Survey of "Text-Summarization using Transformer based models"

Kalyani Tonchar[1] , Sakshi Raut[2] , Madhura Peshwe[3] , Sani Rathod[4] , Rutvik Pagrut[5] ,

Parag Thakare[6]

[1,2,3,4,5]Computer Engineering Students. [6]Assistant Professor.

[1,2,3,4,5,6] Jagadambha College of Engineering and Technology, Yavatmal, Maharashtra, India.

ABSTRACT : In this research paper, we address the pressing challenges posed by the abundance of online information and the need for efficient data extraction. We advocate for automatic text summarization as a crucial solution, enabling the swift retrieval of essential insights from extensive textual documents while preserving their core meaning. Text summarization is a natural language processing (NLP) technique that aims to convert a longer piece of text into a shorter version. We explore the diverse applications of text summarization, including its role in search engines, business analysis, and market reviews, emphasizing its time-saving benefits. Our investigation encompasses various summarization approaches, from extractive methods to abstractive techniques, as well as a range of summarization techniques, from structured to linguistic analysis. Additionally, we highlight the burgeoning research in text summarization across diverse linguistic landscapes, such as Indian languages. This paper offers a comprehensive overview of text summarization evolution and contemporary significance, serving as an indispensable resource for those navigating the complexities of information management in the digital age. Automatic text summarization is a helpful tool for quickly and easily picking out the most important information from long texts. It saves time and effort by giving readers the key points without the need to read the entire text.

Keywords: Summarization, natural language processing, machine learning, extractive text summarization,abstractive text,summarization, and deep learning models.

1.INTRODUCTION

Text summarization refers to the technique of Long piece of text is converted into the short pieces of text. The aim is to create a meaningful and fluent summary having only the main points outlined in the document.In the age of information abundance, the need for effective data extraction and summarization has never been more pressing. A summary, by definition, distills vital information from one or multiple sources into a concise, digestible form, significantly reducing reading time while retaining the essence of the original content. Automatic text summarization serves as a powerful tool in this context, aiming to present source text in a shorter, semantically meaningful version.

This research paper embarks on a journey to explore the multifaceted realm of text summarization, shedding light on its significance in various domains, including news articles, email communication, mobile news updates, and information retrieval for diverse professional sectors such as business and government. The proliferation of summarization tools and systems, both proprietary and open-source, reflects the growing demand for abstracted information in today's fast-paced digital landscape.

Historically, the concept of automatic text summarization dates back to the late 1950s, with the emergence of the first automatic summarization systems. These systems excelled in selecting crucial sentences from documents, thus streamlining comprehension of voluminous content. The overarching goal of automatic text summarization is to transform lengthy documents into concise

informative summaries that capture the core content.

The methods employed in text summarization can be broadly categorized into two groups: extractive and abstractive. Extractive summarization involves the identification and extraction of significant sentences or phrases from the source text, without altering the original content. Conversely, abstractive summarization requires a deeper understanding of the text, employing linguistic techniques to interpret and rephrase the content. Abstractive methods possess the unique capability to generate entirely new sentences, enhancing focus, reducing redundancy, and achieving optimal compression rates.

In response to the escalating volume of text data on the internet, this paper presents a comprehensive framework for text summarization.Leveraging both morphological elements and semantic information, this framework offers an efficient solution for summarizing web-based content. As the demands of modern life continue to limit the time available for reading extensive materials, the importance of automated text summarization tools cannot be overstated.

Natural Language Processing (NLP) plays a pivotal role in this endeavor, enabling computers to understand and extract meaning from human language. NLP technologies empower applications ranging from automatic summarization and translation to sentiment analysis and entity recognition. By decoding the nuances of language, NLP bridges the gap between human communication and machine understanding, facilitating the development of efficient and automatic text summarization systems.

In a world where information overload is the norm, the pursuit of more efficient and accessible knowledge dissemination is not just a goal but a necessity. This research paper embarks on a quest to explore and advance the field of text summarization, offering insights and methodologies to empower users with the ability to navigate the vast ocean of information efficiently and effectively

2. LITERATURE REVIEW

The primary objective of this paper is to develop a systematic process for condensing documents by identifying significant content and summarizing it while retaining the original meaning. Text summarization involves a series of well-defined steps, including pre-processing, sentence segmentation, feature extraction, clustering, and the generation of summaries. The system primarily operates using an extractive summarization approach, which involves calculating the frequency weight-age of each word in every sentence across the entire document. Additionally, the system authenticates the parts of speech of the words and assigns a score to each sentence based on this information.

Furthermore, the paper delves into the landscape of text processing systems, which are valuable for handling unstructured textual documents from various sources. These text processing systems utilize a range of information extraction methods, such as dictionary-based approaches, pattern matching techniques, and various other strategies, to effectively process and extract valuable insights from unstructured textual data.

In a world inundated with information, the ability to condense and summarize documents without sacrificing their original meaning is of paramount importance. Text summarization, as explored in this paper, provides a structured approach to achieving this goal, making it easier for readers to access essential content while saving time and effort.

3.OBJECTIVES

The primary objective of this study was to develop a text summarization process that condenses a document by identifying and preserving significant content while maintaining the original meaning. The text summarization process involves several key steps, including pre-processing, sentence segmentation, feature extraction, the use of clustering techniques, and the generation of a summary.

This system is based on the extractive summarization approach, where it calculates the frequency weight-age of each word within a sentence throughout the entire document. It also authenticates the parts of speech of words in the sentences and assigns total scores to the sentences based on this information.

