iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44848

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Analysis on Text Summarization

Akash Divekar
(MCA, Institute of Management & Computer Studies (IMCOST) / Mumbai University, India)

Abstract: ***As we enter the 21st century, with the advent of mobile phones and access to information stores, we seem to be surrounded by more information, less time, or the ability to process it. The creation of automated summaries was a clever human solution to this complex problem. However, the application of this solution was very complicated. In fact, there are a number of problems that need to be addressed before the promises of an automated text can be fully realized. Basically, it is necessary to understand how people summarize the text and build a system based on that. However, people are different in their thinking and interpretation that it is difficult to make a "gold standard" summary in which product summaries will be tested. In this paper, we will discuss the basic concepts of this article by providing the most appropriate definitions, characterization, types and two different methods of automatic text abstraction: extraction and extraction. Special attention is given to the method of extraction. It consists of selecting sentences and paragraphs that are important in the original text and combining them into a short form. It is mentally simple and easy to use.***
Keywords: Abstractive approach, Automatic text summarization, Extractive approach, Natural language processing, Text summarization*.*

I. INTRODUCTION

The rapid emergence of the WWW has made dozens of articles on a variety of topics available to users [1] [2]. In order to use these texts effectively, you need to be able to get a summary of them. However, it is very difficult for people to make a handwritten summary of all available text. Automatic Text Summary (ATS) provides a solution to this problem of overload [2]. Therefore, ATS has become an important and timely tool for the user to quickly understand the vast amount of information [3]. Automated summaries included in the field of language processing, the process of dealing with a large amount of information by integration is only important. It often happens in everyday communication and is an important and professional skill for some people. Automatic abstraction of text is intended to provide a concise representation of the content according to the information the user wants to obtain [4]. With a summary of the document available, users can easily determine its compatibility with their interests and find the desired documentation with very few mental responsibilities involved. [5]. In addition, the goal of automated text summaries is to summarize documents into a shorter version and to retain important content [3].

Text summarizing methods can be divided into two main methods of abbreviating and abbreviating [6]. Automatic text summarization is a method that compresses a large text into a short text that encompasses important information. The computer program renders the text and returns the summary of the original text. This is done by reducing text duplication and by extracting text content [9]. Generally, the summary should be much shorter than the source text. This feature is defined by the compression rate, which measures the average length of the summary and the length of the original text [3]. At present automated text abstraction has benefited from the expertise of many fields of research: information retrieval and extraction, natural language production, speech studies, machine learning and technical studies used by professional summaries [7]. One needs a summary especially because it reduces reading time and makes the selection process easier during the text process search.

II. METHODOLOGY

Technique used for Text Summarization
Importing files:

1) punkt (Punkt Sentence Tokenizer)

This tokenizer separates a book into a neglected of sentences, by utilizing an unaided calculation to produce a model for shortened form words, collocations, and words that start sentences. It essential to be prepared on a huge range of plaintext in the objective language before it very well may be used.

The NLTK information bundle includes a pre-prepared Punkt tokenizer for English.

2) stopwords

Natural Language Processing with PythonNatural language processing (nlp) is a research field that presents many challenges such as natural language understanding.

Text may contain stop words like 'the', 'is', 'are'. Stop words can be clean from the text to be managed. There is no worldwide list of representation, the score is computed by aggregating the evidence from different weighted indicators. Stop words in nlp research, however, the nltk component comprises a list of stop words.

You didn't write that awful page. You're just trying to get around data available of it. Beautiful Soup is here to help. Since 2004, it's been saving programmers hours or days of work on quick-turnaround screen scraping projects. Beautiful Soup is a Python library considered for quick improvement projects like

Depending upon the number of documents accepted as input by a summarization process, automatic text summarization can be categorized as single document summarization and multi-document summarization as shown in Fig. 1 below.

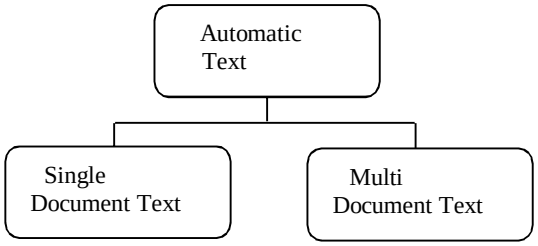


Figure 1. Automatic Text Summarization Models

In the model Single Document Text Summarization, a summary is produced from single input document. The single document summarization process flow can be depicted in Fig. 2.

