Science, Elsevier, Vol.46:354-363, 10.1016/j.procs.2015.02.031, 2015.
21. Pallavi D. Patil, P. M. Mane, "Improving the performance for single and multi-document text summarization by LSA & FL", IJCST, vol.3, issue-4, 2015.
22. S. A. Babar, S. A. Thorat, "Improving text summarization using fuzzy logic and LSA", IJIRAE, vol.1,issue-4,2014.
23. R. Abbasi-ghalehtaki,H. Khotanlou, M. Esmaeilpour, "A combinational method of fuzzy, particle swarm optimization and cellular leaning automata for text ssummarization", IEEE, 978-1-4799-3351-8/14, 2014.
24. Pallavi D. Patil, N. J. Kulkarni, "Text summarization using Fuzzy logic", IJIRAE, vol.1, issue-3, 2014.
25. R. J. Shinde, S. H. Routela, S. S. Jadhav, S. R. Sagare, "Enforcing Text summarization using fuzzy logic",IJCSIT, vol.5, issue-6, 2014.

26. Y. J. Kumar,N. Salim, A. Abuobieda, A. T. Albaham, "Multi document summarization based on news components using fuzzy cross-document relations", Applied Soft computing, Elsevier, Vol.21:265-279, 10.1016/j.asoc.2014.2014.03.041.
27. A. R. Kulkarni, S.S. Apte, "A domain specific automatic text summarization using Fuzzylogic", IJCET, vol.4, issue-4, 2013.
28. A. Ladekar,A. Mujumdar, P. Nipane, S. Tomar, Kavitha S., "Automatic text summarization using : Fuzzy GA-GP", IJERA, vol.2, issue-2, 2012.
29. R. S. Dixit, S.S. Apte, "Improvement of text summarization using fuzzy logic based method", IOSR, vol.5, issue-6, 2012.
30. L. Suanmali, N. Salim,M. S. Binwahlan, "Fuzzy genetic semantic based text summarization", IEEE, 10.1109/DASC.2011.192, 2011.
31. M. S. Binwahlan, N. Salim,L. Suanmali "Fuzzy swarm diversity hybrid model for text summarization", Information Processing & management, Elsevier, Vol.46 (5): 571-588, 10.1016/j.ipm.2010.03.004, 2010.
32. S. Alzahrami, N. Salim, C.K. Kent, "The development of cross-language plagiarism detection tool utilising fuzzy swarm-based summarization", IEEE, 978-1-4244-8136-1/10, 2010.
33. Hsun-Hui Hunag, Horng-Chang Yang, Yau-Hwang kuo, "A fuzzy-rough hybrid approach to multi-document extractive summarization", IEEE, 10.1109/HIS.2009.41, 2009.
34. L. Suanmali ,N. Salim, M. S. Binwahlan, "Feature-based sentence extraction using fuzzy inference rules", IEEE, 10.1109/ICSPCS.2009.156, 2009.
35. L. Suanmali ,N. Salim, M. S. Binwahlan, "Fuzzy logic based method for improving text summarization", IJCSIS, vol.2, issue-1, 2009.
36. L. Suanmali ,N. Salim, M. S. Binwahlan, "Sentence features fusion for text summarization using fuzzy logic", IEEE, 10.1109/HIS.2009.36, 2009.
37. L. Suanmali ,N. Salim, M. S. Binwahlan, "Fuzzy Swarm based text summarization", Journal of computer science, 2009.
38. L. Suanmali ,N. Salim, M. S. Binwahlan, "Automatic text summarization using feature based fuzzy extraction", JurnalTeknologiMaklumat, 2008.
39. F. Kyoomarsi, H. Khosravi, E. Eslami, P. K. Dehkordy, A. Tajoddin, "Optimizing text summarization based on fuzzy logic", IEEE, 10.1109/ICIS.2008.46, 2008.

40. Arman Kiani-B, M.-R. Akbarzadeh-T., M.H. Moeinzadeh, “Intelligent extractive text summarization using fuzzy inference system”, IEEE, 2006.
41. Hsun-Hui Hunag, Horng-Chang Yang, Yau-Hwang kuo, “Fuzzy-rough set aided sentence extraction summarization”, IEEE, 2006.
42. Arman Kiani-B, M.-R. Akbarzadeh-T., “Automatic text summarization using: Hybrid Fuzzy GA-GP”, IEEE, 2006.
43. Chang-Shing Lee, ZHi-Wei Jian, Lin-Kai Huang, “A fuzzy ontology and its application to news summarization”, IEEE, 10.1109/TSMCB.2005.845032, 2005.
44. R. Witte, S. Bergler, “Fuzzy coreference resolution for summarization”, 2003.
45. G. Carenini, R. T. Ng, X. Zhou, “Summarizing Email Conversations with Clue Words”, ACM, 978-1-59593-654-7/07/0005, May 2007.
46. Chin- Yew Lin, “ROUGE: A Package for automatic evaluation of summaries”,2004.
47. G. Ravindra, N. Balakrishnan, K. R. Ramakrishnan, “Automatic Evaluation of Extract Summaries using Fuzzy F-score measure”, ACM transactions on Database Systems, 2013.
48. Günes Erkan, Ann Arbor, Dragomir R. Radev, “LexRank: graph-based lexical centrality as salience in text summarization”, Department of EECS, University of Michigan.
49. Bhavana Lanjewar, “Automatic text summarization with context based keyword extraction”, International Journal of Advance Research in Computer Science and Management Studies, 2015.
50. Kang Wu, Ping Shi, Da Pan, “An approach to automatic summarization for Chinese text based on the combination of spectral clustering and LexRank”, IEEE *Access 2016.*
51. H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.
52. S. Brin and L. Page. 1998. “The anatomy of a large-scale hypertextual Web search engine.” Computer Networks and ISDN Systems, 30(1–7).
53. N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, Nagercoil, 2016, pp. 1-7.
54. Mihalcea R, Tarau P. “TextRank: Bringing order into texts.” *Association for Computational Linguistics, 2004.*

55. Gunes Erkan, Dragomir R. Radev “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization” *Journal of Artificial Intelligence Research 22* (2004) 457-479.
56. Gong, Yihong, and Xin Liu. "Generic text summarization using relevance measure and latent semantic analysis." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.
57. Ganesan, K. A., C. X. Zhai, and J. Han, "Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions", *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, 2010.
58. Lin, Chin-Yew, and F. J. Och. "Looking for a few good metrics: ROUGE and its evaluation." *NTCIR Workshop*. 2004.