IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

Text Summarization Using NLP

Prof. Priyanka Dhumal[1], Sudarshan Sutar[2], Indraneel Surve[3], Mirza Munawwar[4], Vishal Nanaware[5]
Assistant Professor, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
Zeal College of Engineering and Research, Pune, India

Abstract: *In the This research introduces a groundbreaking text summarization approach by combining BERT for extractive summarization and GPT for abstractive summarization. The synergy of these models results in a hybrid system that leverages the precision of extraction and the linguistic fluency of abstraction. Experimental results demonstrate the model's efficacy in producing high-quality summaries, showcasing its potential impact on information synthesis across diverse domains.*

Keywords: GPT, BERT

I. INTRODUCTION

In today's information age, we are inundated with an ever-expanding volume of textual data. From news articles and research papers to social media posts and legal documents, the sheer quantity of text available can be overwhelming. Amid this data deluge, the need for efficient methods to distill, condense, and extract meaningful information from text has become increasingly critical. This is where Text Summarization using Natural Language Processing (NLP) emerges as a transformative technology.

Text summarization is the process of automatically generating a concise and coherent summary of a longer text, while retaining its essential information and meaning. This technology is a cornerstone of NLP, a field at the intersection of artificial intelligence and linguistics that focuses on enabling computers to understand, interpret, and generate human language. The goal of text summarization is to make large volumes of text more manageable and accessible, catering tothe time constraints and information overload faced by individuals, researchers, and organizations.

Text summarization can be broadly categorized into two main approaches: extractive and abstractive summarization. Extractive summarization involves selecting sentences or phrases directly from the source text that are deemed most important or representative of its content. Abstractive summarization, on the other hand, goes a step further by generating summaries that may not be verbatim extracts but convey the same ideas using different words and structures, often resembling a human-authored summary.

For text summary, there are essentially two different methods:

- Extractive Summarization
- Abstractive Summarization

Extractive summarization is a technique in Natural Language Processing (NLP) that involves selecting and combining key sentences or passages directly from the source text to create a concise summary.Unlike abstractive summarization, which generates new sentences to capture the essential meaning, extractive summarization relies on identifying and extracting the most important information already present in the original content.

Abstractive summarization is a Natural Language Processing (NLP) technique that involves generating a summary of a document by paraphrasing and rephrasing the content in a way that captures the essential meaning while often introducing new language constructs. Unlike extractive summarization, which selects and combines existing sentences from the source text.

II. LITERATURE REVIEW

In the paper [1], Angel Hernandez-Castaneda , Rene Arnulfo Garcia-Hernandez By automatically generating the summarization process could become more adaptable and applicable to a wider range of domains and languages.

In the paper [2], Jiawen Jiang, Haiyang Zhang , Chenxu Dai1 has The benefits of a hybrid summarization approach, potential technical innovations, and experimental validation of their proposed enhancements.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

In the paper [3], Ayesha Ayub Syed , The most advanced neural models for abstractive text summarization, highlighting the role of neural networks in this task.
In the paper [4], Aniqa Dilaw,Muhamm-ad Usman Ghani Khan. their main aim A loss function is introduced to normalize the inconsistency between word-level and sentence level attentions.
In the paper [5], Ángel Hernández-Castañeda , René Arnulfo García-Hernández had Create A Practical Application For High-Quality Document Summaries.
In the paper [6] Youhyun Shin al had Multi-Encoder Transformer For Korean Abstractive Text Summarization
In the paper [7], Divakar Yadav,Rishabh Katna had a comprehensive overview of the current status of text summarizing approaches, techniques, standard datasets.
In the paper [8], Asmaa Elsaid , Ammar Mohammed. Using recent deep learning models to adopt them in Arabic summarization studies is an essential demands
In the paper [9], P. Mahalakshmi, Addition To Text Summerization Image Description Is Generated For The Visualized Entities That Exist In The Images
In the paper [10], M.F. Aklima Akter Lima, had This paper outlines extractive and abstractive text summarization technologies and provides a deep taxonomy of the ATS
In the paper [11], Heewon Jang Wooju Kim had Improve the quality of the summary statement by proposing a reward function used in text summarization based on RL .
In the paper [12], Jingwei Cheng , Fu Zhang , Xuyang Guo. they This paper proposes an automatic text summarization model, which extends traditional sequence to-sequence (Seq2Seq) neural text summarization modeL
In paper [13], Muhammad Yahya Saeed, Muhammad Awais, Approach Analyzes The Mesh Of Multiple Unstructured Documents And Generates A Linked Set Of Multiple Weighted Nodes By Applying Multistage Clustering.
In paper [14], Aswani Radhakrishnan , Dibyasha Mahapatra, Alex James had developed A Hardware Accelerator For Information Retrieval Using Memristive TF-IDF Implementation
In paper [15] Pratik K. Biswas Aleksandr Iakubovich, Indigenously Developed Method That Combines Topic Modeling And Sentence Selection With Punctuation Restoration.

