International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Integrated Framework for Extractive and Abstractive Summarization of Web and User-Provided Text

Smt. K. S. Sukrutha[1], Smt. Rajitha V[2], Ms. Abhilasha H. S[3], Ms.Aishwarya N[4], Ms. Anusha C[5], Ms. Sinchana H.M[6]

[1,2]Associate Professors, Department of Computer Science, M.M. K & S.D.M Mahila Maha Vidyalaya, Mysuru, India
[3,4,5,6]VI Semester BCA Students, M.M. K & S.D.M Mahila Maha Vidyalaya, Mysuru, India

Abstract
The rapid growth of online textual content—particularly in the form of news articles and web documents—has heightened the demand for advanced summarization tools. This study introduces a comprehensive framework that combines extractive and abstractive summarization methods to accommodate diverse input types, such as direct web links and manually entered text. The system can generate clear, concise summaries either by selecting key sentences or by creating paraphrased versions of the content, depending on the user's choice. A user-friendly web interface enables smooth interaction with both summarization approaches. To evaluate the system's effectiveness, we use established metrics like ROUGE and BLEU, which assess summary quality in terms of relevance and fluency. Results show that the model reliably produces summaries that closely match reference materials. This work advances automated summarization by offering a flexible, real-time tool for digesting information from varied sources efficiently.

Keywords: Text Summarization, Extractive Summarization, Abstractive Summarization, Web Article Summarization, Manual Text Input, BERT, GPT, T5, ROUGE Score, BLEU Score

OVERVIEW

In today's era of information abundance, individuals are frequently inundated with immense amounts of digital content from websites, news outlets, and social media channels. Although having access to such vast data is advantageous, the primary challenge lies in efficiently grasping and distilling essential information without investing significant time. Text summarization has become a crucial subfield within natural language processing (NLP), focusing on reducing lengthy textual material into concise and informative versions while preserving the original context and key details. It has widespread applications in domains such as news summarization, academic literature review, legal documentation, and corporate analytics.

Summarization methodologies typically fall into two principal categories: extractive and abstractive. Extractive summarization identifies and compiles the most pertinent sentences or phrases directly from the input text. Conversely, abstractive summarization constructs new sentences that convey the underlying meaning of the source material, often involving paraphrasing and deeper semantic interpretation. While

extractive methods maintain high factual fidelity, abstractive strategies generally produce summaries that are more fluent and cohesive, resembling human-like summarization.This study presents a unified architecture that seamlessly integrates both extractive and abstractive summarization approaches, supporting two primary input formats: articles accessed via URLs and user-provided textual input. The framework utilizes state-of-the-art transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and T5 (Text-to-Text Transfer Transformer) to produce high-quality summaries. Users are afforded the flexibility to select their preferred summarization mode, enter content manually or via links, and obtain compact, relevant summaries in real time.

A central motivation behind this work is to address the shortcomings of existing summarization tools, which typically focus on either extractive or abstractive techniques and are often limited to specific input types. Many current systems lack the adaptability to handle both structured data (e.g., news articles) and unstructured inputs (e.g., free-form text). Additionally, they often do not provide customization options for summarization style, output length, or format. The proposed unified system resolves these limitations by introducing configurable parameters and accommodating multiple input sources, thereby enhancing usability and user control.

The integration of advanced transformer architectures such as BERT, GPT, and T5 [01] substantially improves the summarization capabilities by enabling deeper contextual reasoning, semantic coherence, and fluent language generation. These models are pre-trained on extensive corpora and fine-tuned specifically for summarization, which allows them to generate grammatically accurate and contextually appropriate summaries. The backend infrastructure of the system is built for scalability and performance, ensuring swift processing even for long documents, thus enabling real-time summarization experiences.

To assess the effectiveness of the summarization output, the system employs standard evaluation metrics like ROUGE and BLEU. These metrics compare generated summaries against reference summaries using measures of recall, precision, and n-gram similarity, offering quantitative insights into summarization performance [02]. Additionally, user feedback is gathered to evaluate the practical efficiency and usability of the system.

