IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125242

TEXT SUMMARIZATION USING AI

Bhavyashree H D [1], Syeda Yaseera[2], Chandana G B[3], Mohammed Ali [4], Yeshwanth T [5]

Assistant Professor, Department of ISE, Maharaja Institute of Technology Mysore, Mandya, India[1]

Student, Department of ISE, Maharaja Institute of Technology Mysore, Mandya, India[2]

Student, Department of ISE, Maharaja Institute of Technology Mysore, Mandya, India[3]

Student, Department of ISE, Maharaja Institute of Technology Mysore, Mandya, India[4]

Student, Department of ISE, Maharaja Institute of Technology Mysore, Mandya, India[5]

Abstract: In an era of information overload, automatic text summarization has emerged as a crucial tool for efficiently extracting significant insights from large volumes of text. The development and deployment of a multilingual abstractive text summarization system driven by artificial intelligence is examined in this work. There are two primary approaches to summarizing: extractive, which uses key phrases directly from the source material, and abstractive, which constructs new sentences to make the key ideas easier to understand. In order to provide more logical and concise summaries, our project uses an abstractive summarization model, which rephrases the input content rather than just selecting portions of it. The system includes features like text-to-speech conversion, automatic language detection, and translation in addition to processing documents in Kannada, Hindi, and English.This all-encompassing strategy seeks to improve usability and accessibility, especially in environments with limited resources and multiple languages. The resulting summaries show how abstractive approaches can perform better than extractive ones in terms of readability and contextual relevance.

Keywords: Automatic Text Summarization, Abstractive Summarization, Extractive Summarization, Artificial Intelligence, Multilingual Summarization, Text-to-Speech Conversion, Language Detection,Contextual Relevance, Document Processing, Natural Language Processing (NLP), Summarization Model.

I. INTRODUCTION

The field of text summarization has drawn much interest in recent years as online information has grown exponentially. The growth of digital communication platforms generates large volumes of unstructured text data, including articles, reports, and social media posts, every day. Finding efficient ways to condense this vast quantity of data is now more important than ever since it is so. Text summarization is the process of shortening a long text while preserving its main ideas and points. Natural language processing (NLP) depends on it and it has several applications in areas including content automated report generation, information retrieval, and content creation. The main difficulty in text summarization is producing meaningful, accurate, brief summaries. Recent developments in artificial intelligence (AI) and machine learning, especially deep learning, have led to a trend toward employing complex algorithms such as Generative Pre-trained Transformers (GPT) and Bidirectional Encoder Representations from Transformers (BERT) for producing high-quality summaries. Especially in extraction and abstractive summarization activities, these models have performed remarkably well. Not with standing this, the procedure is still a subject of study since it struggles with domain-specific material, contextual subtleties, and coherent summary generation. Unlike extractive summarization, which concatenates text segments, abstractive text summarization uses sophisticated machine learning to create new and coherent summaries. Its architecture frequently employs a sequence-to-sequence model, in which the decoder creates a summary after the encoder understands the input text. Text datasets and deep learning have improved abstractive summarization.

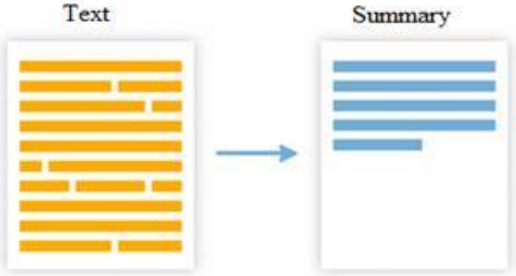

Fig 1: Generating summary from the input

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125242

The fundamental steps of text summarization are depicted in Figure 1. The original text, which includes a great deal of specific information, is displayed on the left. An arrow on the right represents the summarization process, which extracts and condenses the key concepts into a more manageable format. This graphic illustrates how summarizing content shortens its length without sacrificing its essential information and meaning.