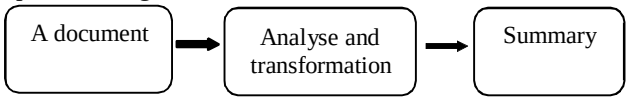


Figure 2. Single Document Text Summarization

However, in Multi Document Text Summarization, a summary is produced from multiple input documents dealing with the same topic as illustrate in Fig. 3.

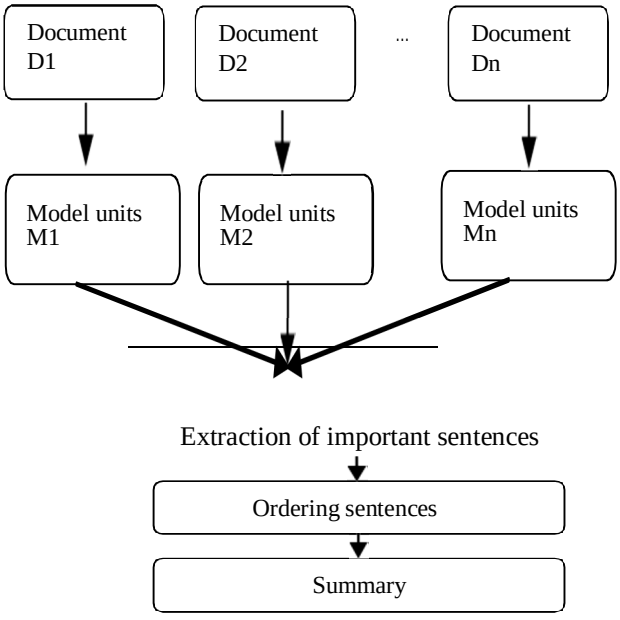


Figure 3. Multi Document Text Summarization

In 1995, Radev and McKeown [13] were the first to develop a system for producing abstract texts. Multi-document summarization is one of the major challenges in the current system of summaries because the task of summarizing multiple texts is much more difficult than the task of summarizing single texts where rewriting [1] is a major problem in summarizing multiple texts.
Summary output can be of two types: Output abbreviations and tense abbreviations. Summary of quotations is generated by extracting all the sentences from the source text. The value of sentences is determined by the mathematical and linguistic features of the sentences [9]. Intense summaries are produced by rearranging the sentences in the source text. The goal of paraphrasing includes understanding key concepts in a text and conveying those ideas in clear natural language. It uses language techniques to explore and interpret text and finds new concepts and terminology that best describes it by producing a new short text that conveys the most important information from the original text [9].

A. Extractive Method

The purpose of the exclusion method is to construct a summary by extracting key sentences from the original text [2]. The extracted sentences will then be compiled to produce a summary and to maintain order as in the original text and without modifying the source text [11]. Much of the work in text summaries focuses on quoted summaries because they are conceptually simple and easy to process. In general, there are three types of sentence extraction in the production of a summary: mathematics, method of language learning and machinery [10].

1) Statistical Method

With Statistical ingenuity, the summary is made without comprehension, but instead relies on the Statistical distribution of certain structures [10]. This process aims to find the meanings of key words and determines the value of a sentence by the total weight of the sentence it contains [5].

a) Pre-processing: This is the first step in uploading a given text to a proposed program and splitting it into a compound sentence (it takes raw text as input and uses basic methods to modify or eliminate unused text features to further process text data). Familiarity is a way of turning text into a normal process by performing procedures, such as collecting cases, making tokens, stopping deletions and stems. Therefore, the steps for Preliminary Process are [2] [18]:

- Case-Folding: is the process of converting a given text into lowercase text to avoid duplication of the same word in different contexts. This helps the system to distinguish common words and improve its accuracy [2] [18].
- Tokenization: is the process of dividing a text into each sentence. In a sentence breaker, a dot is considered a comma and a word space is considered [2] [18].
- Stop word removal: the process of removing stop words, that is, words with less semantic knowledge. The most common words and appear in most of the text but do not include much semantic information are called stop words, such as: “the”, “by”, “a”, “an”, etc.
- Separation is based solely on the terms of the element and not on full stops, commas, colonies, semicolons, etc. They are therefore removed from the document and will not be saved in the signature file to continue the process [2] [18].
- Stemming: The purpose of this process is to find the stem or radix of each word (usually, the text of the text contains the repetition of the same word in variety), which emphasizes its semantics [15]. It deals with the same words syntactically, such as plurals, variations of words, etc. [15]. The purpose of this process is to determine the stem or radix of each word, emphasizing its semantics [15]. Stemming can be of two types [2]:

- Derivational Stemming.
- Inflectional Stemming.

Derivational title creates new words from existing names, e.g., “Finish-Last”, “Useful Use”, “Music Music”, etc. However, the subject of flexibility includes words that are common in grammatical variations such as past or present tense or singular or plural, e.g. [2] [18].

b) Analysis:

This phase is traditionally divided into three steps [2] [18]:
- Level: Conceptual analysis concept using summaries.
- Selection: Conversion using the "Mathematical Tasks" function.

- Ordering: ordering new statements to make an understandable summary.

- Methods of Statistical Technique: Scoring is the process of assigning points for each sentence to determine its significance in summary [2]. Summary of text identifies and extracts key sentences from source text and combines them to form a brief summary. The significance of a sentence can be determined in a number of ways, such as:
- TF-IDF Method (Frequent Term Term-Frequency Document): This method was introduced in 1989 [19]. The term frequency (TF) has the same effect as the frequency of words is high. However, the frequency of the opposite document (IDF) considers low frequency words to give the opposite of the higher value in the scale [19]. The purpose of tf-idf is to reduce the weight of frequently repeated words by comparing their measurement frequency in the document collection. This structure has made tf-idf one of the most widely used terms in the extracting text [14].
- Cue-phrase method: Words can have a positive or negative effect on the weight of a sentence to indicate importance or main idea [3], such as indicators: “in summary”, “concluding”, “paper explains”, “importantly”.
- Title Method: This method states that sentences from a topic are considered very important and are more likely to be summarized. Sentence points are calculated by how many words are often used between sentence and subject. The title method cannot work if the document does not include any title information [12].
- Location mode: It depends on the idea that key sentences are located somewhere in the text or paragraph, such as the beginning or the end of a paragraph [3]. Therefore, important information in the text is often covered by the authors at the beginning of the article. The first sentences are therefore thought to contain the most important content [11].
- Length of sentence: Very short sentences are rarely included in the summary as they convey little information. Extremely long sentences are also not appropriate to represent a summary [20].
- Proper noun: Sentences that contain a proper name representing a unique business such as a person's name, organization or place are considered important in a document [20] [14].

2) Linguistic Approach

This process incorporates grammatical knowledge so that the computer can analyze sentences in chronological order and then decide which sentences to choose based on subject position, verb and noun [10]. It is much harder than mathematical methods.

a) *Machine Learning Approach:* The Machine Learning (ML) method is useful where a set of documents and related references are available [15]. ML aims to learn from the training model in order to find the right class where the object is part. The sentences in each text will be represented by using the vectors of the elements extracted from the text [15] [14]. Therefore, the goal of the training model is to divide sentences into two categories: a sentence labeled as an “abbreviated sentence” when it is part of a reference summary or as an “unabridged sentence” otherwise. This process of learning from text collection and its abbreviations allows the use of a trained model to produce a summary that is released when a new document is presented in the system [14]. Other ML methods used in a single document will be described.

b) *Text Summarization with Neural Networks:* This method involves training the neural network to identify the type of sentences to be included in the summary. The neural network learns important patterns in sentences and should be included in the summary. Typically, this approach uses Feed forward neural network architecture with three layers [11].

c) *Text Summarization with Naive Bayes:* One of the first integrated machine learning activities was the use of the Naive Bayes data learning phase in 1995 [14]. In this method, the naïve-bayes classification function is used to classify each sentence as suitable for exclusion or not [16] [17].

