Biomedical Abstract Simplification Using Large Language Models (LLMs)

G. Thrishul[1], B. Bharath Kumar[2], K. Sai Karthik[3], Ch. Sai Kiran[4], D.Kavitha[5]

[1234]Department of Information Technology, TKR College of Engineering and Technology, Telangana, India
[5]Assistant Professor, Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

---

Abstract - *Biomedical research articles often contain highly complex terminology and sentence structures, making them difficult to comprehend for non-expert readers, patients, and interdisciplinary researchers. This creates a significant accessibility gap between advanced medical research and its practical understanding. To address this challenge, this paper proposes an intelligent biomedical text simplification system based on a transformer-driven sequence-to-sequence architecture. The proposed system leverages a FLAN-T5 encoder–decoder model to convert complex biomedical abstracts into simplified, readable text while preserving core semantic meaning. The system incorporates preprocessing, tokenization, dense vector embeddings, contextual encoding, and beam search–based decoding to generate high-quality simplified outputs. Additionally, multiple simplification levels—mild, medium, and strong—are supported to cater to diverse user requirements. A Stream lit-based user interface enables real-time interaction and visualization of results. Experimental observations demonstrate that the proposed approach effectively enhances readability without significant loss of informational content, making biomedical literature more accessible and user-friendly.*

***Key Words*: Biomedical Text Simplification, Natural Language Processing, Transformer Models, FLAN-T5, Text Preprocessing, Beam Search Decoding, Stream lit Application**

1. INTRODUCTION

1.1 Background and Motivation

The rapid growth of biomedical research has resulted in an exponential increase in scientific publications, clinical reports, and healthcare documentation. While these resources are invaluable for medical professionals, they often contain complex terminology, dense sentence structures, and domain-specific expressions that are difficult to understand for non-expert readers, patients, and interdisciplinary researchers. This lack of accessibility creates a significant barrier between biomedical knowledge and its effective utilization.

Natural Language Processing (NLP) techniques have emerged as a powerful solution to bridge this gap by enabling automated understanding and transformation of textual data. Among these, text simplification has gained considerable attention as it focuses on reducing linguistic complexity while preserving the original semantic meaning. In the biomedical domain, effective text simplification can improve knowledge dissemination, patient education, and cross-domain collaboration, making it a critical research area.

1.2 Problem Statement

Despite advancements in NLP, simplifying biomedical text remains a challenging task due to the presence of specialized vocabulary, long compound sentences, and context-sensitive meanings. Traditional rule-based and statistical approaches often fail to preserve critical medical information or produce oversimplified outputs that distort the original intent. Additionally, many existing systems lack flexibility in controlling the level of simplification, limiting their applicability to diverse user groups.

There is a clear need for an intelligent and adaptive system capable of simplifying biomedical abstracts while maintaining contextual accuracy, semantic integrity, and readability. Such a system should also provide an interactive interface to allow users to experiment with different simplification levels in real time.

1.3 Objectives of the Proposed System

The primary objective of the proposed system is to develop an intelligent and reliable biomedical text simplification framework that can automatically transform complex biomedical abstracts into simplified and easily understandable text. The system aims to reduce linguistic complexity while preserving the original semantic meaning and critical medical information, ensuring that the simplified output remains informative and contextually accurate. By focusing on biomedical abstracts, the proposed approach targets a highly specialized and information-dense form of text that presents unique challenges in natural language processing.

Another key objective of this work is to leverage recent advancements in transformer-based architectures to improve contextual understanding and text generation quality. The proposed system utilizes a FLAN-T5 encoder–decoder model, which is specifically designed to handle sequence-to-sequence tasks efficiently. By exploiting its ability to capture long-range dependencies and contextual relationships, the system aims to generate coherent and fluent simplified text that aligns closely with the intent of the original abstract.

The system also seeks to provide flexibility by allowing users to control the degree of simplification based on their individual requirements. Different user groups, such as medical students, patients, and interdisciplinary researchers, may require varying levels of simplification. To address this need, the proposed system supports multiple simplification levels, enabling users to choose outputs ranging from minimally simplified text to more aggressively simplified versions. This adaptability significantly enhances the practical usability of the system.

In addition, the proposed work aims to bridge the gap between advanced natural language processing models and real-world applications by incorporating an interactive web-based interface. The integration of a Stream lit-based user interface allows users to input biomedical abstracts and obtain simplified outputs in real time. This design choice emphasizes usability and accessibility, ensuring that the system can be easily adopted by users without technical expertise in machine learning or natural language processing.

