iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51815

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 11 Issue V May 2023- Available at www.ijraset.com

Automatic Text Summarization

Sakshi Jawale[1], Pranit londhe[2], Prajwali Kadam[3], Sarika Jadhav[4], Rushikesh Kolekar[5]
[1, 2, 3, 4, 5]*Information Technology Department, Parvatibai Genba Moze College of Engineering*

Abstract: ***Text Summarization is a Natural Language Processing (NLP) method that extracts and collects data from the source and summarizes it. Text summarization has become a requirement for many applications since manually summarizing vast amounts of information is difficult, especially with the expanding magnitude of data. Financial research, search engine optimization, media monitoring, question-answering bots, and document analysis all benefit from text summarization. This paper extensively addresses several summarizing strategies depending on intent, volume of data, and outcome. Our aim is to evaluate and convey an abstract viewpoint of the present scenario research work for text summarization.***
Keywords: ***Natural Language Processing, Text Summarization, Abstractive Summary, Extractive Summary.***

I. INTRODUCTION

To summarize a piece of writing is to present the main points in a concise form. Work on automated text summarization began over 40 years ago [1]. The growth of the Internet invigorated this work in recent years [2], and summarization systems are be-ginning to be applied in areas such as healthcare and digital libraries [3]. Several commercially available text summarizers are now on the market. Examples include Capito from Semiotis, Inxight's summarizer, the Brevity summarizer from LexTek International, the Copernic summarizer, Text Analyst from Mega puter, and Whis-key™ from Conver speech. These programs work by automatically extracting selected sentences from a piece of writing.

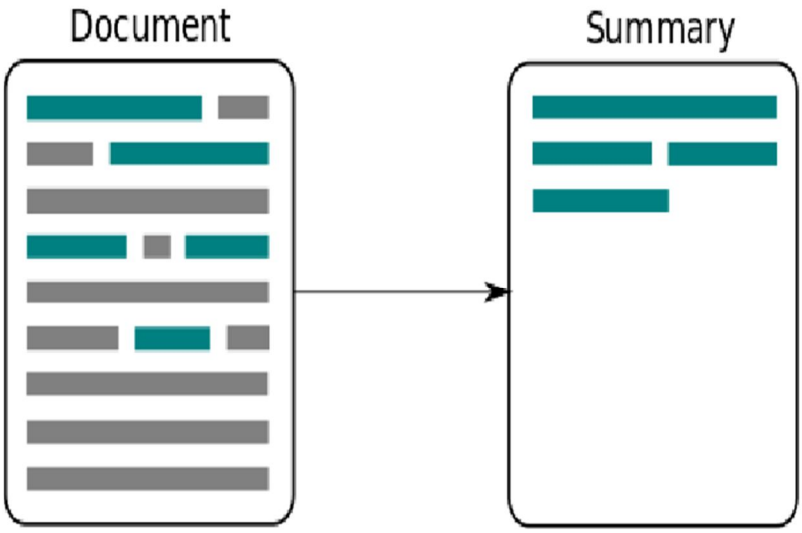


Fig. 1: generating summary from input document

II. LITERATURE REVIEW OF EXISTING SURVEY

We have investigated the existing surveys of the ATS domain, and a few of them are presented to prove the significance of this paper. Most surveys covered the former methods and research on ATS. However, recent trends, applicability, effects, limitations, and challenges of ATS techniques were not present. Table 1 summarizes and compares the existing survey on ATS. Mishra et al. [5] reviewed (2000-2013) years of studies and found some methods such as hybrid statistical and ML approaches. The researchers did not include cognitive aspects or evaluations of the impact of ATS. Allahyari et al. [8] investigated different processes such as topic representation, frequency-driven, graph-based, and machine learning methods for ATS. This research only includes the frequently used strategies. El-Kassas et al. [10] described graph-based, fuzzy logic-based, concept-oriented, ML approaches, etc., with their advantages or disadvantages.

