International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

Automatic Text Summarization and Categorization: Fuzzy Approach

Krisha Mistry[1], Maulik Panchal[2], Varun Varier[3], Panjab Mane[4]

[1-3]B.E. Student, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India.
[4]Professor, Dept. of IT Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India.

Abstract - *Automatic Text Summarization is undergoing comprehensive research and is growing in significance as the online information available is going on increasing. Automatic Summarization of Text is to compact the original large text into a shorter text called as Summary. Automatic summarization of Text consists of two approaches: abstraction and extraction. This paper centers around extraction approach. The main objective of automatic text summarization depending on extraction method is selection and determining of sentences on the bases of various features and their scores. Various documentations like file, pdf, word documents are used as input. The input document then will be preprocessed and will determine to which category input belongs to then the five features will be used to calculate the scores of each sentence. In this paper we have proposed the Fuzzy Logic method for improving the extraction of summary sentences and have used the NLP for better optimization during the preprocessing method and also improves the performance of text categorization method.*

Key Words: **Automatic Text Summarization, Fuzzy Logic, Sentence Score.**

1. INTRODUCTION

Today, in the current era with an increasing growth in complexity and quantity of information available on internet it has become very much important for providing the consumer with an enhanced method for finding reliable and exact information from available data on internet. Text Summarization has become timely and an important tool for supporting and understanding the vast volumes of text that are accessible in documents.

The main goal of Automatic Text Summarization is to present the most relevant information in a compact and condensed version of the original text without changing the actual meaning of the original text and help the user understand the large quantities of information quickly. Automatic Text Summarization is substantially different from human based text summarization because humans can catch and convey the deep meaning and themes of text documents whereas it is quite difficult to incorporate automating such skills. The problem of choosing the most relevant portion of text as well as the problem of producing coherent summaries can be easily solved by Automatic Text Summarization.

The Automatic Text Summarization can be categorized in two ways: abstraction summarization and extraction summarization. Extraction summarization is versatile and takes less time in comparison with abstraction summarization. Extraction is the process of selecting the phrases or sentences having the maximum from the original document and put it together to form a new short and a compact text without changing any meaning of the original document. The abstraction summary uses linguistic techniques to analyse and interpret the text. For a precise summary output, most of the existing text summarization works best on well-structured documents like news, interview, posts and science papers. The extraction summarization includes defining features such as length of sentences, position of sentences, frequency of phrases, numbers of terms appearing in title, number of correct nouns and thematic words. In our project we are using a feature fusion method to find which of the features are more useful. In the paper we have provided Automatic Text Summarization using Fuzzy Logic to have an accurate summary.

2. LITERATURE REVIEW

In earlier traditional systems, the synopses were generated according to the words that repeated frequently in the text. Luhn made the first conventional system [1] in 1958. Rath et al. [2] in 1961 proposed observational confirmations for challenges in generating an efficient summary. Both research utilized one topical capacity called term frequency, along these lines they are portrayed by means of surface level techniques. In the early 1960‟s new methodologies emerged which was known as entity level approaches, the primary methodology of this sort utilized syntactic examination [3]. In [4], wherein key expressions that are utilized managed three extra segments: title, heading words, structural indicators. Right now, proposed strategy utilizes fuzzy standards and fuzzy set for selecting sentences principally dependent on their features. Fuzzy logic technique has a major advantage that is it has powerful reasoning capabilities. The [5] paper suggests that decision making ability of humans is not a traditional multi valued technique but is a combination of logic with fuzzy truths, fuzzy connectives and fuzzy rules of inference. Fuzzy set proposed by method for Zadeh [6] is a numerical gadget for adapting to vulnerability, imprecision, and uncertainty. Fuzzy logic in text summarization required multiple studies and investigation which were precisely concluded by Zadeh, Witte and Bergler [7] offered a fuzzy hypothesis based absolutely on strategy to co-reference goals and its application to content outline. Kiani and Akbarzadeh [8] proposed system for outlining printed content utilizing blend of Genetic Algorithm (GA) and Hereditary Programming (GP) to streamline rule sets and membership function of fuzzy frameworks. The trademark feature extraction procedures are utilized to procure the basic sentences inside the content. In feature extraction strategy some of the sentences have more significance and a couple

have substantially less so they must have balance weight in calculations, and we utilize fuzzy good judgment to determine this difficulty through membership function for each feature. Yan Liu et al [9] have proposed a synopsis structure by means of profound deep learning model. The system comprises of concepts extraction, summary generation and reconstruction validation. A query arranged extraction method has been packed data disseminated in numerous records to concealed units' layer by layer. At that point, the entire profound engineering was fine turned by limiting the data misfortune in reconstruction validation port. As indicated by the ideas from deep architecture, dynamic programming was utilized to look for most enlightening arrangement of sentences in the summary. Tests on three benchmark datasets show the viability of the structure and calculations. Jason Weston et al [10] have proposed a directed learning for profound models. Scientists utilized shallow designs previously indicated two different ways of to improve generalization. First is inserting unlabeled information as a different preprocessing step (i.e., first layer preparing) and the second is utilized at the output layer. All the more critically, they have summed up these ways to deal with the situation where, have train a semi-regulated supervised embedding jointly with a supervised deep multilayer architecture installing mutually with a managed profound multilayer design on any (or all) layers of the system, and indicated there could have been genuine advantages for complex errands. F. kyoomarsi et al [11] have introduced a methodology for making text summaries. Utilized fuzzy rationale and word-net, they have retrieved the most important sentences from a unique record. The methodology uses fuzzy measures and induction on the extricated printed data from the record to establish the most significant sentences.