1.4 Organization of the Paper

This paper is structured to provide a clear and systematic presentation of the proposed biomedical text simplification system. Following the introduction, the second section presents a detailed literature survey that reviews existing research and methodologies related to text simplification and biomedical natural language processing. This section highlights the strengths and limitations of current approaches and identifies the research gaps that motivate the proposed work.

The third section describes the overall architecture of the proposed system, detailing each functional module involved in the text simplification process. This includes discussions on text preprocessing, tokenization, embedding generation, transformer-based encoding and decoding, and beam search optimization. The system architecture is explained with reference to the workflow diagram to provide a clear understanding of data flow and component interactions.

2. Literature Survey

2.1 Existing Text Simplification Techniques

Text simplification has been an active area of research in the field of natural language processing, aiming to reduce textual complexity while preserving the original meaning. Early approaches to text simplification relied heavily on rule-based techniques, where predefined linguistic rules were used to replace complex words and restructure sentences. Although these methods provided some level of interpretability, they were highly dependent on handcrafted rules and lacked scalability, particularly for domain-specific texts such as biomedical literature.

With the advancement of machine learning, statistical and corpus-based methods were introduced to overcome the limitations of rule-based systems. These approaches utilized parallel corpora consisting of complex and simplified text to learn transformation patterns. While statistical methods demonstrated improved performance over rule-based techniques, they often struggled with contextual understanding and failed to handle long-range dependencies effectively, which are common in biomedical abstracts.

Recent developments in deep learning have significantly transformed text simplification research. Neural network–based models, particularly sequence-to-sequence architectures, have shown promising results by learning end-to-end mappings between complex and simplified text. The introduction of transformer-based models further enhanced performance by enabling better contextual representation and parallel processing. These models have demonstrated superior fluency and semantic preservation compared to traditional approaches, making them suitable candidates for complex domains such as biomedical text simplification.

2.1 Biomedical Text Simplification Approaches

Biomedical text simplification presents additional challenges due to the presence of specialized terminology, abbreviations, and domain-specific expressions. Several studies have explored the use of domain-adapted word embeddings and medical ontologies to address these challenges. Ontology-driven approaches attempt to replace complex medical terms with simpler equivalents using curated biomedical knowledge bases. However, such methods are often limited by the availability and coverage of domain-specific resources.

Neural approaches have gained popularity in biomedical text simplification due to their ability to learn contextual representations directly from data. Pretrained language models fine-tuned on biomedical corpora have shown improved performance in handling medical terminology and sentence structure. Encoder–decoder transformer models, in particular, have been effective in generating simplified biomedical text while maintaining semantic consistency. Despite these advancements, many existing models lack flexibility in adjusting simplification intensity based on user needs.

2.3 Limitations of Existing Systems

Although significant progress has been made in text simplification, existing systems still suffer from several limitations. Many approaches focus primarily on lexical simplification and fail to adequately address syntactic complexity, resulting in outputs that remain difficult to comprehend. Additionally, some models tend to oversimplify text, leading to the loss of critical biomedical information, which is unacceptable in medical contexts.

2.4 Research Gap Identification

From the reviewed literature, it is evident that while transformer-based models have improved the quality of text simplification, there remains a gap in developing systems that balance readability, semantic preservation, and user adaptability. Existing approaches often prioritize model performance without considering real-world usability or accessibility. Moreover, limited attention has been given to providing multiple levels of simplification within a single unified framework.

The identified research gap lies in designing an end-to-end biomedical text simplification system that combines advanced transformer architectures with flexible simplification control and an interactive user interface. Addressing this gap can significantly enhance the accessibility of biomedical literature and support a wider range of users, from healthcare professionals to non-expert readers.

3. Proposed System Architecture

3.1 Overall System Overview

The proposed biomedical text simplification system is designed as an end-to-end pipeline that transforms complex biomedical abstracts into simplified and readable text using a transformer-based sequence-to-sequence model. The architecture follows a modular design approach, enabling efficient preprocessing, contextual understanding, controlled simplification, and output generation. Each module in the system is responsible for a specific function, ensuring scalability, maintainability, and clarity in data flow.

