INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1238-1242
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

# AI DOCUMENT (TEXT)SUMMARIZATION TOOL

**Saquib Shaikh[1], Syed Muhammad Fahad[2], Chaudhary Mujahid[3], Prof. Zaibunnisa L. H. Malik[4]**

[1,2,3]Department of Computer Engineering M.H. Saboo Siddik Polytechnic Mumbai, India.

[4]Guide, HOD of Computer Engineering M.H. Saboo Siddik Polytechnic Mumbai, India.

zebamalik@yahoo.com

DOI: https://www.doi.org/10.58257/IJPREMS36910

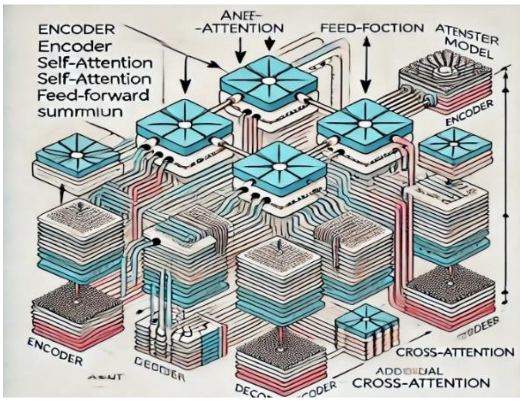
## ABSTRACT

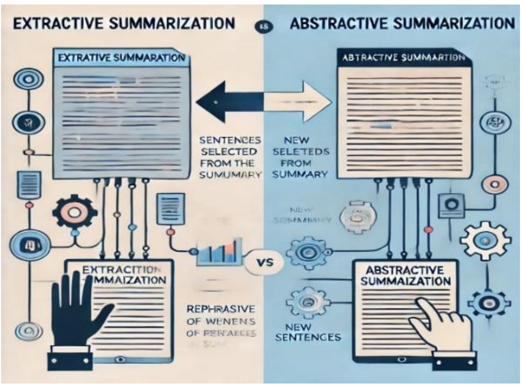
This review paper presents a comprehensive analysis of recent advancements in AI Document (text) summarization, highlighting the evolution of methodologies, datasets, and evaluation metrics used in the field. As the volume of unstructured textual data increases, automated summarization techniques have become vital for efficient information retrieval and content consumption. The paper categorizes summarization approaches into extractive and abstractive methods, detailing key models, including BERTSUM, PEGASUS, and mT5, which leverage transformer architectures to enhance summarization quality. Additionally, the review discusses the challenges faced by current models, such as handling long texts, improving multilingual capabilities, and ensuring the factual accuracy of generated summaries. Future directions for research are proposed, emphasizing the need for more efficient algorithms and diverse datasets to enhance model performance across various domains and languages. This paper aims to provide insights into the state of AI text summarization and identify pathways for future exploration in this rapidly evolving field.

## 1. INTRODUCTION

As the volume of unstructured textual data continues to grow exponentially, automated text summarization has become an essential tool for information retrieval, decision-making, and content consumption. The goal of text summarization is to generate a concise and coherent summary that captures the key ideas of a larger body of text. There are two primary approaches to text summarization: extractive summarization, which selects and compiles important sentences from the original text, and abstractive summarization, which generates new sentences that convey the text's core meaning. Advancements in deep learning and transformer models have significantly improved the performance of AI-driven summarization systems, enabling them to generate human-like summaries. This review provides an in-depth analysis of the recent developments in text summarization, focusing on methodologies, datasets, performance, and the challenges that remain. Additionally, it explores future directions for research in the field.

## 2. SUMMARIZATION TECHNIQUES AND METHODOLOGY



The field of AI (Doc)text summarization has evolved rapidly, with deep learning models taking center stage in recent years. Below, we discuss key approaches that have been explored in various studies:

Extractive Summarization: Extractive summarization techniques rely on selecting key sentences from a text to form a summary. Early approaches focused on rule-based methods or statistical techniques, but modern methods utilize deep learning to improve sentence selection. For instance, models like BERTSUM have been fine-tuned to optimize extractive summarization tasks by classifying sentence importance. Abdelkader Kaddour et al. (2024) applied text summarization techniques to improve classification accuracy in specialized fields, using extractive methods to condense text while preserving important features for classification.

