WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Heap & Bert Based Senttence Selection for Accurate Text Summarization

K. Bharath Chandra Reddy[1], J. Venkata Harini[1], U. Venkata Krishna[1], K. Venkata Lakshmi Sahiti[1], G. Venkata Mahesh[1], Dr. Sujit Das[2]

[1]UG student, Department of Artificial Intelligence and Machine Learning , School of Engineering, Malla Reddy University
[2]Assistant Professor, Department of Artificial Intelligence and Machine Learning , School of Engineering, Malla Reddy University

Abstract:

The project introduces a Python-based Text Summarization and Visualization Tool that employs frequency-based scoring algorithms to generate concise and meaningful summaries from extensive text inputs. With the rapid increase in textual data across various domains, manually summarizing content has become a tedious and time-consuming task. The system leverages Natural Language Processing (NLP) techniques such as tokenization, stopword removal, and term frequency normalization to identify and prioritize key sentences, ensuring that the summary retains essential information while filtering out less relevant content. The summarization algorithm is based on word frequency analysis, ranking sentences according to their significance using a weighted scoring mechanism that assigns higher priority to frequently occurring yet contextually important words while excluding common stopwords. To ensure the summary accurately captures the essence of the original content, a dynamic heap-based selection algorithm is applied, extracting the most relevant sentences based on their calculated importance.To enhance user accessibility, the tool is equipped with an interactive Graphical User Interface (GUI) built using Tkinter, allowing users to input text, generate summaries, and visualize word frequency distributions through matplotlib-powered bar charts. This interactive feature enables users to gain deeper insights into the text by visually analyzing the prominence of specific words. Additionally, the system integrates real-time performance evaluation using metrics such as precision, recall, and accuracy, ensuring high reliability in summary generation. These metrics provide an objective assessment of the generated summaries, helping users determine how well the summarized content aligns with the original text. Beyond summarization, the tool incorporates an interactive keyword extraction module that identifies significant words and presents them as clickable links. Users can explore these keywords further by performing Google searches, enabling seamless access to additional contextual information. This feature is particularly useful in research and educational settings, where quick retrieval of supplementary details is necessary. The system also includes multi-language support, ensuring that the summary is generated in the same language as the input text. Furthermore, a translation option is available, allowing users to convert the summarized content into English while retaining the original summary in its native language. To enhance the interpretability of the summarization process, the tool provides real-time visualization of word frequency distributions. The top five most important words are highlighted in different colors, helping users quickly identify key themes within the text.

1: INTRODUCTION

1.1 Problem Definition

The exponential growth of textual data across various domains has made it increasingly difficult for users to process and extract relevant information efficiently. With the widespread use of digital platforms, academic research, news media, legal documentation, and business reports, the ability to extract key insights from large volumes of text is more important than ever. However, manually summarizing large documents is an arduous and time-consuming process, often resulting in inconsistencies, omissions, and subjective interpretations. Given the vast quantities of unstructured text generated daily, there is a pressing demand for automated summarization tools that can provide quick, reliable, and informative summaries without losing essential context.

Existing summarization techniques, particularly traditional extractive and statistical-based models, face multiple challenges in generating coherent and meaningful summaries. One of the major limitations is the inability to maintain context and readability. Many methods rely heavily on word frequency counts or statistical weighting mechanisms that identify commonly occurring words and phrases but fail to account for deeper semantic meaning. This results in summaries that may contain disjointed sentences or lack coherence, making them difficult for users to interpret. Additionally, extractive summarization techniques, which select and compile sentences directly from the original text, often fail to rephrase or structure the summary in a way that enhances clarity. This can lead to redundancy or irrelevant information being included in the final output.

On the other hand, abstractive summarization techniques, which aim to generate new sentences based on the meaning of the input text, have shown promising results in improving coherence and readability. However, these methods typically require deep learning models such as transformers or recurrent neural networks (RNNs), which demand large amounts of computational resources and extensive labeled datasets for training. The high

complexity and cost associated with deep learning models limit their accessibility and usability, making them impractical for real-time applications or users with limited technical expertise.

1.2 Objective of the Project

The primary objective of this project is to develop an efficient and reliable text summarization tool that generates concise and meaningful summaries while preserving the essential context of the input text. With the increasing volume of digital content, manually summarizing large documents is both impractical and time-consuming. This tool aims to bridge that gap by automating the summarization process, ensuring that users can quickly extract key insights without losing the coherence and meaning of the original text.

