IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156110

Section-Aware Scientific Paper Summarization System Using Transformers

Monthri Mrudhula[1], A. Gayathri[2]

MTech Student, Department of CSE, University College of Engineering, Osmania University, Hyderabad, India[1]

Assistant Professor, Department of CSE, University College of Engineering, Osmania University, Hyderabad, India [2]

Abstract: The need for automatic summarizing systems has grown as the quantity of scientific papers has expanded. This paper provides a Section-Aware Scientific Paper Summarization System that generates structured summaries by combining transformer models like BART, T5, and LED with BM25-based extractive techniques. In order to improve readability and information access, the system divides summaries into sections titled Introduction, Methodology, Results, and Conclusion. ROUGE measures are used to assess performance, and the findings demonstrate that the suggested method successfully maintains important information while generating succinct and logical summaries.

Keywords: Scientific Paper Summarization, BART, T5, LED, BM25, ROUGE, Natural Language Processing.

I. INTRODUCTION

Researchers and students find it challenging to swiftly extract crucial information from research articles due to the increasing volume of scientific publications. By producing succinct summaries of long documents, automated text summarization aids in overcoming this difficulty. The quality of automatic summarization has been enhanced by recent advances in Natural Language Processing (NLP), particularly transformer-based models. While abstractive approaches create new sentences that encapsulate the primary concepts, extractive methods choose important sentences from a document. Summaries that combine the two methods can be more successful. The Section-Aware Scientific Paper Summarization System presented in this paper combines transformer models like BART, T5, and Longformer Encoder–Decoder (LED) with BM25-based sentence ranking. The approach improves readability and information access by producing organized summaries for scientific articles' Introduction, Methodology, Results, and Conclusion sections. Using ROUGE measures, the quality of the summary is assessed. The primary contributions of this paper are:

- A hybrid extractive-abstractive summarization approach that makes use of BM25 and transformer models.
- Creating summaries for scientific articles by section.
- For summarizing, BART, T5, and LED models are integrated.
- Enhanced coverage of material via hybrid summarization.
- ROUGE measures for performance assessment.

II. RELATED WORK

The creation of transformer-based architectures and hybrid summarizing frameworks has propelled recent developments in Automatic Text summarizing (ATS). To enhance summary quality, coherence, and knowledge retention, previous research has investigated extractive, abstractive, and hybrid approaches. Combining extractive and abstractive strategies can result in more useful summaries than utilizing either strategy alone, as several studies have shown.

A hybrid approach was proposed by Sharma and Singh [1] in which the T5 transformer model was used to refine significant phrases after they were originally detected using the TextRank algorithm. Their results demonstrated the efficacy of hybrid summarizing techniques by demonstrating gains in recall, coherence, and overall summary quality. Similar findings were documented in survey research [6], which came to the conclusion that hybrid techniques successfully strike a compromise between linguistic fluency and information preservation.

Transformer architectures' capacity to capture long-range contextual dependencies has made them the predominant paradigm for abstractive text summarization. Transformer-based models produce more coherent and contextually appropriate summaries than traditional recurrent neural network architectures, as shown by Dhapola et al. [2]. Similarly, Prakash et al. [3] improved abstractive summarization by combining adversarial learning with BERT, which improved grammatical accuracy and semantic consistency. Jyoti et al. [4] revealed promising ROUGE performance on scientific summarizing datasets, further validating the efficacy of the T5 transformer model for producing relevant and fluent summaries.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁂ Peer-reviewed & Refereed journal ⁂ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156110

Specialized transformer architectures have been developed in response to the difficulty of processing lengthy research texts. The Longformer architecture, developed by Beltagy et al. [5], uses sparse attention techniques to effectively manage long sequences while maintaining contextual information. Longformer Encoder–Decoder (LED), an encoder–decoder version, has demonstrated considerable promise for long-document summarizing tasks, which makes it especially appropriate for scientific articles that frequently surpass the input constraints of traditional transformer models.

For superior abstractive summarization, recent research has also investigated sophisticated transformer variations like PEGASUS. While Dasari et al. [10] showed that PEGASUS can produce coherent summaries with little task-specific fine-tuning, Karteek et al. [7] used Masked Language Modeling (MLM) and Gap Sentence production (GSG) objectives to enhance summary production. The significance of transformer pre-training techniques in improving summarization performance is highlighted by these studies.

Transformer models have been effectively used in domain-specific applications beyond general text summarization. A transformer-based system for summarizing business reports with both textual and tabular data was created by Faizal et al. [8], illustrating the versatility of transformer topologies across document types. In terms of contextual comprehension, coherence, and summary quality, transformer-based approaches consistently beat statistical and conventional machine learning methods, according to comparative studies by Chakraborti et al. [9].

Three main research trends are indicated by the literature:

- Using hybrid extractive–abstractive frameworks to improve coverage of information [1], [6].
- The extensive application of transformer-based models for producing coherent summaries, including BART, T5, BERT, PEGASUS, and LED [2], [3], [4], [5], [7], and [10].
- The increasing demand for long-document summarizing methods that can process scientific articles [5], [8], [9].

Inspired by these results, the suggested Section-Aware Scientific Paper Summarization System blends transformer-based abstractive generation employing BART, T5, and LED models with BM25-based phrase extraction. Additionally, the suggested framework generates structured summaries divided into Introduction, Methodology, and Results & Conclusion sections, which improves readability and facilitates effective comprehension of scientific literature, in contrast to many current systems that produce a single generic summary.

III. METHODOLOGY

The suggested Section-Aware Scientific Paper Summarization System as shown in Fig. 1. creates structured summaries from scientific research publications using a hybrid extractive-abstractive approach. The approach improves the readability and coherence of summaries while preserving crucial content by combining transformer-based language models with information retrieval algorithms. Document capture, preprocessing, section identification, sentence extraction, abstractive summarization, hybrid enhancement, and evaluation make up the entire workflow.

A. Text extraction and document acquisition

Scientific documents can be entered into the system as raw text or as PDF files. A parser module transforms unstructured PDF documents into a machine-processable format by extracting text from each page. This allows the framework to handle research papers from a variety of scholarly sources while preserving input selection flexibility.

B. Preprocessing Text

Preprocessing is applied to the retrieved content in order to enhance data quality and eliminate unnecessary information. References, figure captions, URLs, email addresses, citation markers, and formatting artifacts are frequently included in scientific articles and do not aid in the creation of summaries. To get rid of these components, regular-expression matching-based noise reduction approaches are used. To make additional processing easier, the cleaned text is then tokenized and normalized.

C. Identification of Sections

The suggested architecture takes a section-aware approach, in contrast to traditional summarizing systems that handle a document as a single text block. Rule-based pattern matching and heading identification approaches are used to identify structural elements like Introduction, Methodology, Results, and Conclusion. This procedure allows the system to produce summaries that accurately replicate the original document structure while maintaining the logical arrangement of scientific documents.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156110

D. Selection of Extractive Content

The BM25 ranking algorithm is used by the system to determine which sentences are significant depending on how relevant they are to the content of the document [11]. Term frequency, inverse document frequency, and document length normalization are taken into account by BM25 when calculating a relevance score. Keyword-guided scoring is used to further enhance section-specific relevance by giving priority to informative phrases from various sections when creating summaries.

BM25 Equation is given as:

$$BM25(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\,(k_1+1)}{f(q_i,D) + k_1\left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}$$

where $f(q_i, D)$ represents the frequency of query term q_i in document D, $IDF(q_i)$ denotes inverse document frequency, $|D|$ is the document length, and $avgdl$ is the average document length.

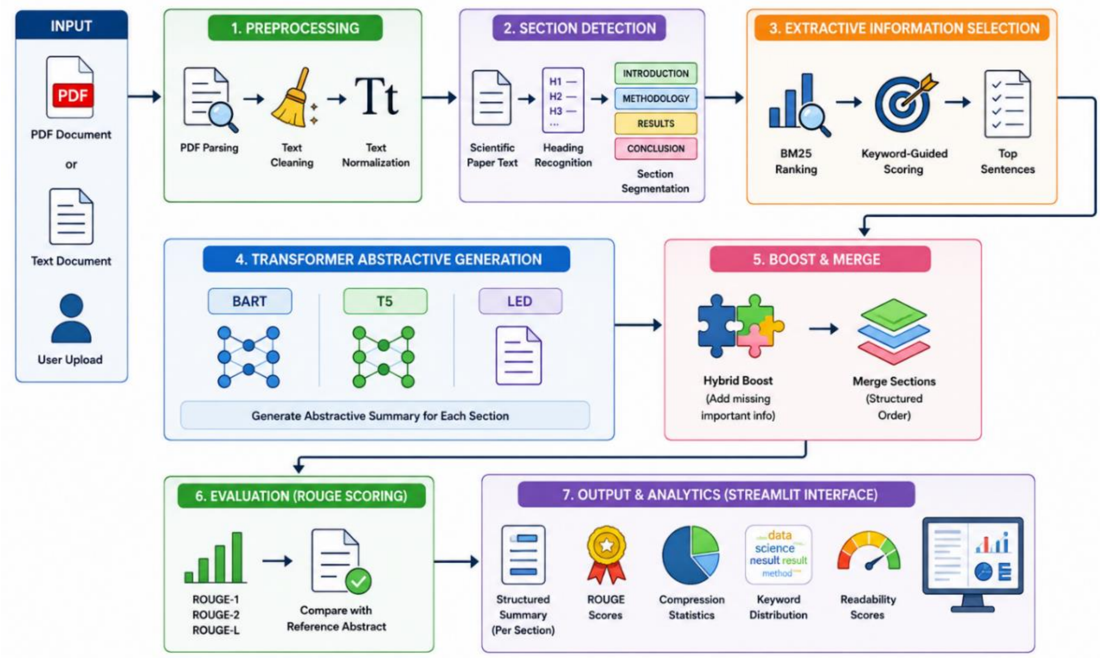

Fig. 1 Proposed System Architecture

E. Abstractive Summarization Using Transformers

Transformer-based models, specifically BART, T5, and Longformer Encoder–Decoder (LED), are used to summarize the chosen sentences. T5 approaches summary as a text-to-text generation job [13], LED effectively manages lengthy research materials using sparse attention mechanisms [5], and BART produces fluent summaries through sequence-to-sequence learning [12]. Together, these models make it possible to provide summaries that are both logical and contextually relevant.

F. Mechanism of Hybrid Enhancement

Transformer models produce fluid summaries, but sometimes during abstraction, crucial information is left out. A hybrid enhancement stage is implemented in order to overcome this constraint. The highest-ranked BM25 sentences are compared with the generated summary, and any important details that are absent from the generated output are added. While maintaining the advantages of abstractive summarization in terms of readability, this method enhances information coverage.

G. Creating Structured Summaries

The improved summaries are arranged into four predetermined sections: Introduction, Methodology, Results, and Conclusion. Structured information presentation enhances understanding and makes it easier for readers to find pertinent parts of a research report. The suggested approach differs from traditional summarizing algorithms that produce a single unstructured summary because of its section-aware representation.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156110

H. Assessment of Performance

ROUGE-based measures, such as ROUGE-1, ROUGE-2, and ROUGE-L, are used to assess the efficacy of the suggested framework. The similarity between generated summaries and reference summaries is gauged by these criteria. Better preservation of sentence structure and semantic information from the reference summary is indicated by higher ROUGE-L scores.

IV. RESULTS AND DISCUSSION

ROUGE-1, ROUGE-2, and ROUGE-L measures were used to assess the effectiveness of the suggested Section-Aware Scientific Paper Summarization System. These metrics measure unigram overlap, bigram overlap, and sequence-level similarity, respectively, to determine how similar the generated summaries are to the reference summaries. The efficacy of the hybrid extractive–abstractive summarization paradigm was assessed using scientific research publications.

TABLE I PERFORMANCE OF THE PROPOSED SYSTEM

Model	ROUGE-1 (F1)	ROUGE-2 (F1)	ROUGE-L (F1)
T5	0.5035	0.4399	0.3200
BART	0.5315	0.4638	0.2766
LED	0.5865	0.4835	0.3769

The findings show that all three transformer models as shown in Table 1 were successful in producing insightful summaries while maintaining the essential ideas of the original texts. Because LED uses sparse attention processes to analyze lengthy research papers well, it received the highest ROUGE scores among the models that were assessed. While T5 generated competitive results with balanced summary quality, BART also showed great performance by producing cohesive and fluid summaries.

V. CONCLUSION

A Section-Aware Scientific Paper Summarization System for producing succinct and organized summaries from scholarly research articles was presented in this work. The suggested system combines transformer-driven abstractive generation utilizing BART, T5, and LED models with extractive sentence selection based on BM25. The method generates summaries that maintain contextual relevance and the logical structure of scientific articles by utilizing section detection and keyword-guided content prioritization.

The experimental evaluation shows that the hybrid approach produces cohesive summaries with better readability and efficiently extracts important information from long research papers. By enabling readers to rapidly recognize important ideas, approaches, and conclusions, the use of section-wise summarizing further improves the usability of the produced output.

Researchers, students, and professionals who need to effectively review vast amounts of scientific literature may find the suggested framework to be a useful resource. Overall, the paper shows how integrating cutting-edge transformer designs with conventional information retrieval methods can help solve the problems related to long-document summarization in academic fields.

VI. FUTURE WORK

Future improvements to the suggested system might include support for multilingual summarization, which would allow research papers published in several languages to be processed. To enhance summary quality, the framework can also be expanded through domain-specific fine-tuning for fields like engineering, healthcare, and law. Furthermore, incorporating sophisticated complex language models and multi-document summarization features could improve the system's accuracy, contextual comprehension, and suitability for extensive literature review assignments.

REFERENCES

[1] N. Sharma and S. Singh, “A Hybrid Model for Text Summarization Using Extractive and Abstractive Approaches”, in *2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO),* 2025, doi: 10.1109/ICRITO66076.2025.11241496.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⪿ Peer-reviewed & Refereed journal ⪿ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156110

[2] S. Dhapola, S. Goel, D. Rawat, S. Vats, and V. Sharma, "Abstractive Text Summarization Using Transformer Architecture", in *2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC)*, pp. 13-17, 2024, doi: 10.1109/AIC61668.2024.10730840.

[3] A. Prakash, G. Rohith, and J. Umamageswaran, "Abstractive Text Summarization Using BERT and Adversarial Learning", in *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 2024, doi: 10.1109/ICTACS62700.2024.10840865.

[4] D. Jyoti, J. Srivastava, and D. P. Mahato, "Implementing T5 for Text Summarization: An Algorithmic Approach", in *Proceedings of the 2025 International Conference on Innovations in Computing (ICOIN)*, 2025, doi: 10.1109/ICOIN63865.2025.10992766.

[5] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer", *arXiv preprint arXiv:2004.05150*, 2020, doi: 10.48550/arXiv.2004.05150.

[6] S. Deo and P. K. Pattnaik, "A Short Survey of Abstractive, Extractive and Hybrid Text Summarization", in *2025 International Conference on Computational Intelligence and Green Computing (IC-CGU)*, 2025, doi: 10.1109/IC-CGU67042.2025.11337962.

[7] V. V. Karteek, P. V. S. Pranav, V. S. Charan *et al.*, "Transformer-Based Model for High-Quality Text Summarization", in *2025 International Conference on Information and Communication Technology (ICICT)*, 2025, doi: 10.1109/ICICT64420.2025.11004690.

[8] F. B., S. Abraham, and S. Thomas, "Automated Business Report Summarization Using Transformer Model", in *2024 International Conference on Advances in Computing, Communication and Systems (ICACCS)*, 2024, doi: 10.1109/ICACCS60874.2024.10716930.

[9] R. Chakraborti, R. Banerjee, and S. Das, "Evaluating the Efficacy of Text Summarization Models: A Comparison of NLP Algorithms", in *2025 International Conference on Emerging Technologies (IEMENTech)*, 2025, doi: 10.1109/IEMENTech65115.2025.10959463.

[10] N. V. K. S. Dasari, A. Sungheetha, and R. Sharma R., "Text Summarization Using PEGASUS Transformer Model in Machine Learning", in *2025 International Conference on Intelligent Computing and Informatics (ICICI)*, 2025, doi: 10.1109/ICICI65870.2025.11069669.

[11] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond", *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333--389, 2009.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871--7880, 2020.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1--67, 2020.

[14] A. Kumar, S. Reddy, and P. Sharma, "Enhancing Summarization of Legal Text Documents Using Pre-Trained Models", in *Proceedings of the International Conference on Emerging Smart Information Computing (ESIC)*, 2025. doi: 10.1109/ESIC64052.2025.10962732.

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pp. 74--81, 2004.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need", in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998--6008, 2017.

[17] A. Cohan *et al.*, "A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents", in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 615–621.

[18] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing", in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2020, pp. 38–45.

[19] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization", in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 1797–1807.

[20] S. Gupta and S. Gupta, "A Survey of Text Summarization Extractive Techniques", *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258–268, 2010.