The system accepts biomedical abstracts as input through a web-based interface and processes them through multiple stages, including text preprocessing, tokenization, embedding generation, transformer-based encoding and decoding, and final output generation. The integration of a FLAN-T5 model at the core of the architecture enables the system to capture contextual semantics and generate coherent simplified text. Additionally, beam search decoding is employed to enhance output quality by selecting the most probable simplified sequences.

3.2 User Interface Module

The user interface module serves as the interaction layer between the user and the proposed system. It is implemented using the Stream lit framework, which provides a lightweight and responsive web-based environment for real-time text input and output visualization. Through this interface, users can submit biomedical abstracts and select the desired level of simplification based on their comprehension needs.

This module is designed to be intuitive and accessible, allowing users with minimal technical background to interact with the system effectively. By enabling instant feedback and simplified text display, the interface bridges the gap between complex backend processing and practical usability, thereby enhancing the overall user experience.

3.3 Text Preprocessing Module

The text preprocessing module prepares the input biomedical abstract for further processing by removing noise and ensuring consistency in text format. This stage involves operations such as lowercasing, removal of special characters, normalization of whitespace, and sentence segmentation. These preprocessing steps are critical for reducing variability in the input data and improving model performance.

By standardizing the input text, the preprocessing module ensures that the transformer model receives clean and structured data. This contributes to better tokenization efficiency and more accurate contextual representations during subsequent stages of processing.

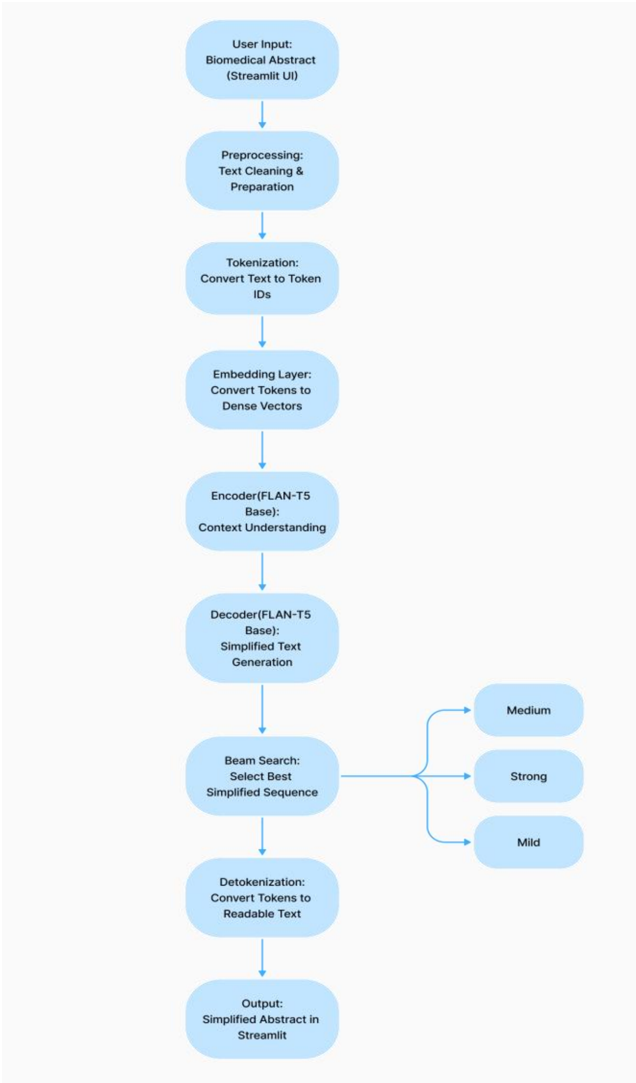

Fig -3: System architecture of the blockchain-based peer-to-peer energy trading platform

3.5 FLAN-T5 Encoder–Decoder Model

The core of the proposed system is the FLAN-T5 encoder–decoder architecture, which performs the task of contextual understanding and text generation. The encoder processes the input embeddings and generates a contextual representation that captures the meaning of the entire biomedical abstract. This representation is then passed to the decoder, which generates the simplified text in a sequential manner.

The use of FLAN-T5 enables the system to leverage instruction-tuned learning, allowing it to adapt effectively to the task of text simplification. This architecture is particularly well-suited for sequence-to-sequence transformations, making it an optimal choice for simplifying complex biomedical language.