There are existing text processing systems that handle unstructured textual documents from various unstructured sources of free text. These systems use methods such as dictionary-based approaches, pattern-matching techniques, and more.

The research in this paper aims to create an automated text summarization system using an extractive approach. It involves calculating the frequency-weight age of words in sentences throughout the document, taking into consideration their parts of speech to determine sentence scores. The system employs a clustering technique, specifically k-means clustering, to extract final summary sentences. This method divides observations into clusters based on their proximity to cluster centroids, effectively partitioning the data space into Voronoi cells. To reduce intra-cluster differences, squared Euclidean distances are used instead of regular Euclidean distances. The goal is to provide users with meaningful and efficient document summaries, addressing the challenge of information overload in the digital age.

3.1 Pre-processing Technique

Various pre-processing techniques are employed to clean noisy and unfiltered text. Noisy and unfiltered text includes erroneous messages, chats, slang, or irrelevant phrases. Common pre-processing procedures include:

1.Parts Of Speech (POS) Tagging: Categorizing words based on speech categories like nouns, verbs, adverbs, adjectives, etc.

2.Stop Word Filtering: Filtering out common words (stop words) either before or after textual analysis.

3.Stemming: Reducing words to their root forms by removing inflections and derivative forms.

4.Named Entity Recognition (NER): Identifying words in the text as names of entities such as person names, locations, company names, etc.

5.Tokenization: Dividing text into tokens, which can be words, phrases, symbols, or other meaningful units.

6.Capitalization: Converting all letters to lowercase to ensure consistency.

7.Slang and Abbreviation: Addressing slang and abbreviations in the text.

8.Noise Removal: Removing unnecessary characters like punctuation and special characters.

9.Spelling Correction: Optionally correcting typos.

10.Lemmatization: Converting words to their base forms.

These pre-processing techniques aim to enhance the quality of the text data before it undergoes further analysis and summarization.

4.ADVANTAGES

Text summarization offers a multitude of advantages, including enhanced information retrieval capabilities by swiftly pinpointing essential details in lengthy documents. It significantly improves time efficiency, particularly valuable in time-sensitive industries. Furthermore, summarization makes content more accessible, catering to individuals with limited time or cognitive constraints. Decision-makers benefit from clear overviews, expediting informed choices. Language agnosticism allows the application of summarization to various languages. Summaries maintain consistency, reducing potential bias associated with human summarizers. They act as navigational aids, enabling content exploration and skimming. Multi-document summarization consolidates information from numerous sources, while customization tailors summarization to specific industries. Advances in language models continually enhance accuracy, and summaries aid in information compression, making them efficient for data storage and sharing. Summarization serves as an educational tool, providing concise overviews of complex topics, and its scalability accommodates large datasets. Furthermore, it ensures consistency in summarization quality across languages, facilitating cross-lingual research and communication.

5.CONCLUSION

In the ever-evolving information landscape, automatic text summarization, a pivotal subfield of Natural Language Processing (NLP), has witnessed a shift in focus towards addressing the information overload prevalent in domains such as bio-medicine, product reviews, education, emails, and blogs, especially in the context of the World Wide Web. This research highlights the essence of automatic summarization, which involves distilling complex texts into concise, meaningful summaries. Extractive summarization methods, rooted in selecting indicative sentences or passages, have seen extensive exploration, including approaches based on Neural Networks, Graph Theory, Fuzzy Logic, and Clustering. Additionally, both extractive and abstractive summarization approaches have been rigorously studied, with extractive methods prevailing due to their practicality, while abstractive methods, despite their potential for generating more human-like summaries, currently require substantial computational resources. As we navigate this digital era, characterized by relentless information expansion on the internet, text summarization role in providing structured, precise summaries becomes increasingly critical. Although summarization research has a history of over half a century, it remains a field ripe for innovation, having transitioned from scientific documents to diverse content types. While abstractive summarization holds promise, extraction-based methods continue to yield reliable results across various applications, effectively condensing extensive textual data into informative, accessible summaries. In conclusion, text summarization stands as an indispensable tool for efficient information processing and comprehension in a world overwhelmed by data, with avenues for further exploration, particularly in the context of Indian languages, beckoning researchers to shape the future of this dynamic field.

6.REFERENCES

1.Babar S.A. and Thorat S.A., "Improving Text Summarization using Fuzzy Logic & Latent Semantic Analysis", International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2014,Vol. 1 Issue 4.

2. M.-T. Nguyen, V. C. Tran, X. H. Nguyen, and L.-M. Nguyen, ''Web document summarization by exploiting social context with matrix cofactorization,'' Inf. Process. Manage., vol. 56, no. 3, pp. 495–515,May 2019.

3.X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, ''Extractive summarization using supervised and unsupervised learning,'' Expert Syst. Appl., vol. 133,pp. 173–181, Nov. 2019.

4. R. Glauber and D. Barreiro Claro, ''A systematic mapping study on open information extraction,'' Expert Syst. Appl., vol. 112, pp. 372–387,Dec. 2018.

5.A. Gupta, I. Banerjee, and D. L. Rubin, ''Automatic information extraction from unstructured mammography reports using distributed semantics,''

6.J. Biomed. Informat., vol. 78, pp. 78–86, Feb. 2018. M. Rajangam and C. Annamalai, ''Extractive document summarization using an adaptive, knowledge based cognitive model,'' Cognit. Syst. Res.,vol. 56, pp. 56–71, Aug. 2019.

7. G. Fuentes-Pineda and I. V. Meza-Ruiz, ''Topic discovery in massive text corpora based on min-hashing,'' Expert Syst. Appl., vol. 136, pp. 62–72,Dec. 2019.