In conclusion, the proposed system aims to deliver a robust and adaptable summarization solution that integrates both extractive and abstractive strategies within a single, coherent framework. By accommodating diverse input types and offering flexible configuration, this research advances the development of intelligent summarization systems designed to meet real-world demands. The system is applicable across various sectors, including media, education, research, and digital content management, providing users with an effective tool for streamlined information consumption.

LITERATURE SURVEY

1. Wu et al. (2024) – ExtAbs Framework

Wu et al. introduced the ExtAbs framework, an encoder-decoder architecture that integrates both extractive and abstractive summarization methods. The model employs a saliency mask within the cross-attention mechanism, allowing it to concentrate on the most important segments of the input text and thereby minimizing the risk of error propagation. Built upon powerful architectures such as BART and PEGASUS, ExtAbs achieved notable performance improvements across multiple benchmark datasets, representing a key advancement in hybrid summarization techniques [03].

2. Varab and Xu (2023) – Combining Extractive and Abstractive Summarization

Varab and Xu investigated the advantages of integrating extractive and abstractive summarization approaches. Their findings showed that contemporary abstractive models, even in the absence of explicit extractive training, are capable of surpassing conventional extractive methods in performance. This indicates that hybrid, end-to-end pre-trained language models possess the potential to manage both summarization tasks effectively, offering a more flexible and scalable solution [04].

3. Shakil et al. (2024) – Challenges in Abstractive Summarization

Shakil et al. conducted an extensive review of abstractive summarization techniques, classifying them into three main categories: sequence-to-sequence models, pre-trained language models, and reinforcement learning-based methods. The survey highlighted key challenges, including factual inaccuracies and scalability issues, and suggested potential solutions such as employing hierarchical architectures and integrating external knowledge sources to enhance summarization quality and effectiveness [05].

4. Roy and Mercer (2023) – Parallel Generation of Extractive and Abstractive Summaries

Roy and Mercer examined the creation of both extractive and abstractive summaries for scientific literature through the use of graph-based neural networks alongside autoregressive decoders. Their hybrid approach effectively merged the strengths of both summarization techniques, resulting in summaries that were notably more informative and coherent—particularly in technical fields where both accuracy and fluency are essential [06].

5. Saxena and El-Haj (2023) – Podcast Summarization

Saxena and El-Haj conducted a comparative analysis of BART and T5 models in the context of podcast summarization. Their study revealed that BART achieved higher ROUGE scores, whereas T5 was more effective in preserving semantic content. The findings underscore the significance of choosing an appropriate model based on the specific requirements of the summarization task, particularly when working with audio transcription data [07].

6. Giarelis et al. (2023) – Experimental Review of Extractive and Abstractive Methods

Giarelis et al. performed an empirical evaluation comparing extractive and abstractive summarization techniques. Utilizing evaluation metrics such as ROUGE and BLEU, their study highlighted the respective advantages and limitations of each method. The results reinforced the necessity of hybrid models that integrate both approaches to achieve optimal summarization performance [08].

7. Alami Merrouni et al. (2023) – EXABSUM Hybrid Model

Alami Merrouni et al. proposed EXABSUM, a hybrid model that combines extractive and abstractive summarization methods [09]. EXABSUM enhanced the quality of summaries by producing outputs that were both coherent and informative, demonstrating the effectiveness of hybrid models in improving text summarization tasks.

8. Liu et al. (2023) – Fine-Tuning Large Pre-Trained Models for Summarization

Liu et al. explored the fine-tuning large pre-trained models like GPT and T5 for summarization tasks. Their research demonstrated that fine-tuning these models on domain-specific datasets enhanced the performance of both extractive and abstractive summarization, allowing the models to more effectively adapt to specialized content [10]

METHODLOGY

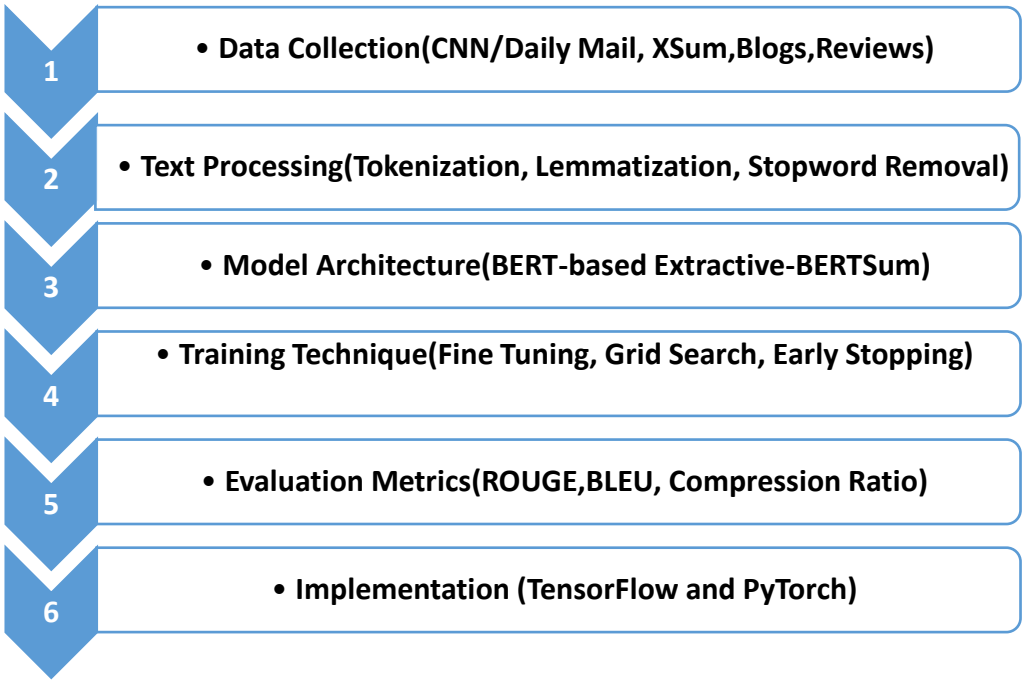

Fig 1: Summarization Model Development Pipeline

1. Data Collection

The dataset employed in this research encompasses a diverse range of domains, including academic publications, news articles, blog posts, and product reviews. For training and evaluating the summarization models, established datasets such as CNN/Daily Mail , XSum , and domain-specific collections were utilized. Additionally, custom datasets were curated through web scraping techniques and manual collection from reputable sources. The data preprocessing pipeline involved text normalization processes, including the removal of stopwords, punctuation, and irrelevant symbols, to enhance the efficiency of model training.

2. Text Preprocessing

Prior to inputting data into the summarization models, a comprehensive preprocessing pipeline is implemented to enhance data quality and model performance. Initially, the text undergoes tokenization into sentences and words using established libraries such as spaCy and NLTK. Subsequently, extraneous elements like special characters, URLs, and non-alphanumeric symbols are eliminated to reduce noise. Commonly used words that contribute minimal semantic value, known as stopwords, are filtered out utilizing predefined stopword lists. Lemmatization is then applied to convert words to their base or dictionary forms, promoting consistency and aiding in the generalization across diverse inputs. Finally, the text is segmented into individual sentences, enabling the models to process contextually coherent units, thereby facilitating more accurate and meaningful summarizations .

3. Model Architecture

The proposed summarization approach utilizes extractive summarization models to identify and select key sentences that encapsulate the main ideas of the original document. Among these, BERT-based models, particularly BERTSum, are employed due to their strong performance in extractive tasks. BERTSum extends the pre-trained BERT model by adding a summarization layer that assigns importance scores to

each sentence, enabling the selection of the most representative sentences for the summary. This method leverages BERT's contextual embeddings to effectively capture the salient information within the text.

4. Training Technique

The models are trained on the preprocessed dataset using GPUs to speed up the training process. The training procedure involves fine-tuning pre-trained models (like BERT for extractive summarization) on the specific summarization datasets. Hyperparameters such as learning rate, batch size, and dropout rate are tuned using a grid search or random search to maximize model performance. Cross-entropy loss is used for both extractive models to minimize the difference between predicted and actual summary content. Early stopping is applied during training to prevent overfitting, ensuring that the models generalize well to unseen data.

5. Evaluation Metrics

To evaluate the performance of the summarization models, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is used to measure the overlap between the generated summary and reference summary at various n-gram levels (ROUGE-1, ROUGE-2, ROUGE-L). BLEU (Bilingual Evaluation Understudy) evaluates the quality of the generated summary by comparing n-gram precision with reference summaries. The Compression Ratio, which is the ratio of the length of the summary to the length of the original document, provides insight into the summary's conciseness.