3.6 Beam Search Optimization

To improve the quality of the generated simplified text, beam search decoding is employed during the output generation phase. Unlike greedy decoding, beam search maintains multiple candidate sequences at each decoding step and selects the most probable sequence based on cumulative likelihood scores.

This optimization technique helps reduce grammatical errors and incoherent sentence structures in the output. By balancing exploration and exploitation during decoding, beam search enhances the fluency and readability of the simplified biomedical text.

3.7 Output Generation Module

The final module of the system is responsible for generating and displaying the simplified biomedical abstract to the user. Based on the selected simplification level, the decoder produces an output that aligns with the desired complexity reduction while preserving essential information.

The generated text is then rendered on the Stream lit interface, allowing users to review and reuse the simplified content. This module completes the end-to-end workflow, ensuring that the system delivers a seamless and efficient text simplification experience.

4. Methodology

4.1 System Workflow and Processing Pipeline

The methodology of the proposed biomedical text simplification system follows a structured and sequential workflow designed to ensure accurate transformation of complex biomedical abstracts into simplified text. The process begins with user input through the web-based interface, where the biomedical abstract is submitted for simplification. Once received, the input text is forwarded to the preprocessing stage to eliminate inconsistencies and prepare the text for model ingestion.

After preprocessing, the standardized text flows through the tokenization and embedding stages, where it is converted into numerical representations suitable for transformer-based processing. These representations are then passed through the encoder–decoder architecture, which performs contextual analysis and generates simplified text. The final output is refined using beam search decoding and displayed to the user. This pipeline ensures smooth data flow, modular processing, and reliable output generation.

4.2 Text Preprocessing Strategy

Text preprocessing plays a critical role in enhancing the performance and reliability of the proposed system. Biomedical abstracts often contain irregular formatting, special symbols, and complex sentence structures that can negatively impact model performance if not handled properly. To address this, the preprocessing stage performs normalization operations such as lowercasing, removal of unnecessary symbols, and whitespace correction.

Sentence segmentation is also applied to improve the model's ability to process long and information-dense abstracts. By breaking the input into manageable linguistic units, the preprocessing strategy ensures that the downstream transformer model can focus on semantic understanding rather than structural inconsistencies. This step significantly contributes to the quality and coherence of the simplified output.

4.3 Model Configuration and Training Strategy

The proposed system employs a FLAN-T5 transformer model configured in an encoder–decoder setup for sequence-to-sequence text simplification. The encoder is responsible for capturing contextual representations of the biomedical abstract, while the decoder generates the simplified version based on the learned context. The model benefits from instruction tuning, enabling it to adapt effectively to the task of biomedical text simplification.

Although the system primarily leverages a pretrained model, task-specific configuration is applied to optimize performance. Parameters such as maximum input length, output length, and decoding strategy are carefully selected to balance simplification quality and semantic preservation. This configuration ensures that the model generates fluent and meaningful simplified text without omitting critical biomedical information.

4.4 Simplification Level Control Mechanism

A key methodological contribution of the proposed system is the incorporation of multiple levels of text simplification. Instead of producing a single fixed output, the system allows users to select the desired simplification intensity based on their comprehension needs. This control mechanism enables the generation of mildly simplified text that closely resembles the original abstract, as well as more aggressively simplified text for non-expert users.

The simplification level selection influences the decoding behavior of the transformer model, particularly during beam search optimization. By adjusting decoding parameters, the system effectively controls sentence complexity, vocabulary choice, and output length. This adaptability enhances the usability and applicability of the system across diverse user groups.

4.5 Implementation Environment

The implementation of the proposed system is carried out using a modern and scalable software stack. The backend processing is implemented in Python, leveraging deep learning libraries for model execution and text processing. The transformer model is integrated into the system using established NLP frameworks that support efficient inference and deployment.

The user interface is developed using Stream lit, enabling rapid prototyping and real-time interaction. The overall implementation environment is designed to support modular development, easy maintenance, and potential future expansion, ensuring that the system can be extended with additional features or integrated into larger biomedical information platforms.

5. Results and Discussion

5.1 Output Analysis of the Simplified Text

The performance of the proposed biomedical text simplification system is evaluated through qualitative analysis of the generated simplified outputs. The system is tested using multiple biomedical abstracts containing complex terminology and long sentence structures. The simplified outputs demonstrate a noticeable reduction in linguistic complexity while retaining the core semantic meaning of the original text. Medical terms are presented in a more readable form, and sentence structures are simplified to improve overall comprehension.