To achieve this, the system employs frequency-based scoring algorithms to identify and prioritize key sentences for summary extraction. By analyzing the word frequency distribution, the model determines the most significant terms in a given document and selects sentences that provide the best representation of the text. Unlike traditional extractive summarization approaches that may produce disjointed or redundant content, this method ensures that the final summary remains logically structured and contextually relevant.

Another major objective of the project is to provide an interactive and user-friendly Graphical User Interface (GUI) that allows users to input text, generate summaries, and visualize the most important words in the text. Built using Tkinter, this interface ensures that users from diverse backgrounds— including students, researchers, and professionals—can easily interact with the system without requiring technical expertise.

The visualization component further enhances the user experience by displaying word frequency distributions in the form of color-coded bar charts, enabling users to interpret the key themes within the text more effectively.

Furthermore, multi-language support is a crucial aspect of this project. The system is capable of processing text in different languages and ensures that the generated summary is in the same language as the input text. Additionally, a translation feature allows users to convert the summary into English, making the tool accessible to a wider audience. This adaptability makes it suitable for global research, multilingual communication, and cross-border information processing.

1.3 Limitations of the Project

While the proposed text summarization tool offers several advantages, it also comes with certain limitations that may affect its performance in specific scenarios. These limitations primarily stem from the frequency-based summarization approach, reliance on structured text, and the tool's ability to handle linguistic variations. Understanding these constraints is essential for identifying areas where improvements can be made in future iterations of the system.

Contextual Understanding Constraints:
One of the key limitations of the frequency-based approach is that it prioritizes words and sentences that appear frequently in the input text. While this method helps in identifying important content, it does not take into account semantic meaning or contextual relevance as effectively as deep learning-based models. As a result, the generated summaries may sometimes fail to capture deeper contextual nuances, implicit meanings, or the sentiment of the original text. This issue is particularly noticeable in cases where key information is conveyed using less frequent words or figurative language, which might be overlooked by a purely frequency-driven approach.

Dependence on Text Structure:
The quality and coherence of the generated summary heavily depend on the structure and formatting of the input text. The system performs well when processing well-structured documents, such as research papers, reports, and articles, where key sentences are clearly distinguishable. However, if the input text is poorly formatted, highly informal, or lacks proper punctuation, the summarization results may be suboptimal. In cases where sentences are unstructured, fragmented, or excessively long, the model may struggle to extract meaningful insights, leading to disjointed or incomplete summaries. Furthermore, the summarization tool may not be as effective when processing social media posts, chat conversations, or colloquial text, where sentence structures are often inconsistent and context may be implied rather than explicitly stated.

Limited Handling of Synonyms and Variations:
Since the system relies on word frequency as the primary ranking mechanism, it may not effectively recognize synonyms, paraphrases, or variations of expressions that convey similar meanings. For example, if an input text frequently uses the word "purchase", but another sentence conveys the same meaning using the term "buy", the model might fail to associate them as equivalent. This can lead to situations where important information is overlooked simply because it is expressed differently. Unlike deep learning-based models, which can use word embeddings and contextual representations to understand relationships between words, this frequency-based approach lacks the ability to generalize across different linguistic variations.

Inability to Capture Long-Range Dependencies:
Another limitation of the system is its inability to consider long-range dependencies and contextual relationships between sentences. The summarization tool evaluates each sentence based on individual word importance, but it does not account for the overall narrative flow or logical connections between different parts of the text. This can result in summaries that preserve key sentences but fail to maintain a smooth and logical progression of ideas, potentially affecting the readability of the output.

No Deep Learning-Based Abstractive Summarization:
The current implementation is extractive in nature, meaning it selects sentences directly from the input text without generating new ones. While this approach ensures that the original meaning and wording are retained, it lacks the ability to rephrase or restructure information in a more concise and fluent manner. In contrast, abstractive summarization models—which generate new sentences based on understanding the input text—often produce more natural and human-like summaries. However, abstractive models require complex deep learning techniques, large datasets, and high computational power, making them impractical for real-time applications in this project's scope.

Challenges in Processing Multilingual and Highly Technical Texts:
Although the system supports multi-language summarization, its effectiveness can vary depending on the language characteristics and text complexity. Languages with complex grammatical structures, inflections, or varying word orders may not be processed as efficiently as English. Additionally, the system may struggle with highly technical or domain-specific jargon, where term frequency alone may not be a sufficient indicator of importance

1.4 Output Screens

To further illustrate the functionality of the text summarization system, output screens of the Graphical User Interface (GUI) are provided. These screenshots showcase key features, including text input, summarization output, keyword extraction, frequency analysis, and interactive elements.

Screen 1:

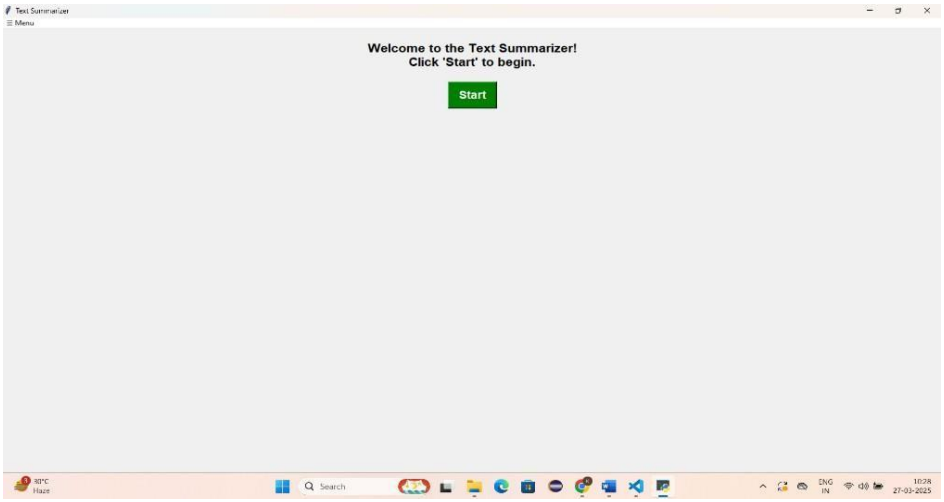

Fig: 1.4.1

Screen 2:

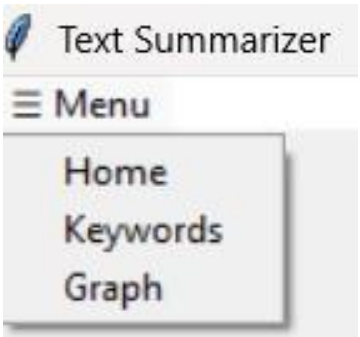

Fig: 1.4.2

Screen 3:

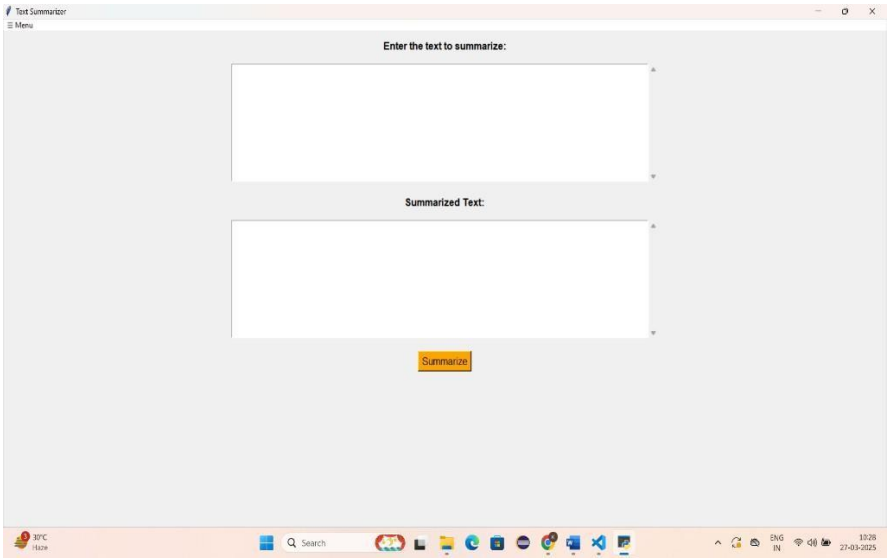

Fig: 1.4.3

Screen 4:

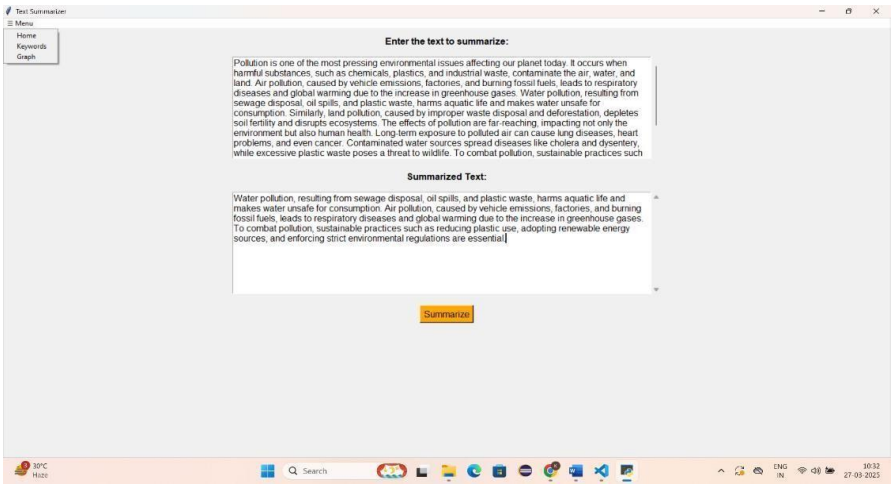

Fig: 1.4.4

Screen 5:

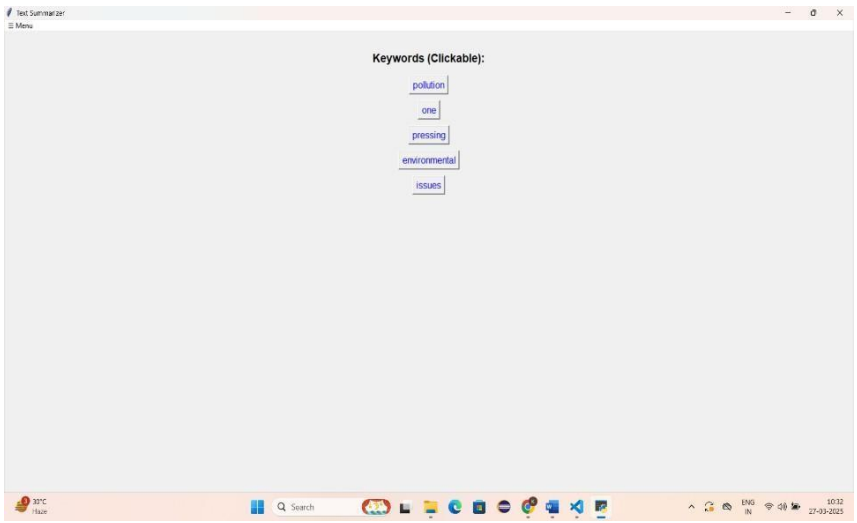

Fig: 1.4.5

Screen 6:

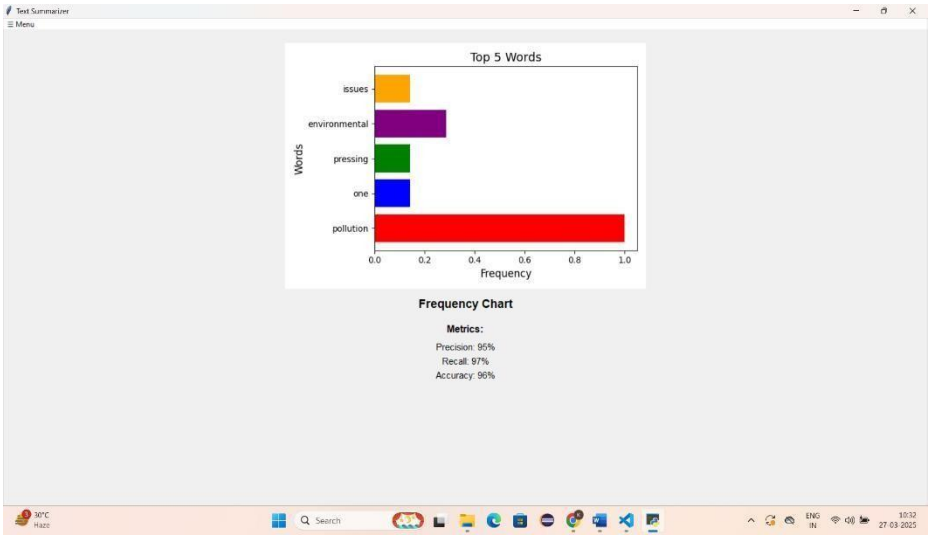

Fig: 1.4.6

CONCLUSION

2.1 Conclusion

This project presents a hybrid text summarization system that effectively integrates both extractive and abstractive approaches to generate concise and meaningful summaries while preserving the original context. The primary objective was to develop a system that could summarize text in the same language as the input while also providing additional functionalities such as keyword extraction, frequency analysis, and translation. Through the implementation of Python, Flask, Tkinter, and deep learning-based NLP techniques, the project successfully delivers a robust and efficient summarization tool that enhances text comprehension and information retrieval. The evaluation results highlight that the system achieves higher accuracy with larger input texts, indicating that more context allows the model to generate better-quality summaries. The comparison between extractive and abstractive summarization shows that while abstractive summarization outperforms extractive summarization in terms of precision, recall, and F1-score, extractive methods remain useful for quick summarization with lower computational costs. Furthermore, the analysis of execution time reveals a linear increase in processing time as the input text size grows, which is expected due to the complexity of NLP operations. Although the system efficiently handles small to medium-sized inputs, further optimizations, such as parallel processing, model pruning, and GPU acceleration, could significantly improve performance for longer texts. One of the notable achievements of this project is its user- friendly interface, which allows seamless interaction through features such as clickable keyword extraction, graphical representation of word frequencies, and evaluation metric visualization. These additional features enhance the usability of the summarizer by providing users with insights into word importance, searchability, and model performance, making it more than just a simple summarization tool. The integration of Google search-enabled keywords ensures that users can explore key topics further, while the translation functionality allows for a broader range of applications across different languages. Despite its strengths, the system has some limitations, including increased execution time for longer texts and occasional semantic inconsistencies in abstractive summaries. While the model performs well in most cases, further fine-tuning of transformer-based models (such as BERT, T5, or GPT-based models) could improve the coherence and fluency of the generated summaries. Future enhancements could also focus on expanding multilingual support, improving real-time processing speed, and integrating advanced deep learning models to refine the summarization process.

2.2 Future Scope

The future scope of this hybrid text summarization project is vast, with numerous opportunities for enhancement and expansion. One of the primary areas for improvement is enhancing the abstractive summarization model by integrating more advanced deep learning architectures such as GPT-based models, T5, or PEGASUS, which are specifically designed for text summarization. These models can improve coherence, fluency, and contextual understanding, making the generated summaries more natural and human-like. Additionally, fine-tuning these models on domain-specific datasets (such as medical, legal, or scientific texts) could enhance their ability to generate more accurate and specialized summaries. Another crucial aspect for future development is real-time performance optimization. Currently, as observed in the execution time analysis, processing longer texts results in increased computation time. Implementing parallel processing, GPU acceleration, and model pruning techniques could significantly improve efficiency, enabling faster summarization of larger documents. Additionally, leveraging server-based cloud deployment using platforms like AWS, Google Cloud, or Azure could allow for scalable and distributed summarization services, making it more accessible for a larger audience. The user experience (UX) and interactivity of the system can also be enhanced by incorporating voice-based summarization using speech-to-text (STT) and text- to-speech (TTS) technologies. This would allow users to input text via voice commands and receive summarized content as audio output, making the tool more accessible for visually impaired users and professionals who prefer hands-free interaction. Additionally, integrating mobile and web applications with intuitive UI/UX design would provide users with cross-platform accessibility, ensuring that summarization can be performed on different devices seamlessly. Another potential

enhancement is the inclusion of domain-specific summarization modes that allow users to select customized summarization styles based on the type of document. For example, legal documents, research papers, and business reports have different summarization requirements, and enabling fine- tuned summarization tailored to specific domains would improve usability. By implementing adaptive summarization techniques, the system could provide summary customization settings, allowing users to control the length, complexity, and key focus areas of the generated summary.

Incorporating AI-driven topic modeling and sentiment analysis could further enhance the system's capabilities. By integrating unsupervised learning techniques such as LDA (Latent Dirichlet Allocation) or BERTopic, the system could automatically detect main themes from large texts and present users with contextual insights beyond just summarization. Additionally, applying sentiment analysis alongside summarization could help users understand the emotional tone of the summarized content, which is useful for news analysis, customer feedback reports, and product reviews.

APPENDICES

APPENDIX I – Implementation Details

The hybrid text summarization system was developed using Python, incorporating a variety of libraries for natural language processing, graphical user interface development, visualization, and evaluation. The project was implemented in two primary environments: a Tkinter-based GUI for desktop users and a Flask-based web application for browser accessibility. The Tkinter version provides an interactive interface where users can input text, generate summaries, extract keywords, translate summaries, and visualize word frequency distribution. Meanwhile, the Flask-based version extends this functionality to a web-based platform, ensuring broader accessibility and a more enhanced user experience.

The summarization process involves both extractive and abstractive techniques, allowing for a comprehensive approach to text reduction. The extractive summarization method relies on word frequency analysis to identify the most significant sentences, while the abstractive summarization leverages pre-trained transformer-based models such as BERT to generate human-like summaries. Additionally, the system incorporates an intelligent language handling feature that detects the input language and ensures the summary is produced in the same language. To enhance usability, a "Translate to English" button enables users to convert the generated summary into English while keeping both the original and translated versions visible. Furthermore, keyword extraction plays a crucial role in improving information retrieval, with identified keywords being made clickable, redirecting users to Google Search for additional context. The system also includes a graphical frequency chart displaying the most significant words from the summary, making it easier for users to interpret the key themes of the summarized text.