Reference	Year	Main Purpose	Limitations
[10]	2009	Critical ways to summarize the texts and provides a taxonomy of the methods of summarization.	Extractive, abstract, NLP with Machine learning and Deep learning are missed.
[11]	2014	A hybrid approach can do both the extractive and abstractive methods efficiently.	Avoided complex processing like NLP approaches.
[12]	2014	Reviewed works between (2000-2013) year and proposed a hybrid statistical method.	This paper doesn't include cognitive aspects, including visual-ization techniques and evaluations of the impact.
[13]	2016	Describes two definite summarization techniques, which are abstractive, extractive.	Introduces techniques and methods only.
[14]	2017	A study based on automated keyword extraction and text summarization.	They do not briefly review every approach they included; they missed some feature extraction model.
[15]	2017	Topic Representation, frequency-driven, graph-based, and the effectiveness and Limitations.	The recent approaches are not surveyed.
[16]	2017	Processes of extractive methods and multilingual text summa-rization are discussed.	A precise classification and idea about feature scores and extraction is missing.
[17]	2020	Methods, processes, primary structure, strategies, datasets, measurements of ATS.	A detailed classification and description of feature extraction are missing.
[18]	2020	To handle multi- documents for summarization based on recent research work and comparison.	Does not represent any brief discussion of any topic.

This research did not include abstractive or hybrid techniques provided a taxonomy of text summarization methods and a variety of techniques. Although the author has covered some time consuming processes of ATS, recent, more efficient methods such as machine learning were missed. Abualigah et al. [18] conducted research on how to handle multiple documents and massive web data for text summarization

III. AUTOMATIC TEXT SUMMARIZATION APPROACHES

A. *Extractive Text Summarization*

1) Splitting the source document into sentences and then create an intermediate representation of the text which highlights the task. Intermediate representation has two main types, such as Indicator representation and topic representation [15] .

2) Assigning scores in each sentence for specifying their importance depending on their performance after the representation creation. Topic representation scores on the topic words the text content. On the other hand, indicator representation scores depend on the features of the sentence

3) Selecting the highest-scoring sentences to form the summary In extractive text summarization, two approaches of machine learning are applied - supervised and unsupervised machine.

a) *Supervised Learning Methods:* In supervised learning methods, the first step is to learn how to label documents by training to identify summarized and non-summarized documents.

b) *Unsupervised Learning Methods:* With unsupervised learning methods, the summarization process can be performed without any help, such as selecting the introductory sentences of the document from the user. These methods only require advanced algorithms such as graph-based, concept-based, fuzzy logic, and latent semantics to take user input and work automatically . These approaches are beneficial for extensive data.

B. *Abstractive Text Summarization*

Abstractive text summarization is the development and automation of the traditional method of text summarization . The abstractive process identifies key sections and the main ideas of a text document by paraphrasing them. The abstractive summarization process follows some common steps as follows:

1) Analyzing main contents from the text documents utilizing a vocabulary set different from the source

2) Paraphrasing the relevant data that fit in the semantics for creating a summary which contains all the actual points of the source document utilizing NLP models The abstractive summarization approaches are of two types, one is a structure-based approach, and the other one is a semantic-based approach. A brief discussion of these two types based on NLPs is provided below:

a) *Structure-Based Methods:* The structure-based approach continuously filters the most critical data from documents by applying abstract or cognitive algorithms. The algorithms for tree-based, template-based ontology, rule-based ontology are the most commonly used .
b) *Semantic-Based Methods:* The semantic-based approach attempts to refine the sentences by implementing the NLP on the entire document. This approach can easily find the noun and verb phrases using some methods.

IV. PRE-PROCESSING TECHNIQUES

IN ATS Several pre-processing are performed to clean the noisy and unfiltered text. Erroneous messages and chats, including slang or trash phrases, are known as ‘‘noisy’’ and ‘‘unfiltered text’’. The approaches mentioned below appear to be some of the most often utilized pre-processing procedures:

1) *Parts Of Speech (POS) Tagging:* The technique of grouping or organizing text words according to speech categories such as nouns, verbs, adverbs, adjectives, etc., is known as speech tagging .
2) *Stop Word Filtering:* Based on the context, stop words are screened out either before or after textual analysis. A, an, and by are illustrations of stop words that can be analyzed and eliminated from plain text .
3) *Stemming:* Stemming eliminates inflections and derivative forms to a set of words categorized as primary or root forms. By using linguistic strategies such as affixation, text stemming transforms words to consider different word forms .
4) *Named Entity Recognition (NER):* Words in the input text are recognized as names of items (i.e., person name, location name, company name, etc.) .
5) *Tokenization:* Tokenization is a text pre-processing technique that divides text flows into tokens, which can be words, phrases, symbols, or other meaningful pieces. The goal of this technique is to examine the words in a document.
6) *Capitalization:* Diverse capitalization in different documents can be problematic and thus requires to convert every letter into lowercase letters in a document. All text and document words are then merged into a single feature space using this method .
7) *Slang and Abbreviation:* Slang and abbreviation are two different types of text anomalies that are addressed in the pre-processing stage. A support vector machine is an acronym a shortened form of a word or phrase made up mainly of the first letter of the terms.
8) *Noise Removal:* Most textual data contain many more characters, such as punctuation and special characters. While important punctuation and special characters are required for human interpretation of documents, they can cause problems with classification algorithms .

V. FEATURE EXTRACTION

In Ats Feature extraction is a technique for discovering topic sentences, essential data traits or attributes from the source documents. ATS follows two phases to locate the important sentences in the text: extracting features and text representation approach. This section describes the most often used extraction features and text representation approaches for generating sentences for text summarization.

Features Collecting the essential features is the first phase of the feature extraction process. It is necessary to represent the sentences as vectors or score them to find a vital sentence from a document. Some features are used as attributes to define the text for this task. The most prevalent features for calculating the score of a sentence and indicating the degree to which it belongs to a summary are given below:

1) *Term Frequency (TF):* The TF metric is used to determine the importance of terms in a single document . As one of the most fundamental properties of ATS, it is commonly employed to represent a word’s weight.
2) *Term Frequency-Inverse Sentence Frequency (TF-ISF):* The most relevant feature extraction approach based on the text summarization survey measures the term frequency-inverse sentence frequency amongst the sentences in all documents [175]. The weights, which seem to be reasonable indications for meaningful sentences, are generated using this method. Calculating is a quick and straightforward process.
3) *Position Feature:* It is usually considered that the beginning and last sentences would provide more information about the document. Researchers have such a better chance of being included in the summary as a result of this. The feature’s binary or regressive score value could be anywhere from [0.1] [176].
4) *Length Feature:* A sentence’s length can indicate whether it is summary-worthy. In summation, it may be wrong to assume that a sentence is worthy of mention based on its length. Compared to the size of other sentences in the source material, very long and comparatively short sentences are usually not included in the summary [177].

Text Representation

The text representation models are now utilized to represent the input documents in a better shape. In NLP, text representation approaches imply translating words into numbers so that computers can comprehend and decode patterns within a language. Generally, these approaches develop a connection between the chosen phrase and the context word from the document. Some popular text presentation methods such as bag-of-words, n-gram, and word embedding are discussed below:

1) *N-gram:* N-gram is an ideal approach for multi-language operations because it does not require any linguistic preparation. An n-gram is a collection of words or characters with N components.
2) *Bag of Words (BoW):* The most primitive sort of numerical text representation is the bag-of-words model . A phrase, such as a term itself, can be expressed as a bag-of-words vector . In a text document, it is a shortened and simplified rendition of the substance of a sentence. Computer vision, NLP, Bayesian spam filters, document categorization, and information retrieval utilizing machine learning are all areas where the BoW technique is used.. The following are some of the issues related to BoW: If the new phrases include new words, the vocabulary will expand, as will the length of the vectors. Furthermore, the vectors would have a significant number of elements.
3) *Term Frequency-Inverse Document Frequency (TF-IDF:* IDF measures how important the word is, whereas Phrase Frequent (TF) measures how frequently a term appears in a text. The IDF value is needed because merely computing the TF is not sufficient to comprehend the significance of words. The inverse document frequency. Term frequency-inverse document frequency is the name given to the combination of TF and IDF (TF-IDF). However, TF-IDF has several drawbacks: it directly calculates texts' resemblance in the word-count space, which might be slow with large vocabularies. Also, it is presumed that the counts of various terms give independent evidence of similarity.
4) *Word Embedding:* Word embedding is a type of feature learning. Each word or phrase in a lexicon is mapped to an N-dimensional vector of absolute values. Various word embedding algorithms have been proposed to convert ngrams into comprehensible inputs for machine learning systems.