The transformer-based FLAN-T5 model effectively captures contextual relationships within biomedical abstracts, enabling coherent and fluent text generation. Compared to the original input, the simplified text exhibits improved readability, reduced sentence length, and clearer expression of ideas. These observations indicate that the proposed system successfully achieves its primary objective of enhancing accessibility without compromising informational value.

In addition, the generated outputs maintain logical flow and contextual consistency across sentences, which is critical in biomedical text interpretation. The simplification process avoids abrupt sentence fragmentation and preserves the explanatory structure of the original abstract. This demonstrates the model's ability to balance simplification with semantic continuity, making the output suitable for educational and informational purposes.

5.2 Impact of Simplification Levels

One of the key strengths of the proposed system is its ability to generate simplified text at multiple levels of complexity. When mild simplification is selected, the output remains close to the original abstract while reducing minor linguistic complexities. This level is particularly suitable for readers with basic biomedical knowledge, such as undergraduate students or interdisciplinary researchers.

In contrast, medium and strong simplification levels produce more concise and reader-friendly outputs by further reducing sentence complexity and substituting difficult terminology with simpler expressions. These levels are effective for non-expert users, including patients and general readers, who may lack familiarity with biomedical language. The availability of multiple simplification levels significantly enhances the flexibility and practical usefulness of the system.

Furthermore, the differentiation between simplification levels allows the system to adapt to diverse comprehension needs without requiring separate models or pipelines. This unified approach ensures consistency in output quality while offering customization. Such adaptability is essential in biomedical communication, where the same content may need to be interpreted differently depending on the target audience.

5.3 Discussion on System Effectiveness

The results highlight the effectiveness of transformer-based architectures in handling complex biomedical text simplification tasks. The use of beam search decoding contributes to improved grammatical structure and output coherence compared to greedy decoding approaches. Additionally, the integration of a user-friendly interface enables real-time interaction, making the system practical for everyday use.

The modular design of the system further enhances its effectiveness by allowing individual components to function independently while contributing to the overall workflow. Each stage, from preprocessing to output generation, plays a role in maintaining text quality and readability. This structured design ensures that errors or inconsistencies at one stage do not significantly degrade the final output.

However, the system also faces certain limitations. In some cases, highly specialized biomedical terms may still appear in the simplified output due to the need to preserve semantic accuracy. Despite this, the overall performance of the system demonstrates a strong balance between simplification and information retention. These findings validate the suitability of the proposed approach for improving the accessibility of biomedical literature.

6. Conclusion and Future Scope

6.1 Conclusion

This paper presented an intelligent biomedical text simplification system based on transformer-driven encoder–decoder architecture. By leveraging the FLAN-T5 model, the proposed system effectively transforms complex biomedical abstracts into simplified and readable text while maintaining semantic integrity. The modular design, combined with preprocessing, beam search optimization, and user-controlled simplification levels, ensures both performance and adaptability.

The qualitative results demonstrate that the system enhances readability and accessibility of biomedical content, making it suitable for a wide range of users. The integration of a Stream lit-based interface further strengthens the practical applicability of the system by enabling real-time interaction and visualization of simplified outputs.

6.2 Future Enhancements

Future work can focus on extending the system to support full-length biomedical articles rather than abstracts alone. Incorporating domain-specific medical ontologies and evaluation metrics such as readability scores could further improve output quality. Additionally, multilingual support and integration with healthcare information systems could broaden the impact and usability of the proposed solution.

REFERENCES

[1] Y. Zhao, X. Wang, and Z. Liu, “Neural text simplification with semantic consistency,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 738–745, 2019.

[2] C. Scarton, G. Paetzold, and L. Specia, “Simplification of scientific abstracts using BERT and reinforcement learning,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3004–3016, 2020.

[3] Y. Guo, Z. Jin, and Q. Zhang, “Lay summarization of biomedical texts with readability control,” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 415–425, 2021.

[4] B. Ondov, M. Berger, and A. Johnson, “Improving readability of biomedical literature using GPT,” Journal of Biomedical Informatics, Elsevier, vol. 132, pp. 104123, 2022.

[5] Y. Guo, Z. Jin, and Q. Zhang, “PLABA: A dataset for lay-friendly biomedical abstracts,” Proceedings of the 2023 Conference on Computational Natural Language Learning (CoNLL), pp. 210–220, 2023.