B. Abstractive Method

The abstract text abstractive method is intended to produce important information about the text in a new way, by translating and examining the source text and creating a short, close summary of what one can produce. The summary will contain compressed sentences or may include novel sentences that are not explicit in the original source text [21] [22] [23]. It produces a live summary with a very clear and precise logical structure compared to the abbreviations produced by the extraction method [12]. However, this approach is difficult because it uses a language approach to understand the original text [12] and requires a deeper understanding of NLP activities. It is broadly divided into two categories: a structural-based approach and a Semantic-based approach [3].

1) Structured Approach

The construction-based approach incorporates the most important information from documents through cognitive schemes [3] [11]. Different methods can be used by the Structured Based Approach, such as the Tree based method, the template-based method, the ontology-based method, the lead-and-body method and the Law-based method [3] as shown in Figure 4.

2) Semantic-Based Approach

In a Semantic-based approach, the semantic representation of texts is used to supply the native language production system (NLG). This approach focuses on identifying noun phrases and verbs by processing language data [3] [11]. Different methods can be used Based on Multimodal semantic Model, Information-based method and Semantic-based approach [3] as shown in Figure 4 below.

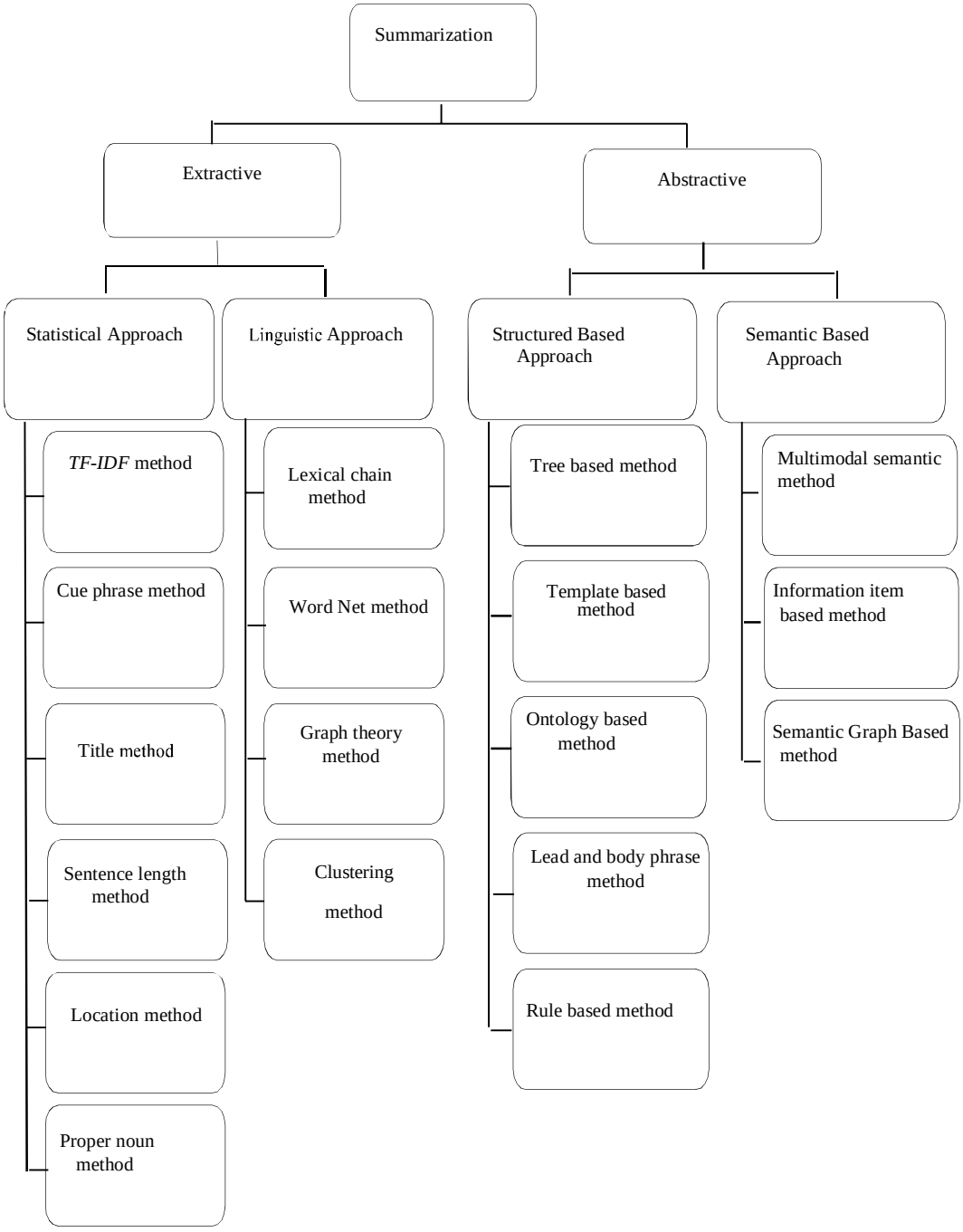


Figure 4. Principles Approaches used in Automatic Text Summarization

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

III. ALGORITHM

A. Convert the Paragraph to Sentences.
Wherever we get the dot(.). We split the sentence.

B. Text Pre-processing.
After converting paragraph to sentences, we need to remove all the special characters, stop words and numbers from all the sentences.

C. Tokenize the Sentence.
We need to tokenize all the sentences to get all the words that exist in the sentences.

D. Find Weight Frequency of the occurrence.
We can find the weighted frequency of each word by dividing its frequency by the frequency of the most occurring word.

E. Replace Words by Weighted Frequency in Original Sentences.
The final step is to plug the weighted frequency in place of the corresponding words in original sentences and find their sum. It is important to mention that weighted frequency for the words removed during prepossessing (stop words, punctuation, digits etc.) will be zero and therefore is not required to be added

F. Sort Sentences in Descending Order of Sum
The final step is to sort the sentences in inverse order of their sum. The sentences with highest frequencies summarize the text

IV. APPLICATION AREA

These are some use cases where Text summarization can be used across the enterprise:

A. Media Monitoring
The problem of information overload and “content shock” has been widely discussed. trading. When you are a financial analyst looking at market reports and news everyday, you will inevitably hit a wall and won’t be able to read everything. Summarization systems tailored to financial documents like earning reports and financial news can help analysts quickly derive market signals from content.
Summarization presents an opportunity to condense the continuous torrent of information into smaller pieces of information.

B. Newsletters
Many weekly newsletters take the form of an introduction followed by a curated selection of relevant articles. Summarization would allow organizations to further enrich newsletters with a stream of summaries (versus a list of links), which can be a particularly convenient format in mobile.

C. Search Marketing and SEO
When evaluating search queries for SEO, it is critical to have a well-rounded understanding of what your competitors are talking about in their content. This has become particularly important since Google updated its algorithm and shifted

V. CURRENT IMPLEMENTATION

Currently implementation is single documented as we are working on a single webpage only search results, understand shared themes and skim the most important points.

A. Internal Document Workflow
Large companies are constantly producing internal knowledge, which frequently gets stored and under-used in databases as unstructured data. These companies should embrace tools that let them re-use already existing knowledge. Summarization can enable analysts to quickly understand everything the company has already done in a given subject, and quickly assemble reports that incorporate different points of view.

B. Financial Research

Investment banking firms spend large amounts of money acquiring information to drive their decision-making, including automated stock

C. Legal Contract Analysis

Related to point 4 (internal document workflow), more specific summarization systems could be developed to analyze legal documents. In this case, a summarizer might add value by condensing a contract to the riskier clauses, or help you compare agreements.

VI. CONCLUSION

Today, the need for automated text summaries has grown due to the rapid increase in the amount of information online. Therefore, it is very difficult for users to physically shorten those big online documents. Automatic text summation solves this problem. It represents one of the applications that is processed in the native language and is becoming increasingly popular in information processing. Allows you to access important information while working with a large collection of documents. A good automatic summary captures the essence of a long work with a short instructive statement that can be read and typed quickly. This solution can be developed using quotation or quotation methods intended to analyze texts and general summaries. Summary of the text in a thought-provoking way is powerful because it produces a symmetrical-related summary that is difficult to produce. However, summarizing text in an easy-to-read way is easier for a person to edit and the computer understands.

This review focuses on the concepts and basic techniques associated with automated text summaries and their most important features. Therefore, many discussions revolve around the extraction method due to its widespread use. However, there are a number of limitations associated with this method, that is, its sentences may be omitted from the context and the analogy references may be violated. Therefore, the main purpose of this research project is to understand the process of summarizing the text in order to develop an automated text summary system with greater accuracy as a future work. This goal can be achieved by using a mixed mathematical method.