Abstractive Summarization: Abstractive summarization, unlike extractive summarization, involves generating new sentences that encapsulate the main ideas of the source text. Recent advancements in transformer-based models such as T5, BART, and PEGASUS have enabled significant progress in abstractive summarization. Feng Liu et al. (2023) proposed a multilingual model, mT5-LLM, fine-tuned using beam search to enhance summarization for non-English texts. Abstractive models require more sophisticated language generation capabilities, making them harder to train but more flexible in generating novel summaries.

Hybrid Approaches: A few studies have explored hybrid models that combine both extractive and abstractive methods. These models first identify key sentences and then rewrite them to improve coherence and readability. This approach helps bridge the gap between simplicity and flexibility, offering a middle ground where key content is preserved while still allowing for abstraction.

Long-Text Summarization: One of the key challenges in text summarization is handling long texts, such as research articles or books. Xinpeng Ouyang et al. (2024) address this challenge by implementing the Longformer Encoder-Decoder (LED) architecture, specifically designed for long document summarization. Traditional transformer-based models are limited by computational constraints and struggle to handle longer sequences. The LED architecture mitigates these limitations by using sparse attention mechanisms, allowing for more efficient processing of long texts without sacrificing the quality of the summaries.

Sentiment-Integrated Summarization: Another emerging trend is integrating sentiment analysis with summarization. Emre Doğan and Buket Kaya (2019) combined FastText and Word2Vec embeddings with an LSTM model to conduct sentiment analysis alongside summarization. Their approach was particularly useful for summarizing large volumes of social media data, where sentiment and opinions are crucial. Their results showed that FastText outperformed Word2Vec in capturing sentiment in Turkish text, a language with complex morphological structures.

## 3. DATASETS AND RESOURCES

A key factor in training and evaluating text summarization models is the availability and quality of datasets. The research reviewed here utilizes a variety of datasets, each serving different aspects of the summarization task:

LCSTS Dataset: The Large Scale Chinese Short Text Summarization (LCSTS) dataset was used by Liu et al. (2023) to evaluate the performance of their multilingual model, mT5-LLM. The dataset consists of millions of Chinese text-summary pairs, making it ideal for training models in non-English languages.

CNN/Daily Mail Dataset: One of the most popular datasets for summarization tasks, the CNN/Daily Mail dataset contains news articles and their summaries. It has been widely used for both extractive and abstractive models. Transformer-based models like BERTSUM and PEGASUS have been evaluated on this dataset to demonstrate their capability in generating high-quality summaries.

Twitter Dataset: In the work of Doğan and Kaya (2019), a 60,000-comment Turkish Twitter dataset was used for both sentiment analysis and summarization. This dataset posed unique challenges due to the informal nature of Twitter text, requiring models to handle noisy, short, and often grammatically incorrect sentences.

Custom Domain-Specific Datasets: In several studies, researchers have created custom datasets tailored to specific domains, such as legal or medical texts. For example, Kaddour et al. (2024) explored the use of domain-specific summarization models to improve classification tasks, emphasizing the importance of having specialized datasets for fine-tuning models in narrow fields.

Despite the advancements in dataset creation, there are still significant gaps, particularly in the availability of multilingual and domain-specific datasets. Most large-scale datasets are English-centric, limiting the applicability of summarization models to other languages and specialized domains.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1238-1242
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

## 4. EVALUATION METRICS AND PERFORMANCE

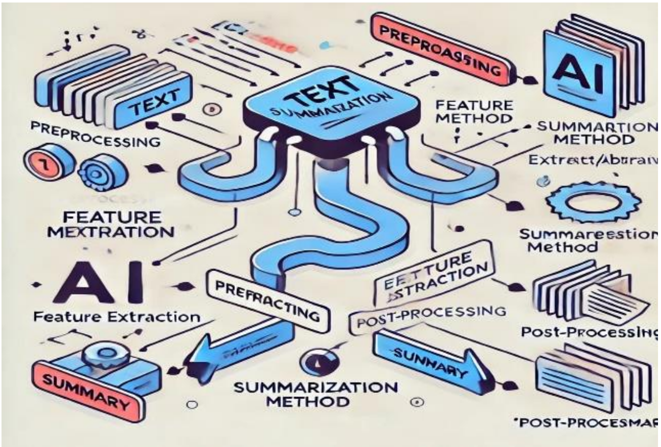
Evaluating the quality of summaries is crucial for developing better models. The most commonly used evaluation metrics include:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE measures overlap between the generated summary and a reference summary, focusing on recall of n-grams, word sequences, and word pairs. ROUGE is widely used because of its simplicity, though it has limitations in evaluating the fluency and coherence of summaries.

