AI Powered Web Content Summarizer

Himaja BM[1], Adithyan V Nair[2]
[1]*Assistant Professor, CSE Department, Sarabhai Institute of Science and Technology*
[2]*Student, Sarabhai Institute of Science and Technology*
doi.org/10.64643/IJIRTV12I10-204454-459

***Abstract*—The exponential increase in digital information necessitates the development of effective tools capable of deriving meaningful insights from extensive text corpora.** **This paper presents the AIWOCS (AI Web Content Summarizer and Analyzer), a web-based application designed to generate concise and context-aware summaries from multiple input sources, including raw text, web URLs, and uploaded documents.**
The system utilizes transformer-based AI models through external inference APIs to perform abstractive summarization. To handle large inputs efficiently, a chunk-based processing mechanism is implemented, enabling scalable summarization without loss of context. Additionally, the system provides keyword extraction and summary evaluation metrics, such as compression ratio and word count analysis.
The application is developed using a Node.js backend and an interactive frontend interface, enabling real-time summarization. The experimental results demonstrate that the system produces accurate, readable, and structured summaries across different types of input data. The proposed system highlights the practical use of modern AI techniques for content analysis and information compression.

***Index Terms*—AI summarization, natural language processing, web content extraction, transformer models, text analysis**

I. INTRODUCTION

The exponential increase in digital content across websites, documents, and online platforms has made it difficult for users to extract relevant information efficiently. Reading lengthy articles, reports, or documents is time-consuming and often impractical for researchers. This creates a need for automated systems capable of generating concise summaries without losing their essential meaning.
Text summarization techniques have evolved from traditional extractive methods to advanced abstractive approaches that use artificial intelligence. Extractive methods select important sentences from the original text, whereas abstractive methods generate new sentences that better capture the context and meaning of the content.
This project introduces AIWOCS, an AI-powered web content summarizer that supports multiple input formats, such as raw text, web URLs, and uploaded files. The system integrates preprocessing, AI-based summarization, and analytical output into a single platform. It aims to provide users with fast, accurate, and meaningful summaries of various types of content.

II. PROCEDURE

A. System Development Stage

The system was developed as a web-based application, consisting of front-end and back-end components. The backend handles data processing, AI integration, and API communication, whereas the frontend provides an interactive user interface for input and output visualization.

B. Implementation Stage

The implementation involves multiple stages, including text extraction, preprocessing, chunk-based summarization, and result generation. The system ensures that different input formats are handled efficiently and converted into readable texts before processing.

C. Output Generation

The system generates structured outputs, including summarized text, extracted keywords, and performance metrics. The outputs are displayed dynamically on the user interface.

III. MATH

The system evaluates the summarization performance using basic mathematical metrics.

Compression Ratio

Compression Ratio is calculated as:

$$Compression\ Ratio = \frac{Original\ Words - Summary\ Words}{Original\ Words} \times 100$$

Length Percentage

$$Length\ Percentage = \frac{Summary\ Words}{Original\ Words} \times 100$$

These metrics help in evaluating how effectively the system reduces the content while preserving meaning.

IV. UNITS

All measurements and calculations in this system were based on standard text-processing units, such as word count, sentence count, and percentage values. These units were used to evaluate the effectiveness and efficiency of the summarization process.

A. Input Handling

The AIWOCS system is designed to support multiple input formats to increase usability and flexibility for different types of users. The three main input methods are raw text input, web URL input, and file upload input.

In the case of raw-text input, the user directly enters the content into the system, which is then forwarded for preprocessing and summarization. For the URL input, the system performs web scraping to extract meaningful textual content from the webpage, removing unnecessary elements such as navigation menus, advertisements, and scripts. For file uploads, the system supports formats such as. txt, .pdf, and .docx, where the file content is parsed and converted into plain text, before further processing.

This multi-input capability ensures that the system can efficiently handle real-world use cases, such as summarizing articles, reports, academic documents, and online content.

B. Text Preprocessing

Text preprocessing is a critical step in ensuring the quality and accuracy of generated summaries. The system performs several preprocessing operations to clean and normalize input data.

These operations include removing HTML tags, scripts, and style elements for web-based inputs. Additionally, unnecessary whitespace, special characters and formatting inconsistencies were eliminated. The text was also normalized to ensure uniformity of structure and readability.

By performing these steps, the system ensures that only meaningful textual data are passed to the AI model. This significantly improved the quality of the generated summary and reduced the noise in the output.

C. Chunk-Based Processing

One of the key challenges in text summarization is the efficient handling of large input data. Transformer-based models have limitations on the maximum input size, making it difficult to process long documents in a single step.

To address this issue, the AIWOCS implements a chunk-based processing mechanism. In this approach, the input text is divided into smaller segments or “chunks” of a fixed size. Each chunk was then processed individually by the AI model to generate partial summaries.