III. METHODOLOGIES

For text summary, there are essentially two different methods:

- Extractive Summarization
- Abstractive Summarization

EXTRACTIVE SUMMARIZATION

Extractive summarization is a technique in Natural Language Processing (NLP) that involves selecting and combining key sentences or passages directly from the source text to create a concise summary. Unlike abstractive summarization, which generates new sentences to capture the essential meaning, extractive summarization relies on identifying and extracting the most important information already present in the original content.

ABSTRACTIVE SUMMARIZATION

Abstractive summarization is a Natural Language Processing (NLP) technique that involves generating a summary of a document by paraphrasing and rephrasing the content in a way that captures the essential meaning while often introducing new language constructs.

IV. SYSTEM DESIGN

The system design for the hybrid text summarization project involves integrating advanced NLP models (BERT for extractive summarization and GPT for abstractive summarization) with a user-friendly mobile application developed using the Flutter framework. The backend is developed in Python and connected to the frontend via Flask. This architecture ensures efficient processing, scalability, and a seamless user experience.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

V. SYSTEM DEVELOPMENT

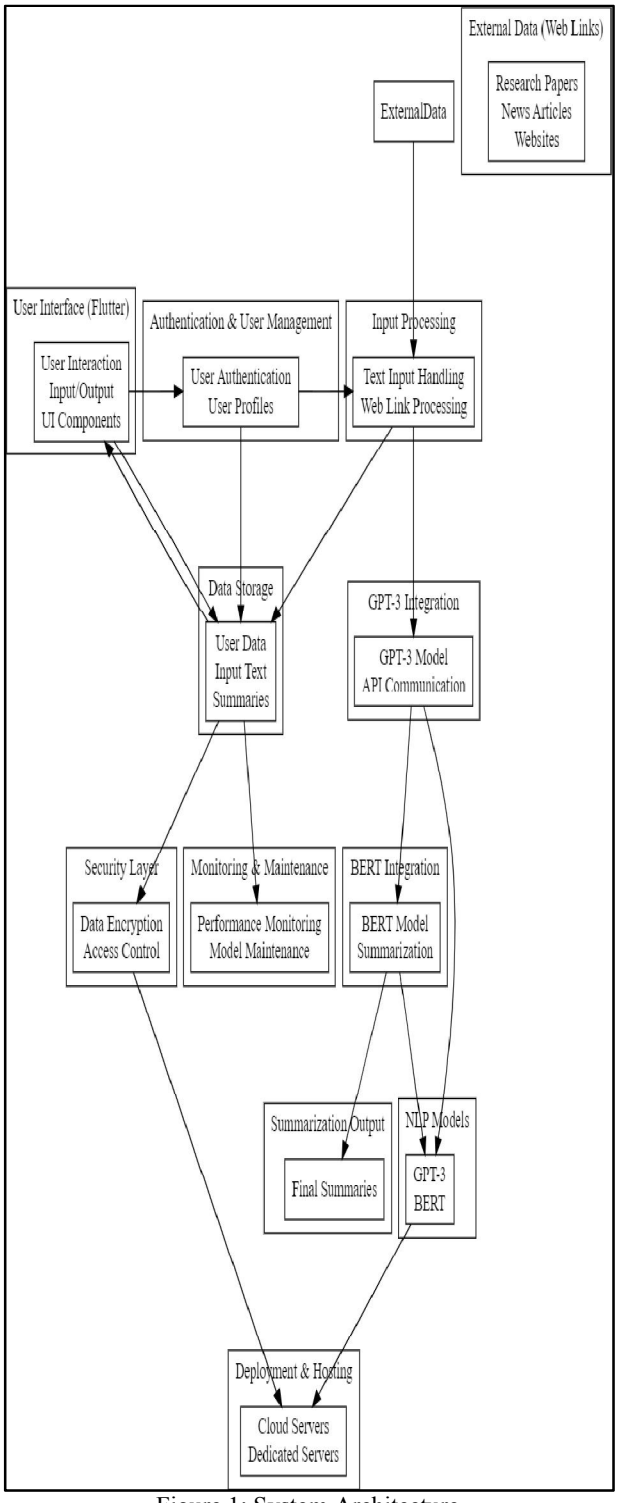

Figure 1: System Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

1. Input Module:

- User Interface: A mobile application built with Flutter that allows users to upload documents in various formats (plain text, PDFs, news articles, research papers).
- File Handling: Mechanisms to read and preprocess different document types, converting them into a standardized text format for further processing.

2. Backend Server:

- Framework: Flask is used to create a RESTful API that connects the mobile frontend with the backend services.
- Processing Logic: Handles requests from the frontend, manages data flow between components, and returns the generated summaries to the mobile app.

3. Extractive Summarization Module:

- Model: BERT (Bidirectional Encoder Representations from Transformers) is used to perform extractive summarization.
- Process: Tokenization and Encoding: Input text is tokenized and encoded using BERT's pre-trained model
- Sentence Ranking: Sentences are ranked based on relevance and importance.
- Selection: Top-ranked sentences are selected to form the extractive summary.

4. Abstractive Summarization Module:

- Model: GPT (Generative Pre-trained Transformer) is used for abstractive summarization.
- Process:Input Preparation: The extractive summary serves as input for the GPT model.
- Text Generation: GPT generates a coherent and concise abstractive summary by paraphrasing and synthesizing the input text.

5. Output Module:

- Summary Display: The generated summary is sent back to the mobile application and displayed to the user.
- Download Option: Users can download the summary in a desired format for offline use.

VI. USER EVALUATION