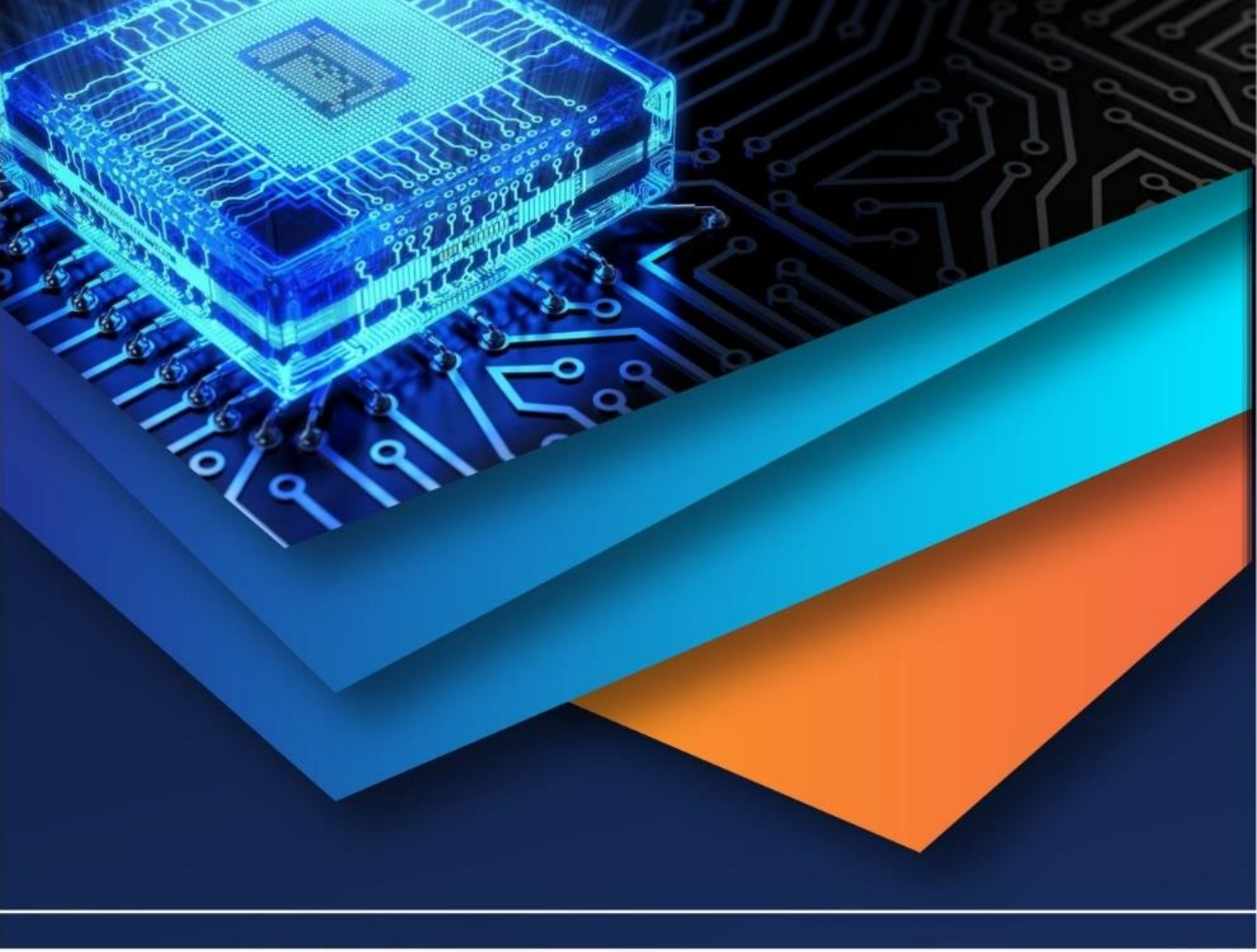
VII. ACKNOWLEDGEMENTS

We thank the Head of Department and all the members of the Master of Computer Application department of the Institute of Management & Computer Studies (IMCOST) for the encouragement that helped us to complete this review paper.

