AI-Powered Research Paper Assistant: Automating Multi-Guideline Formatting and Summarization

by Mr. Harsh Tikone

AI-Powered Research Paper Assistant: Automating Multi-Guideline Formatting and Summarization

Mr. Harsh Tikone
Artificial Intelligence and Machine Learning Department
Thakur College of Engineering and Technology
Mumbai, India
1032210973@tcetmumbai.in

Mr. Bhavya Jain
Artificial Intelligence and Machine Learning Department
Thakur College of Engineering and Technology
Mumbai, India
1032210938@tcetmumbai.in

Mr. Satyam Rudrakanthwar
Artificial Intelligence and Machine Learning Department
Thakur College of Engineering and Technology
Mumbai, India
1032210963@tcetmumbai.in

Dr. Shiwani Gupta
Artificial Intelligence and Machine Learning Department
Thakur College of Engineering and Technology
Mumbai, India
shiwani.gupta@thakureducation.org

***Abstract*- The rapid increase in academic publications has made it challenging for researchers to efficiently summarize, evaluate, and format papers. These tasks often involve time-intensive manual processes, particularly when preparing submissions for varied journal and conference requirements. To address this, we developed the Research Paper Assistant, an AI-driven tool designed to automate key aspects of research paper analysis and preparation. This system integrates advanced Natural Language Processing (NLP) models, such as BERT and GPT, to perform extractive and abstractive summarization, providing precise and coherent summaries. Additionally, the tool's formatting module ensures compliance with diverse journal and conference guidelines. The Research Paper Assistant demonstrated seamless adherence to formatting, achieving 95% adherence to IEEE standards and strong performance during evaluations, achieving 85% accuracy in summarization. By automating these essential processes, the tool significantly reduces the effort required for paper preparation, enhancing productivity for researchers and reviewers. Its versatility makes it valuable across academia, peer review workflows, and publishing, with potential for expansion into additional standards and disciplines.**

Keywords- Research Assistant, Natural Language Processing, Paper Summarization, Paper Formatting, Machine Learning.

I. INTRODUCTION

The growing number of academic publications makes it hard for researchers to manage and analyze large amounts of information. Every year, millions of papers are published, creating challenges in staying updated, evaluating their quality, and preparing papers for submission. Tasks like summarizing findings, checking the structure and quality of papers, and ensuring proper formatting for different journals require a lot of time and effort.

Although tools like text summarizers and citation managers are available, they are often limited in scope and work independently. This creates difficulties for researchers who must use multiple tools to complete their tasks. Many existing solutions are not tailored to meet the specific needs of researchers or handle the unique requirements of academic workflows. This lack of an integrated solution leaves a significant gap. Advancements in Artificial Intelligence (AI), especially in Natural Language Processing (NLP), offer ways to simplify these tasks. However, a tool that combines summarization, evaluation, and formatting into one system has not been developed yet. There is also a need for such a tool to support different formatting styles required by journals and conferences, such as IEEE and others.

This study aims to create the Research Paper Assistant, an AI-powered tool that automates key research tasks. By using advanced NLP models for summarization, simple parameters for evaluation, and tools for multi-guideline formatting, this system provides a complete solution. The goal is to save researchers time and effort, allowing them to focus more on their work's innovation and analysis.

II. LITERATURE SURVEY

Early summarization techniques primarily relied on statistical methods, which identify and rank the most salient parts of a document based on frequency or position. Notable methods include heuristic approaches [1], [2], [3], which prioritize sentence importance based on simple scoring mechanisms such as term frequency or positional importance. Optimization-based models [4], [5] employ mathematical frameworks like Integer Linear Programming (ILP) or Determinantal Point Processes (DPP) to optimize the selection of sentences. While these methods ensure fluency and faithfulness, they often suffer from redundancy and lack the flexibility to paraphrase. Graph-based methods such as LexRank [6], TextRank [7], and PACSUM [8] represent documents as graphs, with nodes as sentences or words and edges as their relationships. While effective for extractive summarization, they are less suited for abstractive tasks.

With the advent of deep learning, summarization transitioned to data-driven approaches capable of capturing semantic nuances and generating more coherent summaries. Sequence-to-sequence (Seq2Seq) models employ RNNs and LSTMs [9], [10] to model summarization as a sequence transformation task. However, they often struggle with long documents like research papers. Reinforcement learning (RL) models like REFRESH [11] and BANDITSUM [12] optimize summaries by incorporating reward signals such as ROUGE scores for relevance and coherence. Graph neural networks

(GNNs) leverage document structure, making them effective for tasks involving citations and multi-document summarization [13]. Pointer-generator networks [14], [15] combine copying mechanisms with abstractive capabilities, enhancing factual accuracy. Recent advancements like DiffuSum [16] explore diffusion models for text generation, providing state-of-the-art performance in extractive summarization.

The development of pre-trained language models (PLMs) marked a paradigm shift in summarization, with models like BERT [17], BART [18], and PEGASUS [19] demonstrating exceptional performance across various datasets. These models are pre-trained on large corpora and fine-tuned for summarization tasks, allowing them to leverage contextual understanding and generate high-quality summaries. Hybrid models combine the strengths of extractive and abstractive approaches to achieve a balance between faithfulness and flexibility. GSUM [20] exemplifies this by using extractive signals to guide abstractive generation, resulting in concise yet comprehensive summaries. Among summarization models, PEGASUS stands out due to its Gap Sentence Generation (GSG) objective, which enhances its ability to generate summaries that effectively capture essential information. PEGASUS achieves state-of-the-art ROUGE scores on datasets like CNN/DailyMail and is highly flexible for diverse summarization tasks. It is also easy to fine-tune for specific domains, making it particularly well-suited for summarizing academic literature with technical terminology and structured information.

To demonstrate how PEGASUS outperforms other models like BART, PRIMERA, GSum, DiscoBERT, and SummaRunner, key gaps in their approaches must be considered. BART is a general-purpose model not specifically optimized for summarization, struggling with domain-specific tasks without extensive fine-tuning [21]. PRIMERA specializes in multi-document summarization but is less effective for single-document tasks and has high computational costs [22]. GSum heavily relies on guidance inputs and performs poorly in unguided summarization scenarios [23]. DiscoBERT focuses on discourse-level coherence but struggles with unstructured texts [24]. SummaRunner is limited to extractive summarization and lacks abstractive generation capabilities [25]. PEGASUS addresses these gaps by leveraging its specialized pre-training objective, domain adaptability, strong abstractive summarization capability, and ability to handle long-context inputs efficiently.

Pre-trained models are widely used in machine learning as they significantly reduce development time and computational costs by leveraging knowledge from large-scale datasets. Models such as PEGASUS, BERT, and GPT capture diverse patterns and representations, enabling better generalization and performance across tasks like summarization, translation, and image recognition. They support transfer learning, allowing fine-tuning for specific applications with minimal additional data and optimized deployment for scalability and efficiency. By providing access to state-of-the-art techniques through open-source platforms like Hugging Face, pre-trained models democratize AI, making advanced capabilities accessible to individuals and organizations with limited resources.

III. METHODOLOGY

The Research Paper Assistant is designed as a modular system with interconnected components that streamline the research paper writing and review process. The paper summarization module employs advanced Natural Language Processing (NLP) techniques using models like BERT for extractive summarization and GPT for abstractive summarization. The extractive approach selects key sentences based on contextual importance, while the abstractive model generates more natural, human-like summaries. These techniques are particularly effective in handling dense scientific literature, enabling users to quickly grasp the paper's contributions. The system offers flexibility by allowing users to choose between extractive and abstractive summarization, depending on their preference for detail.

Another integral component is the paper formatting module, which automates the process of adjusting a document to meet IEEE formatting standards. By leveraging Python libraries such as python-docx, the system ensures consistent adherence to formatting guidelines, modifying fonts, margins, headings, citations, and references accordingly. Although customizable for multiple formatting standards, IEEE is set as the default. This automation eliminates the need for manual formatting, reducing the time and effort required for paper submission preparation. The system enforces compliance with strict IEEE requirements, including specifications for font size, section headers, page margins, and citation styles. These modules, while independently functional, work cohesively to provide a comprehensive tool that spans the entire research paper process from summarization to formatting.

To develop efficient NLP and machine learning models, high-quality and diverse datasets were collected. The training and validation data for the Research Paper Assistant were sourced from publicly available academic repositories such as arXiv, IEEE Xplore, and Google Scholar. These sources provide a broad spectrum of research papers spanning multiple disciplines, including computer science, electrical engineering, biology, and social sciences, ensuring the system generalizes across various fields. The dataset used for formatting model development consisted of research papers that strictly adhered to IEEE standards. Python's python-docx library was employed to extract formatting rules from these papers, which featured complex elements like tables, equations, figures, and multi-level headers. The system was tested with non-IEEE formatted papers to ensure accurate transformation into IEEE-compliant documents.

For the summarization models, approximately 1,000 research papers were selected, covering diverse academic topics. The full-text papers and their corresponding abstracts served as the ground truth for evaluating summarization accuracy. Summaries generated by the models were compared with human-generated abstracts, and the models were fine-tuned accordingly. To improve robustness, the training dataset included different research content types, such as experimental results, theoretical explanations, and literature reviews, ensuring the system performed well across various paper structures.

The algorithmic design of the Research Paper Assistant integrates state-of-the-art techniques for formatting and summarization. The formatting module automates IEEE-compliant formatting using the python-docx library, which programmatically controls document structures. A predefined IEEE template is applied, ensuring adherence to specific guidelines, including Times New Roman font, 10-point font size, margin settings, and citation styles. The system processes unformatted papers and applies the IEEE template while automatically adjusting headings, figure captions, and citations. Additionally, the module includes validation checks to detect and correct formatting inconsistencies, such as incorrect font sizes or improperly formatted citations, minimizing human errors.

For summarization, the system employs both extractive and abstractive NLP models. The extractive approach utilizes BERT, which ranks key sentences based on their importance within a document. BERT's bidirectional understanding of word relationships allows it to identify the most informative sentences, which are then extracted as a summary. The model is fine-tuned using a scientific papers dataset to better recognize domain-specific language. In contrast, abstractive summarization relies on GPT to generate entirely new sentences that capture the core findings of the paper in a coherent and human-readable manner. Unlike extractive summaries that retain the original text, GPT's abstractive model constructs an entirely new summary that conveys the essence of the paper's contributions. Both models are trained and validated using datasets that pair full-text papers with their abstracts, with performance measured using ROUGE metrics to ensure alignment with human-generated summaries.

IV. RESULT AND DISCUSSION

The Paper Formatting Module was tested on 20 research papers requiring adherence to IEEE formatting guidelines, and the results demonstrated high efficiency and accuracy. The module achieved 100% compliance with IEEE formatting standards, including font styles, margins, headings, and citation styles, as verified through automated checks and human reviewers. It effectively handled complex document structures such as multi-level headings, equations, and intricate citation references without errors. Users reported a significant reduction in formatting time, completing tasks in minutes that previously took hours. The integration of the bibtexparser for citation management ensured all references were correctly formatted and synchronized with in-text citations, which was particularly useful for papers with extensive reference lists.

When compared to existing formatting systems such as LaTeX and Overleaf, the proposed system's automation using Python-docx provided faster and more consistent results. While LaTeX offers greater flexibility, it requires manual intervention and is time-consuming for users without technical expertise. In contrast, Overleaf, though more user-friendly, had a lower compliance rate with IEEE standards. The proposed system balanced automation and accessibility, eliminating the need for technical knowledge while maintaining high formatting standards. As demonstrated in the evaluation, the proposed system achieved 100% compliance within 10 minutes, whereas LaTeX required approximately 120 minutes despite offering high flexibility, and Overleaf required 90 minutes but had a lower compliance rate of 65%.

The Paper Summary Module was evaluated using a dataset of 50 research papers spanning multiple academic disciplines. Human reviewers assessed system-generated summaries for readability, conciseness, and content retention. The extractive summarization approach, powered by BERT, retained approximately 85% of critical information by directly extracting key sentences from the text. While effective in capturing important technical details, the summaries occasionally lacked fluidity due to the rigid nature of extractive summarization. This issue was particularly noticeable in highly technical papers, where specialized jargon or complex sentences were extracted without smooth transitions, leading to disjointed summaries.

In contrast, the abstractive summarization approach, driven by GPT, produced more coherent and readable summaries that closely resembled human-written abstracts. However, it sometimes omitted important technical details, particularly in sections related to methodology and experimental results. The focus on readability and brevity occasionally came at the cost of depth in complex academic papers. Fine-tuning the model could improve its ability to balance readability with comprehensive content retention.

When compared to existing summarization models such as TF-IDF-based and rule-based approaches, the use of BERT and GPT in the proposed system resulted in more accurate and coherent summaries. Traditional TF-IDF models struggled with context and coherence, particularly in longer texts, leading to lower content retention and readability. In the evaluation, BERT-based extractive summarization achieved an accuracy of 85%, with high content retention but medium readability and longer summary length. GPT-based abstractive summarization had a readability rating of high but retained only 75% of content, producing shorter summaries. In comparison, TF-IDF-based summarization demonstrated lower accuracy (65%), readability, and content retention, making it a less effective choice for academic research papers.

V. CONCLUSION

This paper introduced an AI-powered Research Paper Assistant designed to automate the summarization, scoring, and formatting of academic papers, significantly reducing the time required for these essential tasks and allowing researchers to focus more on innovation and critical thinking. By leveraging advanced techniques in Natural Language Processing and machine learning, the system provides flexible summarization options through extractive and abstractive methods, while the scoring module offers reliable metrics for evaluating the quality and impact of research. Future improvements will aim to enhance the scoring system to assess novelty and originality more effectively, as well as incorporate LaTeX formatting capabilities to better serve researchers who prefer this widely-used typesetting system. By continuously evolving and integrating user feedback, the Research Paper Assistant aspires to become an indispensable tool in the academic landscape, streamlining the research process and fostering greater innovation across various fields.

REFERENCES

[1] A. Nenkova and K. McKeown, "Automatic summarization," *Foundations and Trends in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2011.

[2] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 404–411.

[3] G. Erkan and D. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.

[4] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. SIGIR*, 1998, pp. 335–336.

[5] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. Eur. Conf. Information Retrieval (ECIR)*, 2007, pp. 557–564.

[6] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proc. Association for Computational Linguistics (ACL)*, 2011, pp. 510–520.

[7] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *Proc. Human Language Technology Conf. (HLT-NAACL)*, 2010, pp. 912–920.

[8] A. See, P. Liu, and C. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 1073–1083.

[9] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016, pp. 484–494.

[10] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proc. International Conf. Learning Representations (ICLR)*, 2018.

[11] Z. Zhang, Y. Guo, and S. Zhang, "DiscoBERT: Distilling discourse information for extractive summarization," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 1415–1426.

[12] K. Liu and H. Chen, "Multi-document summarization based on evolutionary algorithms," in *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 3, pp. 356–368, 2009.

[13] M. Zopf, "Estimating summary quality with pairwise rankers," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 2059–2069.

[14] S. Narayan, S. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 1747–1759.

[15] W. Zhu, Y. Wang, and L. Zhang, "Enhancing extractive text summarization with entailment-based pretraining," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 3744–3755.

[16] R. Fabbri, I. Li, and D. Shapira, "SummEval: Re-evaluating summarization evaluation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 145–155.

[17] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3730–3740.

[18] A. Glavas and G. Šnajder, "Event-centric extractive summarization for multi-document news summarization," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5063–5073.

[19] R. Zhang, J. Zhou, and D. Wang, "BERTSUMEXT: Extractive summarization using BERT," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3295–3305.

[20] M. Lewis, Y. Liu, and N. Goyal, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 7871–7880.

[21] J. Zhang, Y. Wei, and H. Wang, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. International Conf. Machine Learning (ICML)*, 2020, pp. 11328–11339.

[22] S. Li, R. Li, and C. Sun, "PRIMERA: A data-efficient transformer-based model for multi-document summarization," in *Proc. Conf. Empirical Methods in*

Natural Language Processing (EMNLP), 2021, pp. 3405–3416.
[23] R. P. Schuster, M. Glavas, and I. Vulic, "Multi-XScience: A large-scale dataset for multi-document summarization of scientific articles," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2022, pp. 5180–5191.
[24] K. Ganesan, "ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 2450–2460.
[25] A. Das and S. Ghosh, "A survey on automatic text summarization: Progress, trends, and challenges," in *IEEE Access*, vol. 8, pp. 143589–143619, 2020.

AI-Powered Research Paper Assistant: Automating Multi-Guideline Formatting and Summarization

ORIGINALITY REPORT

20%	18%	18%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	2%
2	Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, Tianming Liu. "ChestXRayBERT: A Pretrained Language Model for Chest Radiology Report Summarization", IEEE Transactions on Multimedia, 2021 Publication	2%
3	Wei Li, Hai Zhuge. "Abstractive Multi-Document Summarization Based on Semantic Link Network", IEEE Transactions on Knowledge and Data Engineering, 2021 Publication	1%
4	www.iis.sinica.edu.tw Internet Source	1%
5	www.techscience.com Internet Source	1%
6	Shah, Dhvani K., Vinayak A. Bharadi, V.J. Kaul, and Sameer Amrutia. "End-to-End Encryption Based Biometric SaaS: Using Raspberry Pi as a Remote Authentication Node", 2015 International Conference on Computing Communication Control and Automation, 2015. Publication	1%
7	ijitr.com Internet Source	1%
8	deepai.org Internet Source	1%

9	Junhua Ding, Haihua Chen, Sai Kolapudi, Lavanya Pobbathi, Huyen Nguyen. "Quality Evaluation of Summarization Models for Patent Documents", 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS), 2023 Publication	1 %
10	Tsutomu Hirao, Masaaki Nishino, Yasuhisa Yoshida, Jun Suzuki, Norihito Yasuda, Masaaki Nagata. "Summarizing a Document by Trimming the Discourse Tree", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015 Publication	1 %
11	Yusheng Su, Xu Han, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Peng Li, Jie Zhou, Maosong Sun. "CSS-LM: A Contrastive Framework for Semi-Supervised Fine-Tuning of Pre-Trained Language Models", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021 Publication	1 %
12	Chaobo Zhang, Junli Wang, Hongda Qi, Chungang Yan, Changjun Jiang. "Multi-View Metrics Enhanced Heterogeneous Graph Neural Network for Extractive Summarization", 2021 China Automation Congress (CAC), 2021 Publication	1 %
13	Achmad F. Abka, Kurniawati Azizah, Wisnu Jatmiko. "Transformer-based Cross-Lingual Summarization using Multilingual Word Embeddings for English - Bahasa Indonesia", International Journal of Advanced Computer Science and Applications, 2022 Publication	<1 %

14	Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, Zhiyong Wang. "TLDW: Extreme Multimodal Summarization of News Videos", IEEE Transactions on Circuits and Systems for Video Technology, 2024 Publication	<1 %
15	Xinyuan Song, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Ming Li, Tianyang Wang, Xuanhe Pan, Jiawei Xu. "Transformer: A Survey and Application", Open Science Framework, 2024 Publication	<1 %
16	Fenia Christopoulou, Eleftheria Briakou, Elias Iosif, Alexandros Potamianos. "Mixture of Topic-Based Distributional Semantic and Affective Models", 2018 IEEE 12th International Conference on Semantic Computing (ICSC), 2018 Publication	<1 %
17	ijritcc.org Internet Source	<1 %
18	Luobing Dong, Meghana N. Satpute, Weili Wu, Ding-Zhu Du. "Two-Phase Multidocument Summarization Through Content-Attention-Based Subtopic Detection", IEEE Transactions on Computational Social Systems, 2021 Publication	<1 %
19	iasscore.in Internet Source	<1 %
20	ijeecs.iaescore.com Internet Source	<1 %
21	zenodo.org Internet Source	<1 %
22	David Grimsman, Mohd. Shabbir Ali, Joao P. Hespanha, Jason R. Marden. "The Impact of Information in Distributed Submodular	<1 %

Maximization", IEEE Transactions on Control of Network Systems, 2019
Publication

---

23 Maida Shahid, Summra Saleem, Aniqa Dilawari, Usman Ghani Khan. "Telegenic Data Analysis for Compact Description and Genre Classification based on Re-ranking", 2019 International Conference on Communication Technologies (ComTech), 2019 — <1%
Publication

---

24 ar5iv.org — <1%
Internet Source

---

25 dergipark.org.tr — <1%
Internet Source

---

26 ijarcs.info — <1%
Internet Source

---

27 kmap.arizona.edu — <1%
Internet Source

---

28 www.lexisnexis.com — <1%
Internet Source

---

29 dokumen.pub — <1%
Internet Source

---

30 eprints.uwe.ac.uk — <1%
Internet Source

---

31 ir.ia.ac.cn — <1%
Internet Source

---

32 Jian Gu, Pasquale Salza, Harald C. Gall. "Assemble Foundation Models for Automatic Code Summarization", 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), 2022 — <1%
Publication

---

33 Sotudeh Gharebagh, Sajad. "Effective Domain-Specific Text Summarization", Georgetown University, 2024 — <1%

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		