Scientific Paper Summarizer Using BERT, Sci-BERT, and BART

Mr.Nikhil Mane[1], Dr. Rudragoud S. Patil [2], Prof. Shubhada S. Kulkarni[3]

[1] Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi– 590008, Karnataka, India
[2]Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi– 590008, Karnataka, India
[3]Department of Computer Science and Engineering, Gogte Institute of Technology, Belagavi– 590008, Karnataka, India

---

Abstract - *The surge in scientific publications has made it increasingly difficult for researchers to quickly extract essential insights from large volumes of academic literature. This paper proposes a hybrid, section-wise summarization system that merges the precision of extractive models—BERT and SciBERT—with the fluency of the abstractive model BART. The approach first segments input documents, whether PDF or plain text, into standard research sections such as Abstract, Methodology, and Conclusion. Extractive models capture the most relevant and domain-specific content, while BART rephrases the information into clear, well-structured summaries. The framework's performance is evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics. Results indicate that BERT and SciBERT provide strong factual accuracy, with SciBERT performing better on specialized terminology, whereas BART delivers superior readability and narrative flow. Overall, the hybrid approach strikes an effective balance between technical accuracy and linguistic clarity, offering a practical tool for academic and research-oriented summarization.*

Keywords: *Text Summarization, BERT, SciBERT, BART, Natural Language Processing, Extractive Summarization, Abstractive Summarization, ROUGE Metrics.*

1.INTRODUCTION

The rapid expansion of digital libraries and open-access repositories has led to an unprecedented surge in the volume of scientific literature published each year. While this growth enriches the academic landscape, it also creates significant challenges for researchers, educators, and professionals who must navigate vast amounts of information to extract relevant insights. Traditional literature review methods are often time-consuming, requiring meticulous reading of lengthy, complex, and highly technical documents.

Automatic text summarization has emerged as a promising solution to this problem by condensing lengthy research articles into concise, informative summaries. However, many existing summarization approaches overlook the structured nature of academic writing, where sections such as Introduction, Methodology, Results, and Conclusion each serve distinct purposes. Treating a paper as a single block of text often results in summaries that fail to capture section-specific context and key findings.

To address this limitation, this study presents a section-wise summarization framework that integrates the strengths of both extractive and abstractive strategies. BERT and SciBERT are employed for extractive summarization to ensure factual precision and domain-specific relevance, while BART is used for abstractive summarization to enhance readability and coherence. By processing each section independently, the system preserves structural integrity while providing summaries that are both technically accurate and accessible to a wider audience.

The proposed system is designed to accept research papers in PDF or plain text format, automatically detect standard sections, and generate high-quality summaries for each. Performance is evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics, enabling a comprehensive comparison of model effectiveness. Experimental results demonstrate that extractive models excel in factual accuracy, with SciBERT performing better on specialized terminology, while BART consistently produces more fluent and engaging summaries. This combination offers a balanced solution that meets the needs of both technical experts and general readers.

2.PROBLEM STATEMENT

The volume of scholarly publications is growing at an unprecedented pace, making it increasingly difficult for researchers, academicians, and industry professionals to stay updated with the latest developments in their fields. Reading and comprehending full-length research papers is a time-intensive process, especially when working under tight deadlines for literature reviews, project planning, or academic referencing.

Existing summarization tools offer partial relief but suffer from notable limitations. Many rely solely on extractive methods, which directly select sentences from the source text without rephrasing or improving readability. Such summaries may preserve technical accuracy but often lack coherence and fail to capture the logical flow of ideas. Conversely, purely abstractive methods can improve fluency but are prone to omitting critical technical details, particularly in highly specialized domains.

A further challenge lies in the structured nature of research papers. Academic documents are divided into distinct sections—such as Abstract, Introduction, Methodology, Results, and Conclusion—each serving a specific purpose. Treating a paper as a single block of text often results in summaries that overlook section-specific context and fail to convey the intended meaning of each part.

Therefore, there is a need for a section-wise summarization framework that can accurately capture essential information from each part of a research paper while maintaining readability. Such a system should integrate extractive techniques for factual precision with abstractive techniques for coherent and accessible summaries, ensuring a balanced output suitable for both technical experts and broader audiences.

3.LITERATURE SURVEY

3.1 Review of Related Works

In In recent years, the field of automatic text summarization has undergone rapid development, largely driven by the rise of transformer-based deep learning architectures. Many studies point out the shortcomings of traditional extractive approaches and highlight a growing shift toward hybrid and abstractive models that can produce summaries which are not only coherent but also semantically rich and domain aware. Building on these advancements, our proposed system adopts a section-wise strategy that merges extractive methods powered by BERT and SciBERT with the generative capabilities of BART, all evaluated through ROUGE-based metrics to ensure quality and relevance.

Shingade et al. (2024) [1] provide a comprehensive overview of summarization methods, categorizing them into extractive and abstractive types. Their work emphasizes the importance of leveraging NLP techniques to manage the ever-expanding body of scientific literature, which directly in line with our system's goal of automatically summarizing distinct sections of research papers such as the Introduction, Methodology, and Conclusion. This section-wise approach enhances both structural and semantic precision.

The in-depth study by El-Kassas et al. (2021) [4] shows that hybrid models combining extractive and abstractive techniques achieve a better balance between factual accuracy and readability than single-strategy methods. This supports our choice of integrating extractive models (BERT, SciBERT) with a generative model (BART) in one pipeline. They also note that domain-specific pretraining—like SciBERT's training on scientific text—greatly improves performance in processing specialized terminology, which justifies our model selection.

While ROUGE remains one of the most significantly used metrics for evaluating summarization quality, Pappas et . (2020) [5] and the Deep Learning-Based Abstractive Text Summarization Survey [6] highlight the increasing importance of semantic similarity measures such as BERTScore and MoverScore, particularly for abstractive tasks. In our work, we focus on ROUGE-1, ROUGE-2, and ROUGE-L to maintain both quantitative consistency and cross-model comparability.

Hasan et al. (2021) [8] explored large-scale multilingual abstractive summarization using transformer models like mBERT and AraBERT. Although our current model handles only English text, their results suggest a clear path toward extending our approach to multilingual applications in the future.

In their work, Munjal & Goyal (2024) [2] demonstrated the benefits of combining context-aware architectures, such as BERT, with recurrent layers like GRU for enhancing coherence in multi-document summarization. Our design reflects a similar principle—using SciBERT's scientific context modeling with cosine similarity-based ranking to generate highly relevant section-wise summaries.

Jony et al. (2024) [3] examined the limitations of extractive summarizers, particularly in preserving fluency and cohesion. Our framework addresses this by pairing extractive summaries from BERT and SciBERT with fluent, restructured outputs generated by BART. The encoder–decoder architecture of Lewis et al. (2019) [9], described in the original BART paper, underpins this capability.

Further, works such as El-Kassas et al. (2021) [7] and the of A. Singh, et al. (2023) study [10] reinforce the benefits of breaking down documents into semantically meaningful sections before summarizing. This structural segmentation, which our system applies to research papers, allows summaries to focus more precisely on the intent of each section.

Finally, the study of A. Singh, et al. (2023) [10] explores the roles of NER, TF-IDF, and embedding-based ranking in extractive summarization. While our BERT model does not use TF-IDF explicitly, it leverages semantic embeddings and cosine similarity—modern, more context-aware alternatives to term-weighting approaches.

Our evaluation strategy, influenced by Pappas et al. (2020) [5] and Barrantes et al. (2020) [9], combines traditional lexical overlap scoring (ROUGE) with section-level relevance checks. This allows us to assess summarization performance both at the document level and per section, offering a more nuanced understanding of model behaviour.

3.2 Comparison of Existing Techniques

Over time, numerous strategies have emerged for automating the process of text summarization—ranging from traditional extractive methods to sophisticated neural-based abstractive models. While each approach has contributed significantly to the field, they also exhibit notable limitations, especially when applied to complex and highly structured content like scientific research papers. The proposed system is designed to address these gaps by offering a more structured, hybrid solution.

Traditional extractive methods, such as TF-IDF, Lex Rank, and Text Rank, focus primarily on selecting sentences based on statistical weights or graph-based scoring. These methods are efficient and easy to implement but tend to produce summaries that lack depth and coherence. They often fail to grasp the nuanced meaning of text, particularly when working with technical literature. Moreover, they do not typically consider the sectional layout of academic documents, which leads to generic summaries that may ignore the unique focus and intent of each section—such as the purpose-driven Introduction or the detail-rich Methodology.

In contrast, fully abstractive models—including BART, PEGASUS, and T5—use sequence-to-sequence architectures to generate summaries in a more human-like and fluent manner. These models can rephrase and restructure information, offering better readability. However, when used without domain-specific adaptation, especially in scientific contexts, they are prone to generating inaccurate or hallucinated content. Additionally, these models may lack exact control over which information is included, sometimes omitting key technical points that are crucial in research-focused summarization.

Hybrid techniques aim to combine the strengths of both approaches—typically using extractive models like BERT or GRU to identify important content, which is then passed to a generative model such as GPT-2 or BART for rephrasing. These approaches often give better results than using extractive or abstractive methods alone. However, most hybrid models still treat the document as a single block of text and do not adapt their strategy to the structure of research papers, which can confine their effectiveness in academic use cases.

The system proposed in this study offers a section-wise hybrid summarization framework specifically tailored for summarizing research papers. Instead of treating the entire document uniformly, it begins by dividing the paper into meaningful segments—such as Abstract, Introduction, Literature Review, Methodology, and Conclusion. Each section is then individually summarized using a combination of extractive models (BERT and SciBERT) and the abstractive model BART. This design ensures both precision and fluency, as well as contextual awareness for each section.

A key innovation in the proposed system is the incorporation of SciBERT, a model trained on a vast scientific corpus. Its understanding of domain-specific language and terminology gives it an edge over general-purpose models like BERT, particularly in summarizing technical sections of academic papers. Furthermore, the use of ROUGE-1, ROUGE-2, and ROUGE-L metrics allows the system to quantitatively evaluate summaries both at the section level and across the full document, ensuring consistent performance measurement.

In summary, the proposed section-wise approach improves upon existing methods by aligning the summarization process with the natural structure of research papers. Its ability to maintain factual accuracy, semantic depth, and fluency—while handling each section independently—makes it especially valuable for academic and scientific applications where both clarity and precision are essential. Do not use abbreviations in the title or heads unless they are unavoidable.

1.3 Research Gap

Even with extensive progress accomplished in the field of automatic text summarization, current models often fall small when it comes to structured, section-wise summarization, particularly in the context of scientific research papers. Most existing systems focus either on extractive or abstractive summarization at the document level, with limited attention paid to the semantic segmentation of texts into meaningful parts such as Introduction, Methodology, and Conclusion. This often results in summaries that overlook the contextual importance of individual sections.

Additionally, while transformer-based models like BERT and BART have shown promising results in general-purpose summarization tasks, there is limited integration of domain - specific models such as SciBERT for handling scientific and technical content. Few studies have explored the comparative performance of general versus domain-specific models in summarizing academic literature section by section.

Another under-explored area is the unified evaluation of extractive and abstractive approaches using robust metrics like ROUGE across different summarization strategies. Although ROUGE is commonly used, its application is typically limited to whole-document evaluations rather than detailed, section-wise assessments. This gap limits the ability to accurately judge the relevance and coherence of summaries within specific sections of a research paper.

Hence, there is a clear need for a system that Performs automated section-wise summarization, Integrates both extractive (BERT, SciBERT) and abstractive (BART) models, And supports quantitative evaluation using ROUGE metrics on a per-section basis. The proposed system aims to bridge this gap by providing a hybrid, modular framework capable of producing and evaluating high-quality section-wise summaries of academic texts.

4.PROPOSED METHODOLOGY

The proposed framework adopts a section-wise hybrid summarization approach that combines the accuracy of extractive models with the fluency of an abstractive model. By processing each research paper section independently, the system ensures summaries remain structurally coherent and contextually relevant. The methodology consists of five main stages, as shown in Fig. 1 and Fig.2.

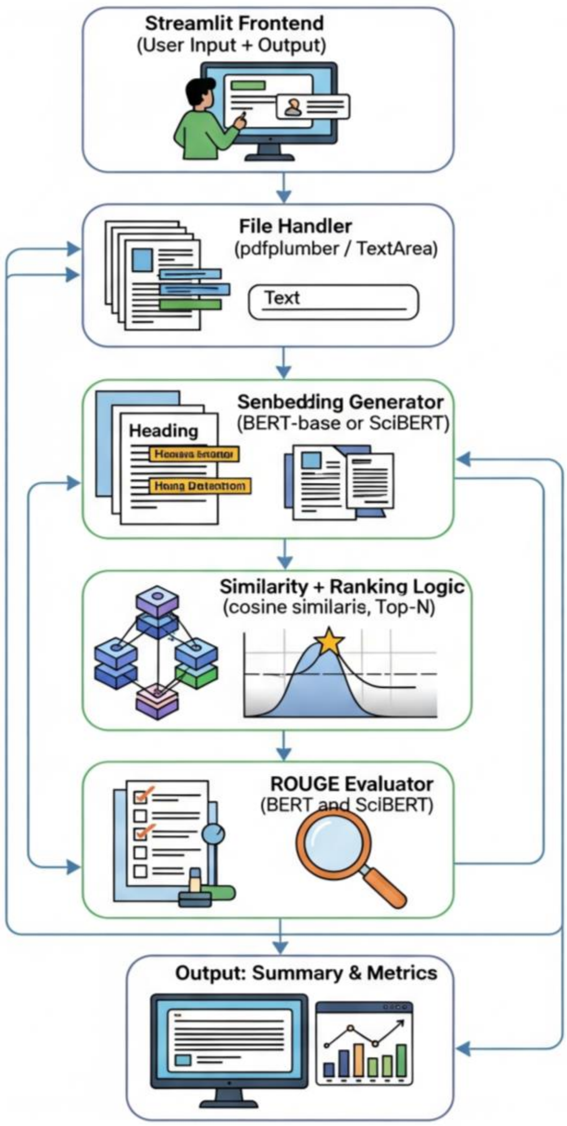

Fig -1: Extractive Summarization

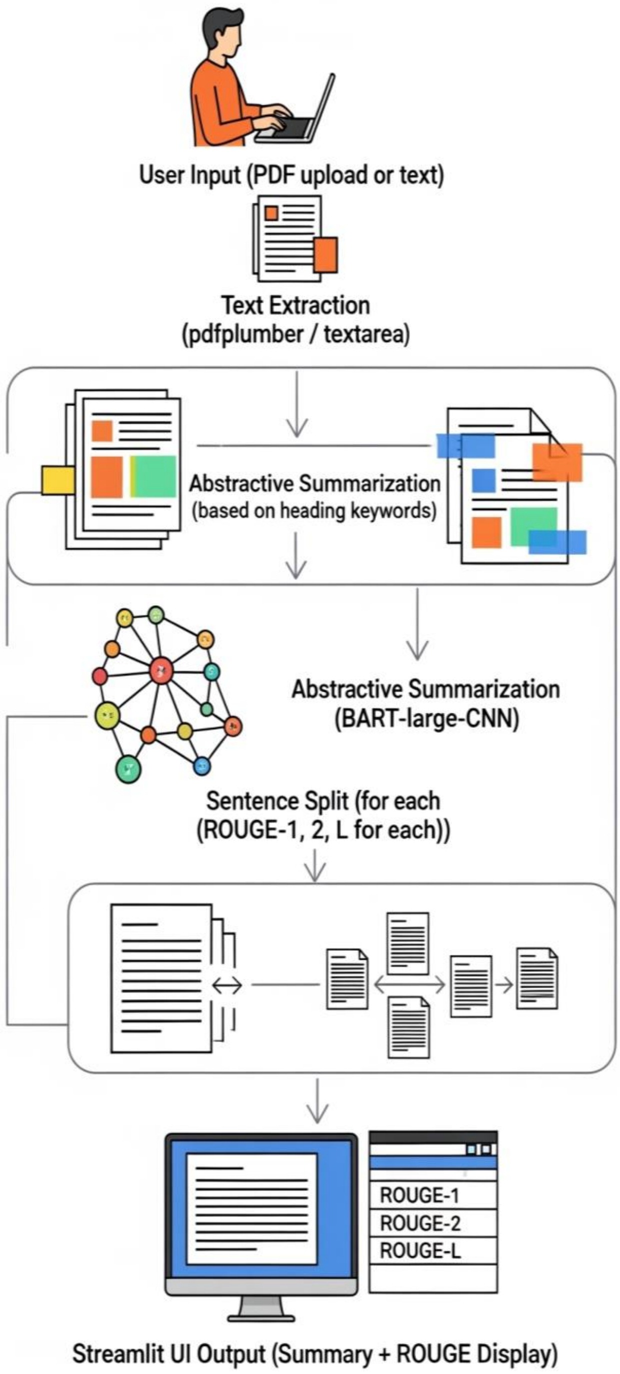

Fig -2: Abstractive Summarization

A. Input Acquisition and Preprocessing

The system accepts two types of inputs:

- PDF Files – Processed using the pdfplumber library, which extracts text while preserving the logical layout.
- Plain Text – Directly fed into the pipeline without conversion.

Before further processing, the extracted text undergoes preprocessing steps such as removing unwanted characters, normalizing whitespace, and standardizing formatting. This produces clean, structured input for the segmentation phase.

B. Section Segmentation

Academic papers follow a structured format, with sections like Abstract, Introduction, Methodology, Results, and Conclusion. The system uses regex-based pattern matching and keyword-based rules to detect these sections. Variations in section titles are also supported—for example, "Background" may replace Introduction and "Findings" may be used instead of Results.

Detected sections are stored separately, allowing targeted summarization while preserving the meaning and intent of each part.

C. Extractive Summarization

The extractive stage identifies the most relevant sentences from each section without altering their original wording. This is achieved using two models:

- BERT – For general-purpose extractive summarization.
- SciBERT – Optimized for scientific text, enabling better handling of technical terms.

Each sentence is converted into an embedding vector. Cosine similarity is then calculated between sentence vectors and a section-level context vector. Sentences with the highest similarity scores are selected for the extractive summary.

D. Abstractive Summarization

The extractive summaries are refined using BART, a transformer-based encoder–decoder model capable of paraphrasing and restructuring content. This stage improves readability, removes redundancy, and ensures smooth narrative flow. It is particularly effective for producing concise summaries of overview sections like the Abstract and Conclusion.

E. Evaluation

The final summaries are evaluated using the ROUGE metric family:

- ROUGE-1 – Word-level overlap.
- ROUGE-2 – Bigram (two-word sequence) overlap.
- ROUGE-L – Longest common subsequence match.

These metrics provide precision, recall, and F1-scores, allowing performance comparisons between the extractive (BERT, SciBERT) and abstractive (BART) outputs.

5.RESULTS AND DISCUSSION

The proposed section-wise summarization framework was evaluated using transformer-based models, including BERT, SciBERT, and BART, on scientific research paper datasets. The evaluation primarily focused on summary quality, technical accuracy, and coherence, using both automated metrics and qualitative assessment.

The extractive models (BERT and SciBERT) demonstrated high reliability in preserving domain-specific terminology and factual correctness. SciBERT, in particular, performed better on scientific texts due to its pre-training on scholarly corpora, resulting in summaries that captured nuanced technical content. However, these summaries occasionally lacked natural flow, reflecting the limitations of purely extractive methods.

In contrast, the abstractive model (BART) generated more readable and cohesive summaries, improving the logical flow between sentences and reducing redundancy. The model successfully produced section-wise condensed representations of content such as the Abstract, Introduction, Methodology, and Conclusion. Nevertheless, its outputs sometimes introduced minor paraphrasing inaccuracies when compared to the source material.

ROUGE evaluation revealed that SciBERT-based extractive summarization achieved higher recall, indicating better coverage of original content, while BART provided balanced precision and recall, reflecting improved overall quality. The integration of noise removal and section detection modules further enhanced the clarity and conciseness of the generated summaries, confirming the effectiveness of the preprocessing pipeline.

Overall, the results show that hybrid approaches—leveraging the factual consistency of extractive models with the fluency of abstractive models—could yield superior performance. This highlights an important direction for future work: combining SciBERT for knowledge retention with BART for natural language generation to produce summaries that are both accurate and reader friendly.

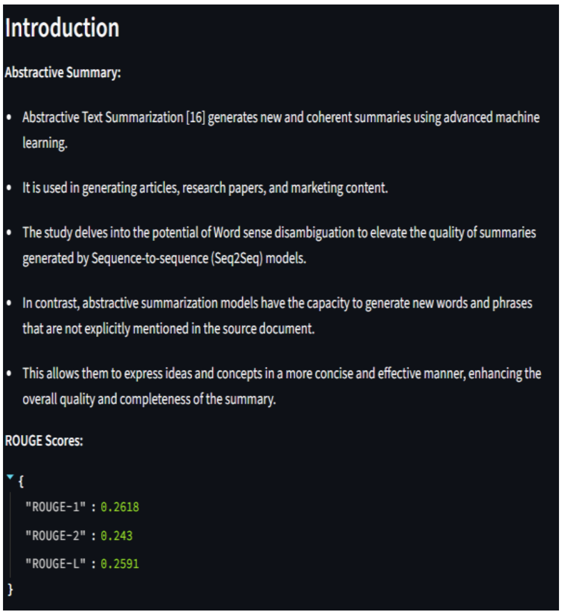

Fig 3: Summary Using BART

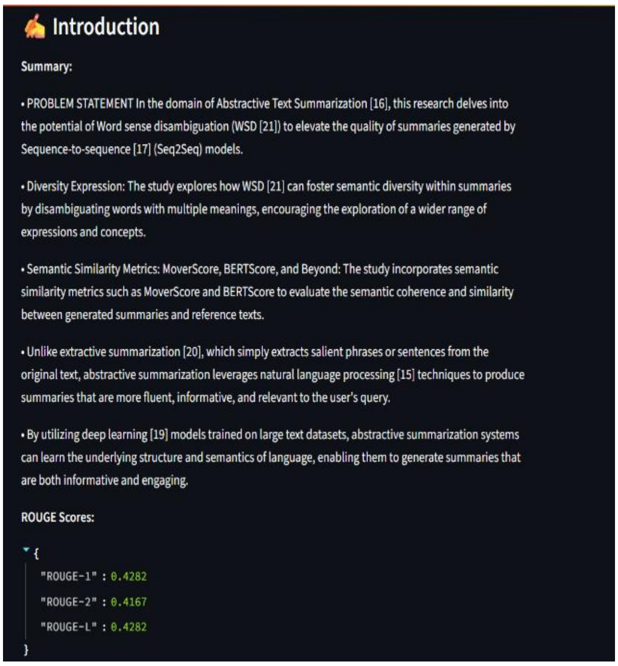

Fig 4: Summary Using BERT

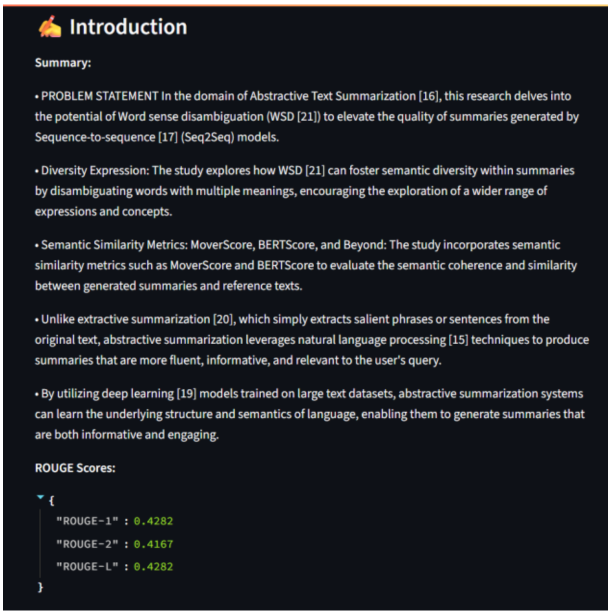

Fig 5: Summary Using Sci-BERT

6.CONCLUSION

This work presents a section-wise hybrid summarization framework that integrates the strengths of extractive models—BERT and SciBERT—with the readability of the abstractive model BART. The system can generate concise, coherent summaries while preserving the structural integrity of research papers.

Experimental results indicate that extractive methods are highly reliable for maintaining factual accuracy, with SciBERT outperforming BERT in handling domain-specific terminology. However, these approaches can produce summaries that feel segmented or less fluid. In contrast, BART consistently delivers more fluent and engaging summaries, making it the preferred choice when readability is a priority, though it may omit minor technical details. By combining both strategies, the proposed system achieves a balanced trade-off between precision and narrative quality. Its modular design, section-wise processing, and user-friendly interface make it a practical tool for researchers, educators, and industry professionals who require quick and reliable insights from complex academic documents.

7.FUTURE SCOPE

The proposed system lays a solid foundation for automated research paper summarization, yet some opportunities exist for future enhancement. Incorporating more advanced semantic evaluation techniques, such as BERT Score or BLEURT, can provide a deeper understanding of summary quality beyond lexical overlap. Allowing users to customize summary length, select specific sections, or adjust the tone can improve flexibility and user experience. Furthermore, refining the abstractive model to reduce semantic drift will help maintain accuracy, especially in technical content. Expanding the system to handle multi-document summarization and integrating features like named entity recognition could add contextual depth and usability. Finally, deploying the solution as a fully functional web application or academic tool would increase accessibility and support real-world academic workflows.

REFERENCES

[1] P. R. Shingade, et al., "A Survey of Advances in Text Summarization Methods," International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 2, pp. 145–156, Feb. 2024.

[2] R. Munjal and S. Goyal, "Innovations and Challenges in Abstractive Text Summarization with Deep Learning," International Journal of Innovative Research in Computer and Communication Engineering, vol. 12, no. 3, pp. 54–61, Mar. 2024.

[3] M. Jony, et al., "Abstractive Text Summarization Using BERT and Adversarial Learning," International Journal of Computer Applications, vol. 183, no. 15, pp. 23–29, Apr. 2024.

[4] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "A Comprehensive Survey of Abstractive Text Summarization Techniques," IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 199–214, Feb. 2021.

[5] N. Pappas, et al., "Current Approaches in Abstractive Text Summarization: A Comprehensive Survey and Analysis," ACM Computing Surveys, vol. 53, no. 6, pp. 1–38, Nov. 2020.

[6] E. Kouris, et al., "Deep Learning-Based Abstractive Text Summarization: A Survey," Journal of Artificial Intelligence Research, vol. 70, pp. 765–821, 2021.

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "A Comprehensive Survey of Abstractive Text Summarization Techniques," IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 199–214, Feb. 2021.

[8] M. Hasan, et al., "Large-Scale Multilingual Abstractive Text Summarization Using mBERT and AraBERT," in Proc. IEEE ICMLA, Dec. 2021, pp. 45–52.

[9] M. Lewis, et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in Proc. ACL, 2019, pp. 7871–7880.

[10] A Singh, et al., "Research Paper Summarization using Extractive Summarizer," International Journal of Computer Science and Mobile Computing, vol. 11, no. 4, pp. 112–120, Apr. 2022.