VI. MOTIVATION AND APPLICATION OF ATS

1) *Books or Novel Summarization:* ATS is used mainly to summarize long documents such as books, literature, or novels, as short documents are unsuitable for summarization. It is not easy to find context from short texts, whether long documents are a better summary material [19].
2) *Social Posts or Tweet Summarization:* Every day, millions of messages, posts are generated on social networking sites such as Facebook, Twitter, etc. Useful important text summarization can be achieved using ATS [20]. This valuable source of information using the ATS [20].
3) *Sentiment Analysis (SA):* The analysis of people's views, feelings, and judgments regarding events and situations is known as sentiment analysis. SA classifies emotions and mostly opinions from product reviews as ''Positive'' or ''Negative'' using fuzzy logic. ATS is quite helpful for market analysts in summarizing the feelings or thoughts of hundreds of people [21].
4) *News Summarization:* The ATS helps summarize news from many websites, such as CNN and other prominent news portals. ATS extracts the primary emphasis point of the story in a newspaper, which is sometimes used as the story's headline [22].
5) *Email Summarization:* Email communications are unstructured and not usually syntactically well-formed domains for summarization. ATS usually extracts noun phrases and generates a summary of email messages using linguistic methods, and machine learning algorithms [23].
6) *Legal Documents Summarization:* ATS discovers relevant prior instances based on legal questions and rhetorical functions to summarize a legal judgment document. A hybrid approach employs various methods, including keywords, critical phrase matching, and case-based analysis [24].