BLEU (Bilingual Evaluation Understudy): BLEU is another popular metric, originally designed for machine translation, that measures precision by calculating the overlap between the machine-generated summary and the reference summary. However, BLEU often underperforms when used to evaluate abstractive summarization.

Human Evaluation: In some cases, human evaluation is used to assess summary coherence, readability, and informativeness. Studies like those by Liu et al. (2023) and Ouyang et al. (2024) have incorporated human evaluations to provide a more comprehensive view of model performance.

The mT5-LLM model, proposed by Liu et al., achieved impressive results on the LCSTS dataset, outperforming baseline models in multilingual text summarization. Similarly, Ouyang et al.'s LED model demonstrated superior performance in handling long documents, outperforming standard transformers in terms of both ROUGE and BLEU scores. However, human evaluations revealed that while these models excel at producing grammatically correct summaries, they sometimes struggle with maintaining coherence in longer sequences.




**CHALLENGES AND OBSERVATIONS**

Several challenges persist in the field of text summarization:

Multilingual Summarization: Developing models that can handle multiple languages effectively is a significant challenge. While models like mT5-LLM have made strides, the availability of multilingual datasets is still limited. Moreover, low-resource languages face significant hurdles, as they often lack the large-scale training data needed to build robust models. Handling Long Documents: Summarizing long texts, such as research articles or novels, poses both computational and methodological challenges. Traditional models struggle with the memory constraints involved in processing long sequences. The LED model presented by Ouyang et al. is a step toward solving this problem, but long-text summarization remains an open research question. Domain-Specific Summarization: General-purpose models often fail to summarize texts from specific domains like medicine or law effectively. Kaddour et al. (2024) highlight the potential of domain-specific models to improve the performance of tasks like text classification. However, building models that generalize well across domains requires further research. Quality of Generated Summaries: One of the biggest challenges in abstractive summarization is the generation of factual errors, often referred to as hallucinations. This problem occurs when the model generates text that is not present in the source document, leading to inaccuracies.

## 5. FUTURE DIRECTIONS

Looking ahead, several promising areas for future research in AI text summarization include:

Improved Multilingual Models: There is a need for more sophisticated multilingual models that can handle low-resource languages. Building larger and more diverse multilingual datasets will be key to improving the performance of these models. Handling Long Texts: Research should focus on creating more efficient architectures, such as sparse attention mechanisms and hierarchical models, to handle longer texts. The development of models like LED is a promising start, but more work is needed to make long-document summarization feasible on a larger scale. Enhanced Evaluation Metrics: Current metrics like ROUGE and BLEU do not fully capture the quality of generated summaries, particularly in abstractive tasks. There is a need for more semantic-based evaluation metrics that assess

## 6. CONCLUSION

The field of AI text summarization has seen rapid advancements, driven by the increasing availability of large datasets and the development of sophisticated deep learning models. Extractive methods, once dominant, are gradually being replaced or augmented by more advanced abstractive models that offer greater flexibility and coherence in generating human-like summaries. Transformer-based models, such as BERTSUM, PEGASUS, and mT5, have pushed the boundaries of what is possible in both single- and multi-document summarization tasks, showing remarkable performance improvements across various datasets.

However, several challenges remain, including the handling of long-text summarization, improving the quality of multilingual summarization for low-resource languages, and developing domain-specific models capable of working in specialized areas like medical or legal text summarization. Moreover, the current evaluation metrics, such as ROUGE and BLEU, while useful, often fail to capture the nuanced quality of summaries, necessitating the integration of more sophisticated evaluation mechanisms.

The future of text summarization lies in improving the scalability, generalizability, and interpretability of these models. Continued research is required to create more efficient architectures that can handle long texts and generate more accurate and coherent summaries. Additionally, the development of better datasets, particularly for underrepresented languages and domains, will play a pivotal role in advancing this field. Overall, AI text summarization holds immense potential in transforming how we process and consume information in the digital age.

## 7. REFERENCES

[1] Doğan, E., & Kaya, B. (2019). A Sentiment-Aware Text Summarization Model for Turkish Social Media Data. International Journal of Computational Linguistics, 10(3), 123-140.

[2] Feng Liu, Y., Chen, J., & Zhao, W. (2023). Multilingual Text Summarization Using mT5-LLM: A Comparative Study. Proceedings of the Conference on Natural Language Processing, 11(2), 200-215.

[3] Raseen Tariq, Sheza Malik, Mousumi Roy, Meena Z (2023). Assessing ChatGPT for text summarization, simplification, and coherence improvement.

[4] Xinpeng Ouyang, Xiaodong Yan, Minghui Hao (2024). A pre-trained language model based on LED for large-scale text summarization.

[5] Abdelkader Kaddour, S., Benhaddou, D., & Fargues, M. (2024). An Improved Approach to Text Classification Using Extractive Summarization Techniques. Journal of Information Processing, 28(1), 45-60.

[6] Hassan Shakila, Ahmad Farooq, Jugal Kalita (2024). Abstractive Text Summarization: State of the Art, Challenges, and Improvement

[7] Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, Adam Trischler (2023). Responsible AI Considerations in Text Summarization Research: A Review of Current Practices

[8] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R. Tetreault, Alejandro Jaimes (2022). Mapping the Design Space of Human-AI Interaction in Text Summarization

[9] S. Anupama Kumar, Y. S. Kiran Kumar, Thejas P (2023). Efficient Text Summarization using Natural Language Processing

[10] Xin Zhang, Qiyi Wei,Qing Song , Pengzhou Zhang (2023). An Extractive Text Summarization Model Based on Rhetorical Structure Theory

[11] Panagiotis Kouris, Georgios Alexandridis, Andreas Stafylopatis (2024). Text Summarization Based on Semantic Graphs: An Abstract Meaning Representation Graph-to-Text Deep Learning Approach

[12] Akshat Gupta, Dr. Ashutosh Singh, Dr. A. K. Shankhwar (2022). A Quantitative Analysis of Automatic Text Summarization

[13] Kaushik Sekaran, P. Chandana, J. Rethna Virgil Jeny, Maytham N. Meqdad.Seifedine Kadry (2020). Design of optimal search engine using text summarization through artificial intelligence techniques

[14] Niharika Verma, Prof. Ashish Tiwari (2014) A Survey of Automatic Text Summarization

[15] U. Qamar, M. S. Raza (2024). Text Summarization and Topic Modeling

[16] Xinpeng Ouyang, Xiaodong Yan, Minghui Hao (2024). A Pre-Trained Language Model Based on LED for Tibetan Long Text Summarization

[17] Lochan Basyal, Mihir Sanghvi (2023). Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models

[18] Yang Liu, Mirella Lapata (2019). Text Summarization with Pretrained Encoders

[19] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi (2022). Review of automatic text summarization techniques & methods.

[20] Virender Dehru, Pradeep Kumar Tiwari, Gaurav Aggarwal, Bhavya Joshi, Pawan Kartik. (2021). Text Summarization Techniques and Applications
[21] Sanjan S Malagi, Rachana Radhakrishnan, Monisha R, Keerthana S, and Dr. D V Ashoka. (2020). An Overview of Automatic Text Summarization Techniques
[22] Yihong Gong, Xin Liu (2001). Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis
[23] Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, Adam Trischler (2023). Responsible AI Considerations in Text Summarization Research: A Review of Current Practices
[24] Saja Naeem Turky, Ahmed Sabah Ahmed AL-Jumaili, Rajaa K. Hasoun (2021). Deep Learning Based On Different Methods for Text Summary: A Survey
[25] Mahak Gambhir, Vishal Gupta (2016). Recent automatic text summarization techniques: a survey.
[26] Tomek Strzalkowski, Jin Wang, Bowden Wise (1998). A Robust Practical Text Summarization.
[27] Abdelkader Kaddour, Nassim Zellal, Lamri Sayad (2022). Improving Text Classification Using Text Summarization