REFERENCES

[1] M. Gambhir and V. Gupta, “Recent automatic text summarization techniques : a survey,” Artif. Intell. Rev. Springer Sci. Media Dordr., vol. 47, no. 1, pp. 1–66, 2016.
[2] T. P. Sariki, B. Kumar, and R. Ragala, “Effective classroom presentation generation using text summarization,” Comput. Technol. Appl., vol. 5, no. August, pp. 1–5, 2014.
[3] A. Khan and Naomie Salim, “A Review On Abstractive Summarization Methos,” J. Theor. Appl. Inf. Technol., vol. 59, no. 1, 2014.
[4] F. Kiyoumarsi, “Evaluation of Automatic Text Summarizations based on Human Summaries,” Procedia - Soc. Behav. Sci., vol. 192, pp. 83–91, 2015.
[5] M. Chandra, V. Gupta, and S. K. Paul, “A Statistical Approach for Automatic Text Summarization by Extraction,” in International Conference on Communication Systems and Network Technologies, 2011, pp. 268–271.
[6] V. Gupta, Gurpreet Singh Lehal, and G. S. Lehal, “A Survey of Text Summarization Extractive techniques,” J. Emerg. Technol. Web Intell., vol. 2, no. 3, pp. 258–268, 2010.
[7] J.-M. Torres-Moreno, Automatic Text Summarization. British Library Cataloguing-in-Publication Data. ISTE Ltd, John Wiley & Sons, Inc., 2014.
[8] D. Das and A. F. Martin, “A Survey on Automatic Text Summarization,” Lit. Surv. Lang. Stat. II course C. 4, pp. 192–195, 2007.
[9] P. Shah and N. P. Desai, “A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages,” in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016.
[10] B. M. M. Othman, M. Haggag, and M. Belal, “A Taxonomy for Text Summarization,” Inf. Sci. Technol., vol. 3, no. 1, pp. 43–50, 2014.
[11] C. S. Saranyamol and L. Sindhu, “A Survey on Automatic Text Summarization,” Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 6, pp. 7889–7893, 2014.
[12] N. Munot and S. S. Govilkar, “Comparative Study of Text Summarization Methods,” Int. J. Comput. Appl., vol. 102, no. 12, pp. 33–37, 2014.
[13] L. Suanmali and N. Salim, “Literature Reviews for Multi-Document Summarization.”, 2008.
[14] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, “A Review on Automatic Text Summarization Approaches,” J. Comput. Sci., 2016.
[15] J. Neto, A. Freitas, and C. Kaestner, “Automatic Text Summarization Using a Machine Learning Approach,” Adv. Artif. Intell. Bittencourt, G. G.L. Ramalho, Springer-Verlag Berlin Heidelb., pp. 205–215, 2002.
[16] S. Suneetha, “Automatic Text Summarization: The Current State of the art,” Int. J. Sci. Adv. Technol., vol. 1, no. 9, pp. 283–293, 2011.
[17] N. Bhatia and A. Jaiswal, “Literature Review on Automatic Text Summarization: Single and Multiple Summarizations,” Int. J. Comput. Appl., vol. 117, no. 6, pp. 25–29, 2015.

[18] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," Tech. – Int. J. Comput. Sci. Commun. Technol., vol. 2, no. 1, pp. 325–335, 2009.
[19] M. Haque, S. Pervin, and Z. Begum, "Literature Review of Automatic Multiple Documents Text Summarization," Int. J. Innov. Appl. Stud., vol. 3, no. 1, pp. 121–129, 2013.
[20] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," J. Comput. Sci., vol. 8, no. 1, pp. 133– 140, 2012.
[21] M. Bhide, "Single or Multi-document Summarization Techniques," vol. 4, no. 3, pp. 375–379, 2016.
[22] S. Haiduc, J. Aponte, L. Moreno, and A. Marcus, "On the use of automated text summarization techniques for summarizing source code," Proc. - Work. Conf. Reverse Eng. WCRE, pp. 35–44, 2010.
[23] M. S. Patil, M. S. Bewoor, and S. H. Patil, "A Hybrid Approach for Extractive Document Summarization Using Machine Learning and Clustering Technique,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY