

Semantics-Preserving Abstractive Summarization for Long-Form Text

Surya Prasad S H¹ and Kallinatha H D²

^{1,2}Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumakuru, Karnataka, India
Email: hdk@sit.ac.in

Abstract— The explosion of digital content has intensified the demand for automated summarization systems capable of transforming lengthy text into concise and meaningful representations. This work introduces Light Reads, an abstractive summarization framework developed by fine-tuning a T5 transformer model. Trained on a corpus of BBC news articles, the system applies modern Natural Language Processing (NLP) techniques to generate summaries that retain essential information while remaining succinct. To support practical use, Light Reads is equipped with an intuitive Gradio-based interface, enabling users to input text and obtain summaries instantly—making it suitable for applications across journalism, academia, and general information processing. The fine-tuned model shows strong performance, with notable improvements in ROUGE scores compared to baseline systems. This paper outlines the overall system design, fine-tuning methodology, and experimental evaluation, demonstrating that Light Reads produces coherent, information-rich summaries. We conclude by highlighting potential areas for expansion, including multilingual capabilities and enhanced handling of extended or complex documents.

Index Terms— Text summary, NLP, ML model, LLM, and Transformer.

I. INTRODUCTION

Navigating the modern information landscape presents a significant challenge, as the constant stream of news and articles makes it difficult for people to keep abreast of current events [1]. Text summarization addresses this problem by condensing lengthy documents into their essential elements in a digestible format. While extractive approaches identify and compile key sentences from the source material, abstractive summarization creates original phrasing that captures the underlying message [2]. The emergence of transformer-based language models has dramatically transformed abstractive summarization, producing far more natural and contextually aware summaries compared to traditional statistical or RNN-driven techniques [3].

Despite considerable strides in the field, many current systems still fall short of delivering summaries that balance brevity with informational richness [4]. Our work capitalizes on a cutting-edge transformer architecture (Google's T5) that we have customized through training on a curated collection of news articles to produce summaries that faithfully represent the original content's core messages while substantially condensing length [5]. We have incorporated an approachable interface designed for non-technical audiences, enabling anyone to access the tool without specialized knowledge.

The complete system is intended for use cases where rapid comprehension of substantial textual material is necessary—such as busy journalists reviewing multiple publications, scholars examining research literature, or students preparing for exams using extensive course materials [6][7].

This work introduces Light Reads, a refined T5-based solution for generating summaries. We explain how we configured the T5 framework for this task by fine-tuning on the BBC News Summary dataset and integrating mechanisms for controlling output conciseness and promoting variety in generated text to enhance result quality. We also evaluate our approach against comparable existing methods to demonstrate its effectiveness. The paper proceeds as follows: Section II surveys the existing literature on abstractive summarization and transformer-based architectures. Section III outlines the system design and training procedure, covering the model development process and relevant mathematical concepts. Section IV presents the experimental methodology, performance findings, and practical demonstrations through the interface. Section V summarizes key findings and discusses potential enhancements for future iterations.

II. LITERATURE REVIEW

The evolution of text summarization began with rule-based extraction techniques that leveraged metrics like sentence placement and term frequency to identify important passages [2][8]. As neural networks emerged, researchers adopted sequence-to-sequence frameworks such as RNNs and CNNs for generating abstractive summaries, though these architectures encountered difficulties preserving meaning across longer passages and maintaining logical flow [3][4][9]. The introduction of the Transformer architecture combined with large-scale pretrained models fundamentally reshaped the field, establishing what is now considered the leading methodology [10].

A significant breakthrough came with Liu and Lapata's BERTSum, which adapted BERT for both extractive and abstractive tasks, demonstrating impressive performance on major evaluation datasets including CNN/DailyMail. PEGASUS emerged as another influential model, built explicitly for abstractive summarization work and proving particularly effective for news-related tasks [5][6][11]. Research by Shanthakumari et.al. [3] demonstrated that PEGASUS generates summaries with superior coherence and contextual depth compared to earlier generation systems. Beyond model architecture, scholars have investigated enrichment strategies—Asif and Yaseen [4] incorporated semantic similarity assessment with BART to enhance summary precision, while Yadav et.al[5] benchmarked different computational approaches for specialized domains like medical literature. Contemporary innovations include Topic-GraphSum [6], which tackles extended document summarization, and BARTScore [7], which moves beyond conventional assessment techniques to provide more sophisticated evaluation. Important cautionary research by Veselovsky et.al[8] raises questions regarding the reliability of large language models for summary creation and potential threats to dataset authenticity. Domain-specific models like DialogLM [9] for conversational content and T5[11][12] for journalistic pieces underscore the continuing dominance of transformer-based solutions, consistently delivering superior results across diverse applications.

This trajectory reveals a fundamental shift toward transformer-based methods that prioritize customization of established models for particular subject areas, sophisticated generation techniques, and thorough quality measurement. Our contribution reflects this progression by adapting T5 for journalistic content, incorporating user-friendly interactive capabilities, and demonstrating how principled fine-tuning produces practical, accessible summarization systems suited to real-world information management challenges.

III. PROPOSED SYSTEM

The Light Reads framework employs a staged processing architecture that converts source material (such as news content) into shortened summaries through a carefully designed workflow. As demonstrated in Fig. 1, the system comprises three interconnected layers: an initial data processing stage, the core T5 transformer engine, and a presentation interface for user engagement.

The complete process starts when a user provides raw textual content, which undergoes normalization and conversion into machine-readable tokens, feeds into the transformer for summary synthesis, and then delivers results through the visual interface.

The system incorporates quality assessment mechanisms (such as ROUGE evaluation) and adjustable settings for controlling output parameters including conciseness and lexical variety to accommodate individual preferences. This component-based architecture enables autonomous development and refinement of each module while preserving functional cohesion across the entire system.

Light Reads Summarizer

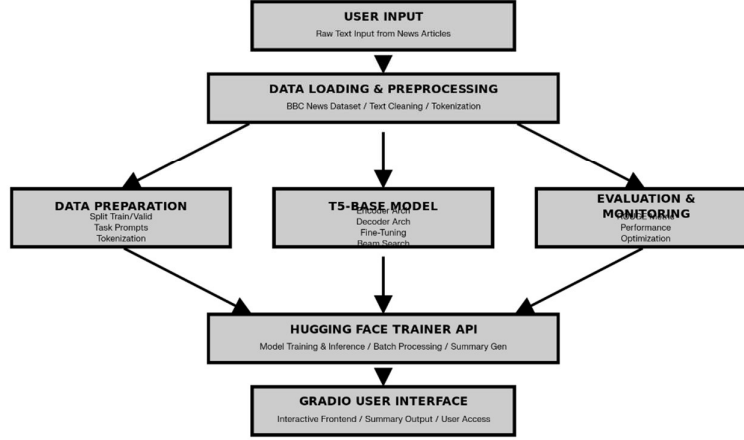


Figure 1. System architecture of the proposed Light Reads text summarization tool.

The Quick-Reads framework operates through an integrated pipeline that efficiently produces condensed text representations. The process initiates by importing the BBC News Summary dataset via Hugging Face's data library, subsequently dividing the material into separate learning and testing portions. Incoming data undergoes preparation and transformation via the T5 tokenization protocol, with contextual instructions incorporated for clarity. The foundational T5 model is retrieved and customized through intensive training against the learning dataset employing Hugging Face's training framework. Critical configuration choices including data batch sizing, gradient adjustment rates, and iteration counts are methodically selected to circumvent model memorization. Throughout the learning phase, effectiveness is tracked using ROUGE assessment methodology, which quantifies commonalities between machine-generated and human reference summaries. Upon conclusion of the learning phase, the refined model is stored for subsequent application. During practical use, the system synthesizes summaries for novel material through algorithmic beam exploration, guaranteeing both variety and caliber in produced content. A Gradio-powered interactive platform permits straightforward article input and summary retrieval. The architecture's adaptable framework facilitates straightforward maintenance, enhancement procedures, and alternative model substitution. This represents a comprehensive text summarization offering balancing computational efficiency, result dependability, and intuitive usability.

To enable straightforward engagement with the summarization capability, we constructed an accessible interface using Gradio technology. The user experience features an uncomplicated layout: an upper text field accepts article material for processing, a "Summarize" button activates the backend system, and a lower display zone presents the synthesized result. This presentation eliminates technical barriers, permitting users to obtain summaries without programming knowledge or model architecture comprehension.

Our experimental validation demonstrates that the customized T5 system performs reliably for journalistic content condensation, surpassing its base configuration and generating summaries strongly corresponding to professional human summaries. The combination of quantitative validation metrics and practical output examples substantiates both the technical robustness and operational utility of the Light Reads framework.

IV. EXPERIMENTAL DETAILS AND RESULTS

Our work utilized the BBC News Summary collection, containing hundreds of journalistic pieces paired with corresponding summaries [1]. These aligned pairs served as supervised examples guiding the model's learning process for generating condensed versions. We selected the T5-base variant, featuring approximately 220 million learnable parameters, as our foundation architecture due to its proven track record on generative language tasks and computational feasibility given our available infrastructure. The customization process occurred within a GPU-accelerated setting (NVIDIA Tesla T4) [11][12]. The model underwent four rounds of iterative learning, a duration determined sufficient based on convergence indicators in our validation monitoring. We implemented a learning velocity of $3e-5$ paired with groupings of 8 sequences, employing gradient accumulation methodology

to effectively manage extended token sequences. The Hugging Face training framework automatically documented performance indicators and preserved intermediate model versions. Throughout the customization phase, we monitored loss metrics across both the learning and evaluation portions of our dataset to confirm effective learning without model memorization. Figure 2 illustrates how loss evolved across the four training cycles. The learning loss exhibited consistent decline, dropping from approximately 0.44 initially to 0.21 in the final stage, signaling accelerated adaptation to target outputs. The evaluation loss similarly diminished (from roughly 0.35 to 0.28), though more gradually, while maintaining proximity to training metrics. This synchronized improvement pattern indicates the model acquired generalizable knowledge rather than memorizing specific examples—had the evaluation loss plateaued or risen while learning continued advancing, this would signal problematic overfitting. The modest separation between both loss values remained encouraging. The ultimate metrics (0.2062 for learning, 0.2767 for evaluation) represent satisfactory model performance. This reliable progression reflects our controlled hyperparameter selections, the manageable scale of our dataset, the consistency of journalistic summaries, and our regularization techniques.

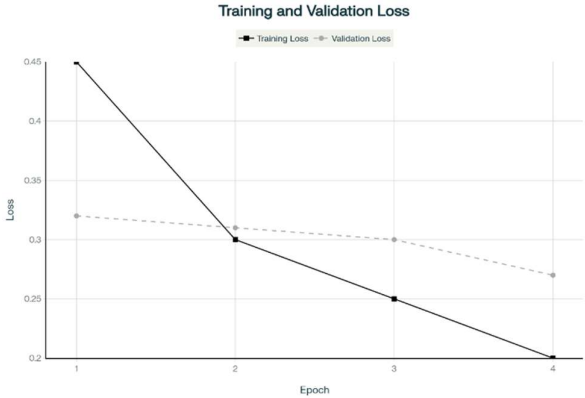


Fig. 2. Training and validation loss over 4 epochs for the T5 summarization model.

Following model customization, we assessed performance using standard evaluation benchmarks on held-out test material. Our refined Light Reads achieved ROUGE-1 performance of 0.9239, ROUGE-2 of 0.8670, and ROUGE-L of 0.9104. For comparison, we simultaneously measured the foundational pretrained model without task-specific adaptation, which yielded baseline measurements of approximately ROUGE-1 0.9038, ROUGE-2 0.8347, and ROUGE-L 0.8870 on equivalent evaluation material. Our customized version successfully recognized and preserved essential narrative elements (such as policy announcements, investment commitments, technological updates) while expressing them through natural language variations. While the system occasionally employed alternative terminology (such as "announced" versus "unveiled"), this stylistic variance is characteristic of abstractive approaches. The generated outputs maintained semantic fidelity to original material and effectively synthesized complex information into digestible form.



Figure 3. ROUGE score improvement with fine-tuning.

Figure 3 illustrates the improvement in ROUGE scores after fine-tuning. The model's generated summary closely aligns with the reference summary, capturing key information (e.g., "50% increase in solar/wind capacity" and "investment in battery storage and grid modernization") while using slightly different phrasing. This is typical of abstractive summarization, where the model rephrases key ideas in a coherent manner. Despite some variations in wording (e.g., "announced a new policy" vs. "unveiled its new policy"), the output maintains factual consistency with the original article and successfully distills the core message.

To facilitate user interaction with the summarization model, we developed a simple Gradio-based interface. The interface features a top text box where users can input the article they wish to summarize. Below the text box, a "Summarize" button triggers the backend model to generate a summary. The resulting summary is displayed in a lower text area labeled "Summary Output". This design enables users to easily obtain summaries without requiring any technical knowledge or coding experience.

In summary, the experimental results demonstrate that our fine-tuned T5 model performs well on news article summarization tasks, significantly improving over its pre-trained state and generating summaries that closely match human-written summaries. The combination of quantitative evaluation metrics and real-world examples confirms the effectiveness and reliability of the Light Reads system.

V. CONCLUSIONS

Light Reads represents an abstractive summarization framework that adapts the T5 transformer architecture to generate condensed versions of news articles. The initiative successfully mitigates information overload challenges by delivering a mechanism to reduce extensive articles into their core elements, thereby expediting the comprehension of critical details. The adapted model achieved strong ROUGE performance metrics, demonstrating that produced summaries retain significant information while maintaining structural consistency. The inclusion of an accessible interface confirms the system's applicability in practical environments, allowing individuals to acquire summaries efficiently without requiring specialized technical expertise. The implementation incorporates contemporary NLP methodologies, including transformer-based architectures and beam search with diversity enhancement, showcasing their practical effectiveness.

In summary, Light Reads exemplifies how customizing a sophisticated pretrained language model yields an effective summarization framework suited to specialized domains. Through the fusion of cutting-edge NLP technologies with user-centric design principles, this work provides a pragmatic utility that assists individuals in managing extensive textual content. The authors anticipate that analogous summarization technologies will grow in significance and pledge continued refinement of Light Reads to accommodate advancing requirements in rapid information assimilation and comprehension.

Further refinement could introduce personalization capabilities enabling users to customize summaries according to their specific needs—adjusting technical depth, emphasizing particular entities, or preserving sentiment—transforming the system from a one-size-fits-all tool into an adaptive solution serving diverse populations. Interactive refinement features would permit collaborative summarization through dialogue, allowing users to request section expansion, passage simplification, or alternative phrasings. Hybrid extractive-abstractive approaches merit investigation, potentially combining selected source passages with abstractive enhancement to preserve important language while reducing computational demands.

This work exemplifies how appropriately configured transformer architectures paired with user-focused design effectively tackle genuine information management challenges. As digital content continues proliferating exponentially, such accessible NLP solutions become increasingly essential. Light Reads demonstrates that powerful, effective summarization tools need not require prohibitive complexity or resources. By integrating technical sophistication with practical usability, this research contributes to developing accessible AI systems that genuinely serve diverse end-users, advancing the broader mission of making advanced language technologies available and beneficial to broad audiences navigating information complexity.

ACKNOWLEDGMENT

The authors wish to thank SIT Tumkur.

REFERENCES

- [1] Hugging Face, "T5: Text-to-Text Transfer Transformer," 2020. [Online]. Available: https://huggingface.co/transformers/model_doc/t5.html.

- [2] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” in Proc. EMNLP, 2019, pp. 3730–3740. DOI: 10.18653/v1/D19-1387.
- [3] R. Shanthakumari, E. M. Roopa Devi, S. Vinothkumar, T. Sabari, M. Sruthi, and T. Subaranjana, “News Article Summarization using PEGASUS model for Efficient Information Consumption,” in Proc. 15th ICCCNT, 2024, pp. 1–6. DOI: 10.1109/ICCCNT61001.2024.10724977.
- [4] M. H. Asif and A. U. Yaseen, “Comparative Evaluation of Text Similarity Matrices for Enhanced Abstractive Summarization on CNN/DailyMail Corpus,” *Journal of Computing & Biomedical Informatics*, vol. 6, no. 1, pp. 208–215, 2023.
- [5] D. Yadav et al., “[Retracted] Qualitative Analysis of Text Summarization Techniques and Its Applications in Health Domain,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3411881, 2022.
- [6] P. Cui, L. Hu, and Y. Liu, “Enhancing Extractive Text Summarization with Topic-Aware Graph Neural Networks,” arXiv:2010.06253, 2020.
- [7] W. Yuan, G. Neubig, and P. Liu, “BARTScore: Evaluating Generated Text as Text Generation,” in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 27263–27277.
- [8] V. Veselovsky, M. H. Ribeiro, and R. West, “Artificial artificial intelligence: Crowd workers widely use large language models for text production tasks,” arXiv:2306.07899, 2023.
- [9] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, “DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization,” in Proc. AAAI, vol. 36, no. 10, 2022, pp. 11765–11773.
- [10] A. A. Syed, F. L. Gaol, and T. Matsuo, “A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization,” *IEEE Access*, vol. 9, pp. 13248–13265, 2021.
- [11] A. G. Etemad, A. I. Abidi, and M. Chhabra, “Fine-tuned T5 for Abstractive Summarization,” *Int. J. Performability Eng.*, vol. 17, no. 10, p. 900, 2021.
- [12] C. M. Muia, A. M. Oirere, and R. N. Ndungu, “A Comparative Study of Transformer-based Models for Text Summarization of News Articles,” 2024.
- [13] J. L. E. K. Fendji, D. M. Taira, M. Atemkeng, and A. M. Ali, “WATS-SMS: A T5-Based French Wikipedia Abstractive Text Summarizer for SMS,” *Future Internet*, vol. 13, no. 9, Art. 238, 2021.
- [14] T. Goyal, J. Xu, and J. J. Li, “Training Dynamics for Text Summarization Models,” arXiv:2110.08370, 2022.