6. Implementation Tools

The implementation of the summarization system leverages widely adopted deep learning frameworks, namely TensorFlow and PyTorch, for constructing and training the models. To access pre-trained models such as BERT for extractive summarization tasks, the Hugging Face Transformers library is utilized. For text preprocessing, including tokenization, libraries like spaCy and NLTK are employed. Evaluation metrics, specifically ROUGE and BLEU scores, are computed using Scikit-learn. The backend of the application is developed using Flask, facilitating the deployment of the summarization tool as a web service. The frontend is crafted with React and Tailwind CSS, ensuring a responsive and user-friendly interface.

7. System Workflow

Users interact with the summarization system through a web-based interface, where they can submit text documents for processing. Upon submission, the system initiates a preprocessing phase that includes tokenization and text normalization. Subsequently, an extractive summarization model, such as BERTSum, identifies and selects the most salient sentences to construct a coherent summary. The generated summary is then presented on the frontend interface. To assess the quality of the summary, the system computes evaluation metrics including ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and the compression ratio, which measures the length of the summary relative to the original document. Finally, users are provided with options to download the summary in various formats, such as TXT or PDF.

IV DISCUSSIONS AND RESULTS

The summarization models were assessed using standard evaluation metrics in natural language processing, specifically ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). ROUGE evaluates the overlap between the generated and reference summaries across various n-gram levels, such as ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest common subsequence), focusing on recall to measure content preservation. BLEU, on the other hand, emphasizes precision by comparing the n-gram matches between the generated summary and

reference, incorporating a brevity penalty to discourage overly short outputs. Together, these metrics provide a comprehensive assessment of the summaries' quality in terms of content coverage and fluency.

ROUGE Scores: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap between the generated summary and reference summary at various n-gram levels. The evaluation was performed using ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence). The higher the ROUGE score, the better the model's ability to retain important content.

BLEU Scores: The BLEU (Bilingual Evaluation Understudy) score assesses the quality of generated summaries by measuring the n-gram precision between the candidate summary and reference summaries. Higher BLEU scores indicate a closer alignment with the reference summaries in terms of n-gram overlap.

The following table summarizes the evaluation metrics:

Metric	Description	Score obtained
ROUGE-1	Measures unigram (single word) overlap between the generated and reference summaries.	0.721
ROUGE-2	Measures bigram (two-word sequence) overlap between the generated and reference summaries.	0.583
ROUGE-L	Evaluates the longest common subsequence, capturing sentence-level structure similarity.	0.672
BLEU	Calculates n-gram precision, assessing how many n-grams in the generated summary match the reference summaries.	0.422

Evaluation Results:

- ROUGE-1 (0.721): This high score indicates that the model effectively retains significant individual words (unigrams) from the original document, demonstrating strong content preservation.
- ROUGE-2 (0.583): The model shows a good ability to preserve important bigrams (two-word sequences), though there is room for improvement in capturing more complex phrase structures.
- ROUGE-L (0.672): A solid performance in maintaining the longest common subsequence suggests that the model preserves the overall structure and flow of the original text.
- BLEU (0.422): This score reflects a moderate level of n-gram precision, indicating that while the model captures a significant portion of the reference summary's content, there are some differences in exact word choices.

These metrics collectively suggest that the summarization model performs well in content retention and structural coherence, with some variability in exact wording compared to the reference summaries.