3. PROPOSED METHODOLOGY

3.1 INPUT DATA AND PREPROCESSING

In our framework, we are first accepting input to a document, for example, docx, pdf, txt which will get put away in our database. After taking information we need to perform preprocessing. The accompanying segments present the pre-handling of the information dataset, which will give all features:

- **Sentence Segmentation:**
  Sentence division is performed by distinguishing the delimiter normally indicated by "." called as full stop. It is utilized to isolate the sentences in the document. For Example:

  Input: Delhi is India's Capital. Modi is living there.

  Output: Delhi is India's Capital

  Modi is living there.

- **Stop Word Removal:**
  Stop words removal is the way toward expelling words which don't pass on any meaning during classification procedure.

  For Example:

  (E.g. Articles {a, an, the} Prepositions {at, by, in, to, from} Conjunction {and, but, as} others {become, everywhere})

- **Stemming:**
  Stemming is a method to lessen a word to its stem or root structure. We utilize porter's stemming algorithm for this reason. A stemming algorithm lessens the word,

  For Example:

  "Playing"," Plays","Played" to the root word "play".

  The yield of this pre-preprocessing step is the contribution to produce the feature matrix

3.2 CATEGORIZATION

Text documents were retrieved which consisted of documents from different subjects. Machine Learning Techniques were then applied on these documents based on TFID equation. The Term Frequency-Inverse Document Frequency(TFID) is used for retrieving information. It states the importance of a word to the document.

Term Frequency, which gauges how oftentimes a term happens in an archive. Since each archive is diverse long, it is conceivable that a term would show up significantly more occasions in long records than shorter ones. In this way, the term recurrence is regularly separated by the document length

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

Inverse Document Frequency, which quantifies how significant a term is. While figuring TF, all terms are considered similarly significant. Anyway, it is realized that specific terms, for example, "is", "of", and "that", may be seen a great deal of times however have little significance. Thus, we need to scale up the uncommon terms.

3.3 FEATURE MATRIX

After Preprocessing, sentences are represented by a set of features. The value of feature ranges between 0 and 1 and there are eight different features as follows:

Title Feature: This element gives the proportion of the likeness between the title sentence and each other sentence of the document. This is controlled by tallying the quantity of matches between the substance words in a sentence and the words in the title. The score for this component is the proportion of the quantity of matches between a sentence and title sentence over the quantity of words in title.

$$S\_F1(S) = \frac{\text{No of title words in sentence S}}{\text{No of words in title}}$$

Sentence Length: We utilize this element to remove short sentences, for example, datelines and author names. The short sentences are not expected to have a place in the outline. Here first the length of the sentence is determined by including the quantity of words in it and afterward it is standardized. The score for the component is given by the proportion of length of sentence over the length of the longest sentence in the document.

$$S\_F1(S) = \frac{\text{No of title words in sentence S}}{\text{No of words in title}}$$

$$S\_F2(S) = \frac{\text{Length of sentence S}}{\text{Length of longest sentence in a document}}$$

Term Weight: This component utilizes the idea of term frequency which has frequently been utilized to ascertain the significance of a sentence. Here by term frequency we mean event of a term inside a report. We summarize the term frequency of all the term in a sentence. The score of this element is given by the proportion of summation of term frequencies of all terms in a sentence over the limit of summation estimations of all sentences in a report.

Weight of the sentence i is calculated by,

$$W_i = \sum_{i=1}^{k} (TF_i)$$

K is the number of words in the sentence, score for this feature is calculated by,

$$S\_F3(S) = \frac{W_i(S)}{\text{Max}[W_i(S)]_{i=1}^{N}}$$

Sentence Position: Position of the sentence in the content, chooses its significance. This component can include a few things, for example, the position of a sentence in the document, paragraph and section and so forth. For the score of this component we think about the initial 5 sentences in a record.

Feature Score is calculated by,

$$S\_F4(S) = 1st\ sentence = \frac{5}{5}; 2nd\ \ sentence = \frac{4}{5}; 3rd\ sentence = \frac{3}{5}; 4th\ sentence = \frac{2}{5}; 5th\ sentence = \frac{1}{5}; other\ \ sentences = \frac{0}{5} \quad (5)$$

Proper Noun: The sentence that contains proper noun, is a significant and it is most likely present in the summary. The score for this component is determined as the proportion of the quantity of number of proper nouns that is present in sentence over the sentence length.

$$S\_F6(S) = \frac{\text{No of proper nouns in sentence S}}{\text{Sentence Length (S)}}$$

3.4 SUMMARIZATION BASED ON FUZZY LOGIC

The objective of summarization dependent on extraction is sentence selection. Our framework comprises of the accompanying primary advances:

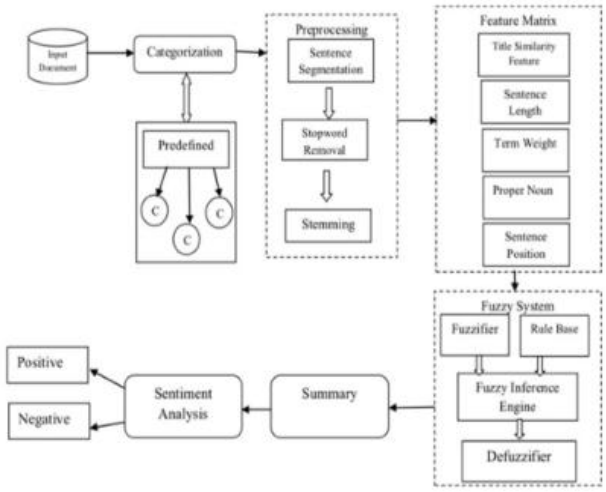
a. Read the source document into the system.

b. For preprocessing step, the framework removes the individual sentences of the document. At that point, separate the information record into singular words. Next, evacuate stop words. The last advance for preprocessing is word stemming.

c. Each sentence is related with vector of five features that depicted in segment 3.3, whose qualities are derived from the substance of the sentence.

d. The score for each sentence is retrieved from its features dependent on fluffy logic method.

e. The sentences with the highest scores are selected.



To get significant sentences we utilized fuzzy logic technique. Fuzzy logic as a rule involves choosing fuzzy inference rules and membership function. The choice of fuzzy rules and membership function influence the performance of system

3.5 GENERATE SUMMARY

In this stage, the retrieved ideal component vector set is utilized to produce the extractive summary of the document. For summary first assignment is acquiring the sentence score for each sentence of archive. Sentence score is acquired by finding the intersection point of user query with the sentence. After this progression positioning of the sentence is performed and the last arrangement of sentences for content synopsis is obtained.

- **Sentence Score:** Sentence score is the proportion of regular words found in query of client and specific sentence to the all-out number of words in the content.
- **Ranking of Sentence:** This is the last step to get the synopsis of content. Here positioning of the sentence is performed based on the sentence score got in past stage. The sentences are orchestrated from highest score to the lowest score based on the acquired sentence score. Out of these sentences top-N sentences are chosen based on compression rate given by the client. To discover number of top sentences to choose from the framework we utilize following formula dependent on the compression rate

4. RESULT AND ANALYSIS

The produced synopsis is then assessed dependent on the Evaluation measurements, for example, Recall, Precision and F-Measure. The maximum Recall, Precision and F-Measure values for the current document is giving as 0.37, 0.86 and 0.50 respectively

5. CONCLUSION

As the information is expanding at an exponential rate, Automatic text summarization is fundamental in time bound circumstances and recovery of precise content .In this system we have extracted five features namely, title feature, sentence length, term weight, sentence position, proper noun. This feature matrix is applied to our proposed work which associates with fuzzy logic. We applied our method for single document summarization which could be extended for multi-document summarization. The input is a text document which belongs to various fields and can be categorized as a document from Artificial Intelligence, Machine Learning etc. Fuzzy inference rules are important to generate the best summaries. Our technique could be reached out for automatic determination of fuzzy inference rules.

REFERENCES

1. H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, pp.159165.1958.
2. G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences" American Documentation, vol. 12, pp.139-143.1961
3. Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization (MIT Press. 1999)
4. H. P. Edmundson., "New methods in automatic extracting" Journal of the Association for Computing Machinery 16 (2). pp.264-285.1969

A. D. Kulkarni and D. C. Cavanaugh, "Fuzzy Neural Network Models for Classification" Applied Intelligence 12, pp.207-215. 2000.

5. L. Zadeh, "Fuzzy sets. Information Control" vol. 8, pp.338–353.1965.
6. R Witte and S. Bergler, "Fuzzy coreference resolution for summarization" In Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS). Venice, Italy: Università Ca‟ Foscari. pp.43–50. 2003.
7. Arman Kiani and M.R. Akbarzadeh, "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP" In Proceedings of 2006 IEEE International Conference on Fuzzy Systems, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada. pp.977-983.2006.

8. Yan Liu,Sheng-Hua Zhong,Wen-Jie Lii "Query Oriented Multi Document Text Summarization via deep learning, "Elsevier Science,2008,PP.3306-3309
9. Jason Weston, Frederic Ratle and Ronan Collobert"," Deep Learning via semi-supervised embedding, "International Conference on Machine Learning,2008, PP.1168-1175
10. F.kyoomarsi, H.khosravi, E.eslami, and M davoudi, "Extraction based summarization using fuzzy analysis", "Iranian Journal of Fuzzy systems,Vol.7,No.3,2010,PP.15-32
11. Laddasunmali, Naomie Salim, Mohamed Salem Binwahlan," Automatic Text Summarization using feature based fuzzy extraction," Jurnal Teknologi Maklumat, vol.20, No.2, Dec 2008,PP.105-115.