Users of the hybrid text summarization system can expect an intuitive and streamlined experience. The mobile application, developed using the Flutter framework, provides a user-friendly interface that allows users to easily upload and process documents in various formats such as plain text, PDFs, news articles, and research papers. The app's design focuses on simplicity and efficiency, enabling users to quickly obtain concise and coherent summaries with minimal effort.Initial user feedback highlights the system's effectiveness in significantly reducing the time and effort required to extract meaningful information from lengthy texts. Users appreciate the accuracy of extractive summarization and the natural, human-like quality of abstractive summarization. Additionally, the ability to handle multiple document formats is seen as a major advantage, making the tool valuable for professionals across different fields, such as academia, journalism, and business.Overall, the user evaluation indicates that the hybrid text summarization system successfully addresses a critical need for efficient information synthesis, offering a practical solution with high usability and significant potential for future enhancements.

VII. EVALUATION AND REFINEMENT

The project began with the aim of creating a robust text summarization system by combining the strengths of BERT for extractive summarization and GPT for abstractive summarization. The initial version focused on implementing core functionalities, including document processing, summarization, and a user-friendly mobile interface developed using Flutter.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

Through these phases of evolution and refinement, the hybrid text summarization system has matured into a sophisticated tool that effectively addresses the needs of users, offering high-quality summaries across diverse document types and improving continuously based on user feedback and technological advancements.

1. Algorithm Enhancements: BERT Optimization: Fine-tune BERT for better extractive summarization by incorporating more diverse training datasets and refining sentence selection criteria.GPT Improvements:*Update to the latest GPT model versions to leverage advancements in language generation, ensuring more natural and coherent summaries.
2. User Interface Enhancements: Design Improvements: Continuously update the UI based on user feedback, focusing on ease of use, visual appeal, and accessibility. Feature Additions: Introduce features such as highlighting key points in summaries, providing summary analytics, and enabling offline summarization.
3. Performance Optimization: Processing Speed: Implement more efficient algorithms and optimize code to reduce processing time, ensuring real-time summarization capabilities. Resource Management: Enhance the system's ability to handle large documents and multiple simultaneous requests without compromising performance.
4. User Feedback Integration: Customization Options: Allow users to adjust summarization settings, such as the level of detail and specific sections of interest. Interactive Summarization: Develop interactive features that allow users to provide feedback on summaries and refine them based on their preferences.

VIII. CONCLUSION

Text summarization has emerged as a crucial application of Natural Language Processing (NLP), addressing the growing challenges of information overload and the need forefficient content distillation. Leveraging BERT and GPT in this context has yielded promising results. In summary, text summarization using NLP represents a valuable technology for coping with the information-rich digital landscape. It empowers individuals and organizations to manage, understand, and utilize vast amounts of textual data efficiently. As NLP techniques continue to advance, the future of text summarization holds the potential for even more sophisticated and tailored approaches, further enhancing its role in information management and accessibility.

IX. FUTURE SCOPE

Multimodal Summarization: Explore the integration of visual and textual information for summarization tasks, enabling the summarization of multimedia content such as images, videos, and audio recordings.
Real-time Summarization: Develop algorithms and techniques for real-time summarization of streaming data sources, such as social media feeds, news broadcasts, and live events, to provide up-to-date summaries as events unfold.
Domain-specific Summarization: Adapt the summarization models and techniques to specific domains such as healthcare, legal, finance, and scientific research, catering to the unique requirements and terminology of each domain.
Interactive Summarization: Implement interactive interfaces that allow users to provide feedback and guidance during the summarization process, enabling personalized and context-aware summaries tailored to individual preferences.
Multi-document Summarization: Extend the capabilities of the summarization engine to generate summaries from multipledocuments or sources, facilitating comprehensive analysis and synthesis of information from diverse sources.
Abstractive Summarization Improvements: Research and develop advanced techniques for abstractive summarization, focusing on improving coherence, fluency, and factual accuracy of generated summaries.

REFERENCES

[1]. Dilawari, M. U. G. Khan, S. Saleem, Zahoor-Ur-Rehman and F. S. Shaikh, "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space," in IEEE Access, vol. 11, pp. 23557-23564, 2023, doi: 10.1109/ACCESS.2023.3249783.

[2]. Á. Hernández-Castañeda, R. A. García-Hernández and Y. Ledeneva, "Toward the Automatic Generation of an Objective Function for Extractive Text Summarization," in IEEE Access, vol. 11, pp. 51455-51464, 2023, doi: 10.1109/ACCESS.2023.3279101.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

[3]. D. Yadav, R. Katna, A. K. Yadav and J. Morato, "Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey," in IEEE Access, vol. 10, pp. 133981-134003, 2022, doi: 10.1109/ACCESS.2022.3231016

[4]. P. Mahalakshmi and N. S. Fatima, "Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques," in IEEE Access, vol. 10, pp. 18289-18297, 2022, doi: 10.1109/ACCESS.2022.3150414.

[5]. Y. -W. Lai and M. -Y. Chen, "Review of Survey Research in Fuzzy Approach for Text Mining," in IEEE Access, vol. 11, pp. 39635-39649, 2023, doi: 10.1109/ACCESS.2023.3268165.

[6]. Elsaid, A. Mohammed, L. F. Ibrahim and M. M. Sakre, "A Comprehensive Review of Arabic Text Summarization," in IEEE Access, vol. 10, pp. 38012- 38030, 2022, doi: 10.1109/ACCESS.2022.3163292.

[7]. M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," in IEEE Access, vol. 9, pp. 156043-156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[8]. Y. Shin, "Multi-Encoder Transformer for Korean Abstractive Text Summarization," in IEEE Access, vol. 11, pp. 48768-48782, 2023, doi: 10.1109/ACCESS.2023.3277754.

[9]. J. Cheng, F. Zhang and X. Guo, "A Syntax-Augmented and Headline- Aware Neural Text Summarization Method," in IEEE Access, vol. 8, pp. 218360- 218371, 2020, doi: 10.1109/ACCESS.2020.3042886. ZCOER, ZCOER, Department of Computer Engineering 2023-24 .

[10]. M. Y. Saeed, M. Awais, R. Talib and M. Younas, "Unstructured Text Documents Summarization With Multi-Stage Clustering," in IEEE Access, vol. 8, pp. 212838-212854, 2020, doi: 10.1109/ACCESS.2020.3040506.