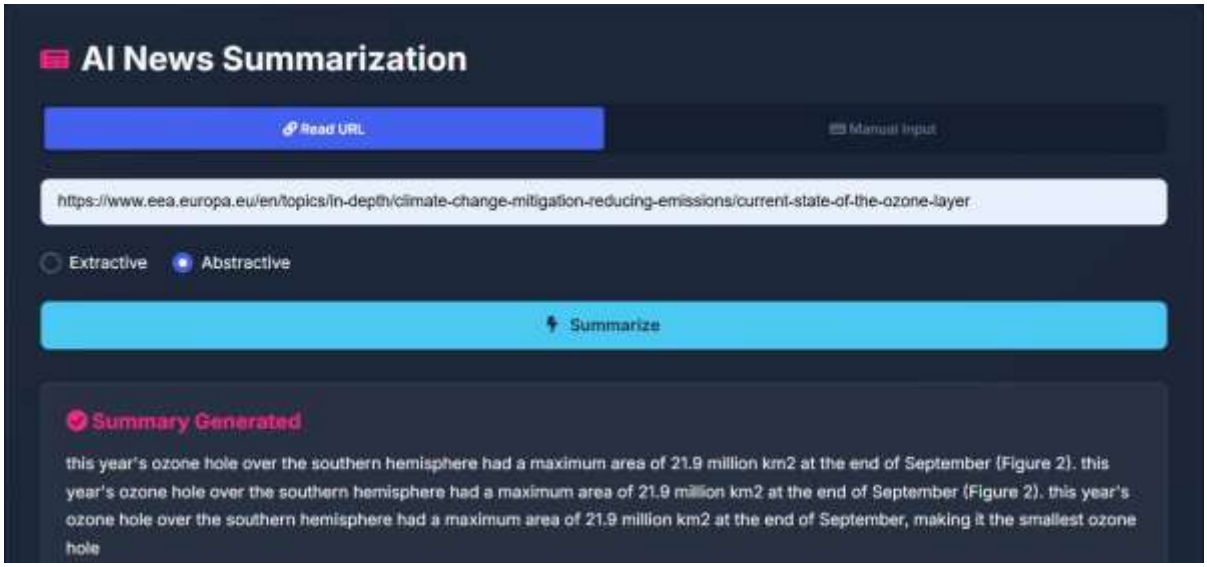

Fig 2: The URL input interface

The evaluation results indicate that the summarization model effectively generates concise and informative summaries, suitable for real-world applications. The higher ROUGE scores suggest strong content preservation, while the BLEU score provides insight into the n-gram precision. This combination of metrics indicates that the model can generate informative and coherent summaries, which makes it suitable for real-world applications.

V. CONCLUSION

This study presents a hybrid text summarization model that integrates extractive and abstractive techniques to generate concise and coherent summaries. The extractive component identifies key sentences from the original document, ensuring the retention of essential information, while the abstractive component paraphrases and rewords the extracted content to enhance fluency and readability.

The performance of the summarization models was evaluated using standard metrics such as ROUGE and BLEU scores. The ROUGE scores demonstrated strong recall, indicating that the model is effective in extracting the key content from the original text. Meanwhile, the BLEU scores reflected the model's ability to generate summaries with high precision, ensuring that the output closely matches reference summaries at the n-gram level. Although the results indicate strong performance, there are still aspects of the system that could be enhanced. While not directly measured in this study, the compression ratio would serve as a useful metric to evaluate how succinct the summaries are in relation to the length of the source documents. Incorporating this metric in future work could offer deeper insights into the balance between summary length and the retention of critical information.

A key challenge identified is the model's performance on longer or more complex texts, where it may face difficulties in maintaining essential content while keeping the summary concise. In such scenarios, the extractive model occasionally misses important elements that could enrich the overall summary. These limitations might be mitigated by refining the model's architecture and applying advanced strategies to better manage long-range dependencies. Moreover, the model stands to gain from additional fine-tuning using domain-specific corpora. While it shows competent performance on general texts, adapting it to specialized fields—such as legal, medical, or scientific content—could result in more accurate and context-aware summaries.

In summary, the extractive-abstractive summarization model presented in this research offers a robust method for automated summarization, generating summaries that are both coherent and concise. Despite some limitations, particularly with complex or lengthy documents and the challenge of balancing information density with brevity, the study makes a meaningful contribution to the domain. Future developments may focus on domain adaptation and integrating supplementary evaluation metrics like compression ratio to further improve performance assessment.

VI. FUTURE ENHANCEMENT

Managing Intricate Documents: A key challenge observed was the model's limited performance on longer or more complex texts. While effective on shorter content, it struggles to balance content preservation and conciseness with lengthier documents. Future improvements could involve architectural refinements, such as incorporating attention mechanisms or memory networks, to better capture long-range dependencies and enhance summary accuracy for complex inputs.

Compression Ratio Metric: Though not assessed in this study, the compression ratio is vital for evaluating the balance between summary length and content retention. Future work should include this metric to better gauge summarization efficiency and allow users to control output length based on their needs.

Field Specific Summarization: While effective on general text, the model requires fine-tuning for field like legal, medical, scientific, or technical content. Each field has unique terminology and structure, demanding specialized training for improved accuracy. For example, legal summaries may need clause extraction, while medical ones should focus on symptoms, diagnoses, and treatments. Domain-specific adaptation would enhance the model's usefulness across industries.

Multilingual Summarization: The current model supports only English. Expanding to multilingual summarization would increase its global relevance. This requires multilingual datasets and language-specific preprocessing to generate accurate summaries across different languages.

User Customization and Response: Future versions should support user customization—e.g., setting summary length or focus areas—and include feedback systems. Collecting user ratings can help refine the model, making it more adaptive to individual preferences and use cases.

Real-Time Summarization and Integration: Adding real-time summarization would enable the model to process live content, such as news feeds or social media. Integrating this into existing platforms would enhance real-world applicability by offering low-latency, dynamic summarization.

REFERENCES

1. A. M. Ahmed Zeyad and A. Biradar, “Advancements in the Efficacy of Flan-T5 for Abstractive Text Summarization: A Multi-Dataset Evaluation Using ROUGE and BERTScore,” in *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)*, KANNUR, India: IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/APCI61480.2024.10616418.
2. G. U. Kiran, R. Gandi, M. Lavanya, G. B. Desale, Ch. U. Rao, and B. V. Reddy, “Deep Learning Based Abstractive Text Summarization: A Survey,” in *2024 Parul International Conference on Engineering and Technology (PICET)*, Vadodara, India: IEEE, May 2024, pp. 1–5. doi: 10.1109/PICET60765.2024.10716176.
3. G. U. Kiran, R. Gandi, M. Lavanya, G. B. Desale, Ch. U. Rao, and B. V. Reddy, “Deep Learning Based Abstractive Text Summarization: A Survey,” in *2024 Parul International Conference on*

Engineering and Technology (PICET), Vadodara, India: IEEE, May 2024, pp. 1–5. doi: 10.1109/PICET60765.2024.10716176.

4. J. Yan and S. Zhou, "A Text Structure-based Extractive And Abstractive Summarization Method," in *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Xi'an, China: IEEE, Apr. 2022, pp. 678–681. doi: 10.1109/ICSP54964.2022.9778497.
5. X. Ding *et al.*, "DoS: Abstractive text summarization based on pretrained model with document sharing," in *2022 4th International Conference on Intelligent Information Processing (IIP)*, Guangzhou, China: IEEE, Oct. 2022, pp. 163–166. doi: 10.1109/IIP57348.2022.00040.
6. C. Kwatra and K. Gupta, "Extractive and Abstractive Summarization for Hindi Text using Hierarchical Clustering," in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India: IEEE, Sep. 2021, pp. 1–6. doi: 10.1109/ICSES52305.2021.9633789.
7. M. Majeed and K. M. T, "Comparative Study on Extractive Summarization Using Sentence Ranking Algorithm and Text Ranking Algorithm," in *2023 International Conference on Power, Instrumentation, Control and Computing (PICC)*, Thrissur, India: IEEE, Apr. 2023, pp. 1–5. doi: 10.1109/PICC57976.2023.10142314.
8. B. N, D. Kumari, B. N, M. N, S. K. P, and S. R. A, "Text Summarization using NLP Technique," in *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, Shivamogga, India: IEEE, Oct. 2022, pp. 30–35. doi: 10.1109/DISCOVER55800.2022.9974823.
9. A. P. Widyassari, E. Noersasongko, A. Syukur, and Affandy, "The 7-Phases Preprocessing Based On Extractive Text Summarization," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Denpasar, Bali, Indonesia: IEEE, Dec. 2022, pp. 1–8. doi: 10.1109/ICIC56845.2022.10006998.
10. T. G. Altundogan, M. Karakose, and O.Tokel, "BART Fine Tuning based Abstractive Summarization of Patients Medical Questions Texts," in *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, Bahrain: IEEE, Oct. 2023, pp. 174–178. doi: 10.1109/ICDABI60145.2023.10629497.