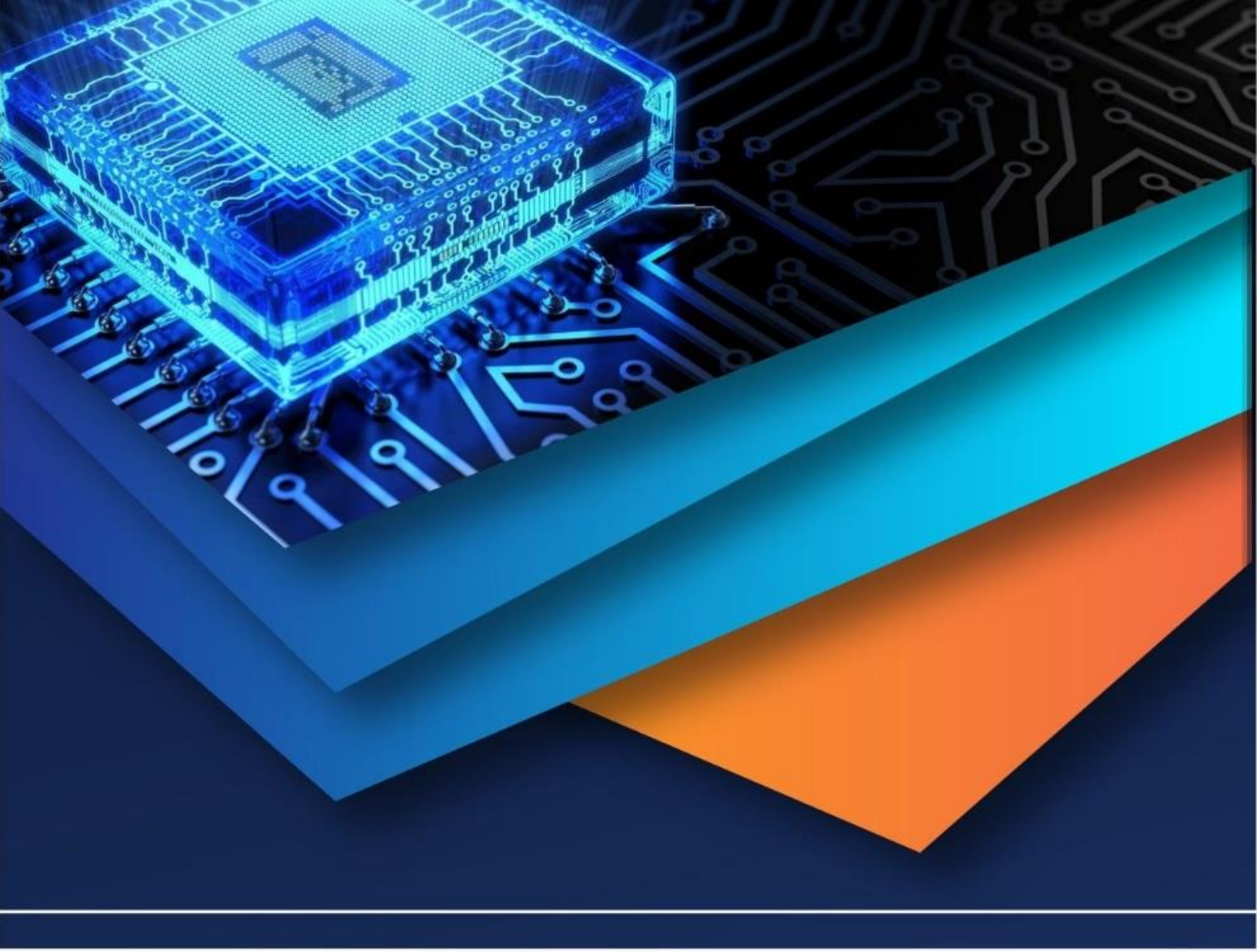
VII. CONCLUSION

Text summarization is a branch of Natural Language Processing (NLP) that focuses on shortening texts and making them more readable for users. With an excess of data accessible on the internet and the necessity to comprehend it in order to save the reader's time, text summary techniques are utilized. This paper provides a quick overview of text preprocessing, used to clean data to do effective summarization. Then it summarizes the many types of text summarizing approaches, categorizing them according to input, output, content, and purpose. The paper's primary emphasis is on extractive and abstractive text summarizing algorithms based on output. Extractive summarization summarizes by simply extracting information from the input text. Abstractive summarization is a more complicated method because it summarizes the text in its language.

The abstractive technique produces better and more semantically connected summaries. Readers would benefit significantly from an overview of the benefits and drawbacks of different techniques, as well as a concise explanation. Text summarization techniques can be applied helpfully depending on the user's needs.

REFERENCES

[1] H. P. Luhn, ‘‘The automatic creation of literature abstracts,’’ IBM J. Res. Develop., vol. 2, no. 2, pp. 159–165, Apr. 1958.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘‘Efficient estimation of word representations in vector space,’’ 2013, arXiv:1301.3781.
[3] Z. S. Harris, ‘‘Distributional structure,’’ Word, vol. 10, nos. 2–3, pp. 146–162, 1954.
[4] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, ‘‘A comprehensive survey on text summarization systems,’’ in Proc. 2nd Int. Conf. Comput. Sci. Appl., Dec. 2009, pp. 1–6.
[5] C. Saranyamol and L. Sindhu, ‘‘A survey on automatic text summarization,’’ Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 6, pp. 7889–7893, 2014.
[6] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. D. Fiol, ‘‘Text summarization in the biomedical domain: A systematic review of recent research,’’ J. Biomed. Informat., vol. 52, pp. 457–467, Dec. 2014.
[7] N. Andhale and L. A. Bewoor, ‘‘An overview of text summarization techniques,’’ in Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA), Aug. 2016, pp. 1–7.
[8] S. K. Bharti and K. S. Babu, ‘‘Automatic keyword extraction for text summarization: A survey,’’ 2017, arXiv:1704.03242.
[9] R. Mihalcea and H. Ceylan, ‘‘Explorations in automatic book summarization,’’ in Proc. 2007 joint Conf. empirical methods natural Lang. Process. Comput. natural Lang. Learn. (EMNLP-CoNLL), 2007, pp. 380–389.
[10] N. V. Kumar and M. J. Reddy, ‘‘Factual instance tweet summarization and opinion analysis of sport competition,’’ in Soft Computing and Signal Processing. Singapore: Springer, 2019, pp. 153–162.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY