# Automatic Text Summarization

Riya Kamble
Atharva College of Engineering
Mumbai, India

Saurabh Shah
Atharva College of Engineering
Mumbai, India

Aalok Nerurkar
Atharva College of Engineering
Mumbai, India

Kanhaiya Prasad
Atharva College of Engineering
Mumbai, India

Prof. Reena Mahe
Atharva College of Engineering
Mumbai, India

***Abstract*- With the fast development of the quantity and complexity of archive sources on the internet, it has come to be increasingly more essential in imitation of providing a modern mechanism for user for finding specific facts in available documents. Text summarization has turned out to be an essential and well timed tool because of supporting and then decoding the tremendous volumes of text available into documents. Summarization" is a method of bringing a lesser version of original text that contains the important information. It can be broadly differentiated into two types which are Extraction and Abstraction. This project focuses on the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis.**

***Keywords—Text summarization, Fuzzy logic Extraction, Latent Semantic Analysis, Semantic approach, Extraction, Abstraction***

## I. INTRODUCTION

Automatic summarization means a mechanically short output is addicted when an input is applied. We should remember that an input is a well structured document. For this even there are opening preprocesses such as Tokenization, Sentence Segmentation, Removing stop words and Word Stemming.

An extractive summarization method is composed of choosing most important sentences, words, paragraphs etc. from the original record and concatenating them into shorter form. An Abstractive summarization is a grasp about the predominant ideas in a file and expresses those thoughts into an obvious simplistic language.

Text Summarization is a lively concern of research among every text regarding the IR and NLP communities. People can keep up with the world affairs by listening to news bites. People can go to the movies largely over the basis of critiques they've seen. People can base investment decisions on stock market updates. With summaries, People can make effective decisions in less time. The motivation right here is in conformity to construct a certain system which is computationally surroundings pleasant and then create summaries automatically.

Text summarization is able to stay categorized in two ways, as abstractive summarization and extractive summarization. Extractive summarization [6] is bendy but consumes much less time namely compared to abstractive summarization. In extractive summarization it considers the whole paragraph into a matrix form, and on the basis of some feature vectors all the indispensable or vital sentences are extracted.

## II. LITERATURE SURVEY

This work [1] is done by Archana AB and Sunitha. C in 2013 and it deals with Text summarization which can be classified into two approaches: extraction and abstraction. This paper focuses on extraction technique. The purpose of text summarization concerning extraction strategy is sentence selection. One of the techniques in conformity is to gain the appropriate sentences assigned partial numerical measure of a sentence for the summary called sentence weighting and then choose the best ones. A summary textual content is a derivative of a source text condensed through selection and/or generalization on necessary content. Query-focused summaries enable customers to find more applicable documents more accurately, with less need to seek advice for the full text of the document. Extractive summarization methods try to locate the most necessary topics of an input document and pick sentences that are related to these select concepts to create the summary. This paper is a Comparative discipline of four methods used for extractive summarization, namely, Neural Network, Graph Theoretic, Fuzzy based method and Cluster based method.

Advantages- Query-focused summaries allow customers in conformity to discover greater applicable archives more accurately.

Disadvantages- Neural Network, Graph Theoretic, Cluster methods are inefficient to use.

This work [2] is done by Josef Steinberger and Karel Ježek in 2009 and it deals with using latent semantic analysis in text summarization. It describes a generic text summarization method which utilizes the latent semantic analysis technique to perceive semantically necessary sentences. Then it proposes twin modern evaluation methods based on LSA, which measure context similarity between an authentic file and its summary. In the evaluation part we compare seven summarizers by a classical content-based evaluator and by the two new LSA evaluators. We also learn an influence regarding summary length on its quality from the angle of the three mentioned assessment methods. LSA is very sensitive on a stoplist

and a lemmatization process. Other weighing schemes and a normalization of a sentence vector on the SVD input is needed. Other evaluations are needed, especially on longer texts than the Reuters documents are.
Advantages- It has an effect regarding summary extent on its quality from the angle concerning the three pronounced evaluation methods.
Disadvantages- It needs other weighing schemes and a normalization of a sentence vector on the SVD input.

This work [3] is done by Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva in 2013 and deals with Text summarization which is the process of automatically creating a shorter version of one or more text documents. Essentially, text summarization techniques are classified as Extractive and Abstractive. Extractive strategies perform textual content summarization by choosing sentences of files according to partial criteria. Abstractive summaries strive to improve the coherence amongst sentences by disposing redundancies and clarifying the contest of sentences. In phrases regarding extractive summarization, sentence scoring is the technique most used for extractive textual content summarization. This paper describes and performs a quantitative and characteristic assessment concerning 15 algorithms for sentence scoring ready in the literature. Three special datasets (News, Blogs and Article contexts) have been evaluated. In addition, instructions to improve the sentence extraction results obtained are suggested. This paper provided the five best results obtained with different test sets, one would obtain a coincidence of four methods as being the best ones: TF/IDF, Word Frequency, Lexical Similarity and Sentence Length. The strategy ‘‘Text-Rank Score’’ was also chosen by as providing good results for two of the three data sets tested. Advantages- This paper describes and performs a quantitative and qualitative assessment of 15 algorithms for sentence scoring available in the literature.
Disadvantages- It does not methods need to be evaluated on provide good test results and other the basis of sentence scoring and structure overload problem by extracting the most important information from a document and which can help a reader to decide whether it is relevant or not. In this paper we advocate an approach of personalized textual content summarization which improves the conventional automated text summarization methods with the aid of accepting the differences within reader’s characteristics.
It uses annotations added by readers as one of the sources of personalization. In this paper we have proposed a method of personalized summarization, which improves traditional summarization strategies by taking various user characteristics including context. The achievement lies in the proposal of the specific raters that take into account terms applicable for the domain or the stage of knowledge of an individual user and the technique of the raters’ aggregate which permits thinking about more than a few parameters or context of the summarization.
Advantages- In this paper we have proposed a method of customized summarization, which extends present summarization techniques by considering various user characteristics along with context.
Disadvantages- It takes into account terms relevant for the domain or the level of knowledge of an individual user and the method of the raters’ combination which is ineffective.

## III. METHODOLOGY

Traditional extraction methods can’t capture semantic relations among standards between a textual content. Therefore, the use of semantic analysis in conformity to capture the semantic articles within sentences including sentence extraction technique brings the elevated summarization method. Our proposed method can improve the virtue of summary with the help of latent semantic evaluation and sentence function extracted by fuzzy logic system. The proposed text summarization method has two main phases. Feature extraction [9] from multiple documents is considered as the initial phase. The characteristic extraction includes, function matrix technology from the applications extracted out of each sentences. Then, the feature matrix is processed with a fuzzy classifier. The rules and regulations generated of the fuzzy classifier are afterwards selected for generating a function matrix primarily based regarding fuzzy score. A feature vector is an n-dimensional vector about numerical functions that represent some of the object. The primary objective of text summarization based totally over extraction approach is the choosing of proper sentence through the need of a user.

## IV. IMPLEMENTATION

Our System consists of 5 modules:

### *A. Preprocessing:*

There are four steps [10] in preprocessing:

*a.* Segmentation: It is a process of dividing a given document into sentences.

*b.* Removal of Stop words: Stop phrases are frequently occurring phrases such as ‘a’ an’, the’ that provides less meaning and consists of noise. The Stop words are predefined and stored within an array.

*c.* Tokenization: The words are assigned tokens or weights according to the usage and importance.

*d.* Word Stemming: converts every word in its root form by eliminating its prefix and suffix which can be used for comparison with other words.

### *B. Feature Extraction:*

The text file is represented by using set, D= {S1, S2, - - - , Sk} where, Si signifies a sentence contained in the file D. The file is subjected to function extraction. The necessary word and sentence features to be used are decided .This employment uses features such as Title word, Sentence length, Sentence position, numerical data, Term weight, sentence similarity, existence of Thematic phrases and proper Nouns . The flow of summarization is as shown in the Fig 1.

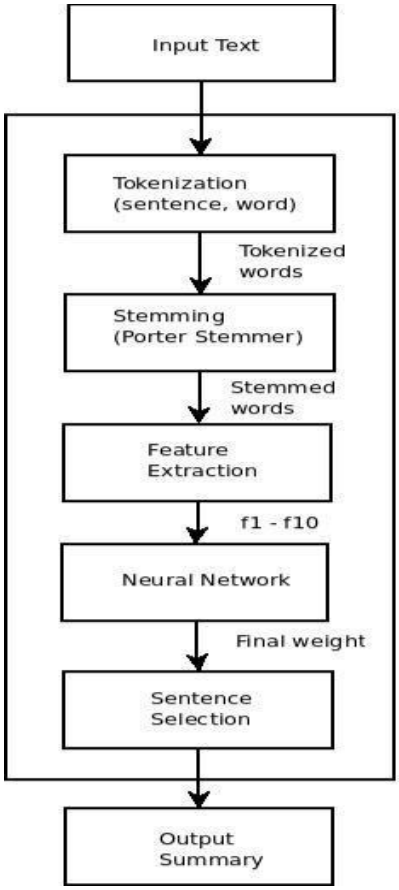


Figure 1: Flow for Summarization

*C. Fuzzy Logic Scoring:*

Thus each sentence is associated with 8 feature vectors. Using [4] all the 8 feature scores, the score for each sentence are derived using fuzzy logic method. The fuzzy sense method uses the fuzzy regulations and triangular membership function. The fuzzy regulations are in the structure of IF-THEN. The triangular membership function fuzzifies each value into one of 3 values that is LOW, MEDIUM & HIGH. Then we apply fuzzy regulations in imitation of deciding whether sentence is unimportant, average or important. This is also known as defuzzification. Sample of IF-THEN rules are described below:

IF (No Word In Title > 0.81) and (Sentence Length > 0.81) and (Term Freq > 0.81) and (Sentence Position > 0.81) and (Sentence Similarity > 0.81) and (No Proper Noun > 0.81) and (No Thematic Word > 0.81) and (Numerical Data > 0.81) THEN (Sentence is important).

Fuzzy Logic for Summarization is as shown in the Fig 2.1

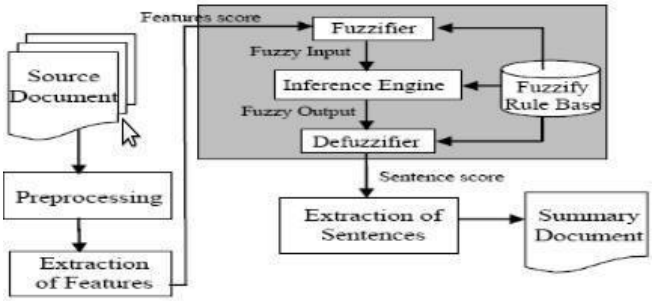


Figure 2: Fuzzy Logic for Summarization

*D. Sentence Selection:*

All the sentences regarding a file are ranked [7] in a descending system based on their scores. Top n sentences of absolute best score are extracted as file summary primarily based on compression rate. Finally the sentences in summary are organized in the order they appear in the unique document.

*E. Latent Semantic Analysis:*

Semantic Analysis is a statistical model of phrase utilization that allows comparisons of semantic harmony between pieces of textual information. It was initially designed to improve the usefulness of record retrieval methods by carrying out recovery on the basis of derived "semantic" content of terms in a query as opposed to performing direct word matching. This approach avoids partial issues of synonymy, in which different words can be used to draw the same semantic concept. The predominant assumption of LSA is that there is partial underlying or "latent" structure in the pattern of phrase usage across documents, and that statistical technique can be used to estimate this latent structure. The term "documents" in this case, can be thought of as contexts in which phrases appear and also could be smaller textual content segments such as individual paragraphs or sentences. Through an evaluation [8] of the associations among phrases and documents, the method produces an illustration in which words that are used in similar contexts will be greater semantically associated.

All summarization methods based on LSA use three main steps. These steps are as follows:

*a.* Input Matrix Creation.
*b.* Singular Value Decomposition.
*c.* Sentence Selection.

*(a) Input Matrix Creation:*

The preceding step of input matrix creation is to create the matrix in the form of different terms of x sentences. Assuming there are m terms and n sentences, the matrix A with size of m x n is created, which is A = [A1, A2, An]. Each column Ai represents adequate vector term of sentence i of the input file. The specified terms can be phrases that have been seen in the sentences, or they can be preprocessed before the creation of the matrix.

In order to reduce matrix size, the rows of the matrix the words can be reduced by preprocessing approaches like stop word removal, using roots of words only, using

phrases instead of words and etc. These preprocessing approaches are mostly language dependent. The cell is filled with the frequency of the word in the sentence.

*(b) Singular Value Decomposition:*

SVD is an algebraic method that can model relationships among words/phrases and sentences.

Singular value decomposition is a mathematical approach which replicates the correlation among phrases and sentences. It disintegrates the input matrix into three sordid matrices as follows:

$A = U \Sigma V^T$

A: Input matrix (m x n)

U: Words x Extracted Concepts (m x n)

$\Sigma$ : Scaling values, diagonal descending matrix (n x n)

$V^T$: Sentences x Extracted Concepts (n x n)

Singular Value Decomposition is as shown in the Fig 3.

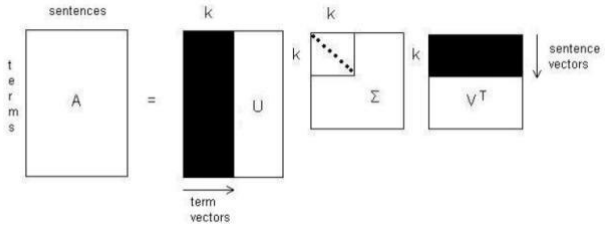


Figure 3: Singular Value Decomposition

*(c) Sentence Selection:*

Using the results involving SVD, different algorithms use extraordinary tactics in imitation of choosing essential sentences. After representing the input document in matrix and doing calculations of SVD, VT matrix, matrix of extracted concepts x sentences, is used for selecting the important sentences. In VT matrix, row order indicates the virtue in relation to the requirements such that the first row represents the close quintessential thought extracted. The cell values of this matrix show the relation between the words and the concept. A greater cell value indicates that the sentence is larger related along the concept. The quantity of sentences to remain accumulated is given as a percentage regarding Summaries within table 1.

| $V^T$ matrix (r = 2) | | | |
|---|---|---|---|
| | Sent0 | Sent1 | Sent2 |
| Con0 | 0,457 | 0,728 | 0,510 |
| Con1 | -0,770 | 0,037 | 0,637 |

Table 1

## V. CONCLUSION

Automatic summarization is a complicated assignment that consists of quite a few sub-tasks. Every sub-task immediately affects the purpose to cause high virtue summaries. In extraction based summarization the necessary portion regarding the process is the identification of essentially applicable sentences of text. Use of fuzzy logic as a summarization sub-task elevated the virtue concerning summary by a greater amount. The results are clearly visible in the comparison graphs. Our algorithm shows better results as compared to the output produced by twin online summarizers. Thus our proposed technique improves the virtue of summary by incorporating the latent semantic analysis into the sentence function extracted fuzzy logic system to capture the semantic relations between ideas in the text.

## ACKNOWLEDGMENT

We would like to express immense gratefulness to our guide Prof. Reena Mahe for her motivation and guidance throughout. We would also like to thank the faculty of Department of Information Technology who greatly assisted our research and for their valuable help.

## REFERENCES

[1] Archana AB, Sunitha. C, "An Overview on Document Summarization Techniques", International Journal on Advanced Computer Theory and Engineering (IJACTE), ISSN (Print): 2319 ′U 2526, Volume-1, Issue-2, 2013.

[2] Josef Steinberger, Karel Ježek , "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň.

[3] Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva ,Fred Freitas ,George D.C. Cavalcanti ,Luciano Favaro , "Assessing sentence scoring techniques for extractive text summarization ",Expert Systems with Applications 40 (2013) 5755-5764 ,2013 Elsevier.

[4] Mrs.A.R.Kulkarni , Dr.Mrs.S.S.Apte "A domain specific automatic text summarization using fuzzy logic ",International Journal of Computer Engineering and Technology (IJCET), ISSN 0976- 6367(Print), ISSN 0976 - 6375(Online) Volume 4, Issue 4, July-August (2013).

[5] Róbert Móro, Mária Bieliková "Personalized Text Summarization Based on Important Terms Identification ",2012 23rd International Workshop on Database and Expert Systems Applications , 1529-4188, 2012 IEEE.

[6] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan "Fuzzy Genetic Semantic Based Text Summarization ", 2011 Ninth International Conference on Dependable, Autonomic and Secure Computing ,978-0-7695-4612-4 ,2011 IEEE .

[7] ZHANG Pei-ying , LI Cun-he , "Automatic text summarization based on sentences clustering and extraction ",978-1-4244-4520-2 ,2009 IEEE .

[8] Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy; "Optimizing Text Summarization Based on Fuzzy Logic ",Seventh IEEE/ACIS International Conference on Computer and Information Science ,978- 0-7695-3131-1 ,2008.

[9] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahla,"Feature-Based Sentence Extraction Using Fuzzy Inference rules ",2009 International Conference on Signal Processing Systems ,978-0-7695-3654-5,2009 IEEE .

[10] Ladda Suanmali, Mohammed Salem Binwahlan and Naomie Salim "Sentence Features Fusion for Text Summarization Using Fuzzy Logic ",2009 Ninth International Conference on Hybrid Intelligent Systems ,978-0-7695- 3745-0 ,2009 IEEE.