These partial summaries are combined and, if necessary, further refined through additional summarization passes. This iterative reduction approach ensures that even very large documents can be summarized effectively without exceeding the model constraints.

This method not only improves scalability but also maintains the coherence and context of the final summary provided.

D. Keyword Extraction

In addition to generating summaries, the system extracts important keywords from the input text. Keyword extraction helps users quickly understand the main topics and themes of content.

The process involves removing common stop words such as “the,” “is,” and “and,” followed by analyzing the frequency of the remaining words. The most frequently occurring and contextually relevant words were selected as keywords.

These keywords are then displayed along with the summary, providing a quick overview of the content and improving the interpretability of the results

E. Summary Metrics and Analysis

To evaluate the effectiveness of the summarization process, the system calculates several performance metrics. These include the word count, sentence count, and compression ratio.

The compression ratio indicates the extent to which the original content was reduced while maintaining its

meaning. Additionally, the system checks whether the summary follows standard guidelines, such as maintaining an optimal length relative to the original text.

These metrics provide users with insights into the efficiency and quality of the generated summary, making the system more informative and analytical in nature.

V. APPLICATION

The AIWOCS was developed with a focus on practical applicability and reproducibility. The methodology used in the system is clearly defined, ensuring that the approach can be replicated and extended by other researchers in the future.

The system integrates multiple technologies, including web scraping, natural language processing, and AI-based summarization into a unified pipeline. This demonstrates the feasibility of combining different techniques to address real-world problems.

Furthermore, the use of transformer-based models highlights the importance of modern AI approaches for text-processing tasks. The system emphasizes scalability, usability, and adaptability, making it suitable for further research and development in the field of automated summarization.

This study adheres to standard academic practices by ensuring originality, proper structuring, and clear explanation of the implemented methods.

VI. CONCLUSION

This study presented the AIWOCS, an AI-powered web content summarization system capable of processing multiple types of input and generating meaningful summaries. The system effectively combines preprocessing techniques, chunk-based processing, and transformer-based AI models to handle both short- and long-text inputs.

The inclusion of keyword extraction and summary metrics enhances the overall functionality of the system, providing users with both summarized content and analytical insights into the content. The results indicate that the system can produce accurate, coherent, and context-aware summaries across various input formats.

Despite certain limitations, such as dependency on external AI APIs and increased processing time for large inputs, the system demonstrates strong potential as a practical tool for content analysis. Future improvements can further enhance performance, add multilingual capabilities, and integrate offline AI models for increased efficiency and independence.

Overall, AIWOCS represents a scalable and effective solution for modern information processing challenges

APPENDIX

Appendix A: Sample Input and Output

To demonstrate the functionality of the AIWOCS, a sample input and corresponding output are provided below.

Sample Input (Text):

Artificial Intelligence is transforming industries by enabling machines to perform tasks that typically require human intelligence. It is widely used in healthcare, finance, and education.

Generated Summary:

Artificial Intelligence enables machines to perform human-like tasks and is widely applied in industries such as healthcare, finance, and education.

Extracted Keywords:

Artificial Intelligence, machines, healthcare, finance, education

Metrics:

- Original Word Count: 30
- Summary Word Count: 18
- Compression Ratio: 40%
- This example illustrates how the system reduces the content while preserving the core meaning.

Appendix B: System Environment

The AIWOCS system was developed and tested in the following environments:

- Operating System: Windows / Linux
- Backend: Node.js with Express.js
- Frontend: HTML, CSS, JavaScript
- AI Integration: Hugging Face Inference API
- Deployment Platform: Cloud-based hosting (Render)

Appendix C: Limitations in Testing

During testing, certain constraints were observed.

- Large inputs require multiple processing steps, increasing response time
- Performance depends on external API availability and latency
- Webpage extraction may vary depending on site structure
- These limitations provide scope for future research.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the faculty members of the Department of Computer Science and Engineering at the Sarabhai Institute of Science and Technology for their continuous guidance and support throughout the development of this project. Their valuable suggestions and encouragement played crucial roles in the successful completion of this study.

REFERENCES

[1] D. Yadav, J. Desai, and A. K. Yadav, "Automatic text summarization methods: A comprehensive review," *arXiv preprint arXiv:2204.01849*, 2022. [Online]. Available: arXiv Paper PDF

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Proces*sing Systems (NeurIPS)*, 2017.

[3] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out Workshop, 2004.

[4] M. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith, "Toward abstractive summarization using semantic representations," in Proc. NAACL HLT Conf., 2015.

[5] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proc. Int. Conf. Machine Learning (ICML), 2020.

[6] Hugging Face, "Transformers documentation." [Online]. Available: Hugging Face Transformers Documentation

[7] Node.js Documentation

[8] Express.js Documentation

[9] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2004.

[10] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in Proc. 1987 INTERMAG Conf., 1987, pp. 2.2-1–2.2-6.