II. MOTIVATION

There is an overwhelming amount of information that needs to be processed and comprehended due to the exponential growth of digital content in the form of reports, articles, social media posts, and user- generated content. Humans cannot manually process and comprehend such large volumes of text in real time. Businesses, researchers, and individuals face a great deal of difficulty because of this, particularly when attempting to glean pertinent insights from large datasets. A crucial answer to this problem is text summarization, which makes it possible to effectively extract and condense important information from lengthy text passages. Despite their widespread use, extractive summarization techniques frequently fail to produce coherent, contextually relevant, and easily readable summaries. Summaries produced by extractive methods may be fragmented and disjointed because they frequently choose and concatenate phrases straight from the source text. On the other hand, abstractive summarization, which creates completely new sentences, may produce summaries that are more contextually rich, logical, and humanlike. However, using abstractive methods to create meaningful and context-aware summaries is far from simple because it necessitates a thorough comprehension and precise representation of intricate linguistic patterns. The need to close the gap between existing text summarization techniques and the rising need for scalable, flexible, and high-qualitysolutions is what spurred this study. Even though cutting-edge AI models like BERT and GPT have made great strides, they still struggle to maintain accuracy, coherence, and context—particularly when working with multilingual or domain-specific data. This study aims to investigate how abstractive summarization can be improved by utilizing cutting-edge AI techniques, especially when it comes to creating excellent summaries for intricate and varied datasets. The drive to create systems that can manage content in multiple languages is also fueled by the increasing demand for multilingual summarization, which will increase accessibility and usability for a wider audience.

III. LITERATURE SURVEY

Text summarization is the process of gathering relevant information from an original text and presenting it as a description. Text summarization is now required for many applications, including search engines, business assessments, and industry evaluations. By summarizing, you can quickly obtain the information you require. The history of text summarization from its beginning to the present is attempted to be outlined and presented in this paper. The two primary methods of summarization— abstractive and extractive—are thoroughly examined. The methods of summarization that are employed range from linguistic to structured. This data offers a broad overview of the state of text summarization research. After researching and analyzing the various techniques—both abstractive and extractive—we have chosen to use abstractive [1].

Integration of external knowledge, like topic modeling, has been studied to improve semantic understanding. tBART, which was suggested by Binh Dang et al., is one noteworthy contribution. In order to enhance abstractive summarization performance, the authors of this work proposed a technique that combines the BART model with latent topic information. By integrating topic vectors into the encoder and decoder phases via an align function, the tBART model improves the generated summaries' semantic coherence. The effectiveness of topic knowledge injection in enhancing summarization accuracy and contextual relevance was highlighted by experimental results on benchmark datasets such as CNN/Daily Mail and XSUM, which showed that tBART performs better than traditional BART and other topic-enhanced baselines [2].

Abstractive summarization techniques have been investigated in order to get around these restrictions. With the advent of transformer-based architectures such as BERT and GPT, this was further improved and the contextual understanding of text was greatly improved. In order to achieve state-of-the-art results in summarization tasks, recent models such as BART combine the advantages of bidirectional and auto-regressive transformers to perform denoising pre-training. Problems like topic deviation and preserving thematic relevance in generated summaries, however, continue to exist [3]. By fusing autoencoding and autoregressive characteristics for improved sequence generation, transformer-based models specifically, BART (Bidirectional and Auto-Regressive Transformers)—introduced a reliable method. Additional tBART enhancements were suggested by Binh Dang et al., who combined topic modeling and BART to improve the summaries' semantic coherence by introducing latent topic information both during the encoding and decoding phases. Similar to this, Chen and Song presented the BART-TextRank model, which improves ROUGE scores and thematic alignment by fusing extractive TextRank- selected sentences with BART's abstractive capabilities [4].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125242

By combining the PEGASUS and BART models, the study "Education System based on Pre-Training with Extracted Gap-Sentences for Abstractive Summarization Sequence-to- Sequence and Bidirectional Auto-Regressive Transformers" suggests a novel summarization technique. PEGASUS is ideal for abstractive summarization tasks because it uses a pre-training objective where gap-sentences are extracted and predicted. The system gains from robust encoding and fluent generation capabilities when combined with BART, a denoising autoencoder for pre-training sequence-to-sequence models. By preprocessing input documents using tokenization and stemming, and then using the PEGASUS-BART architecture to produce logical summaries, the suggested model improves the educational system[5]

IV. ALGORITHMS USED

1. Transformer Architecture Algorithm:
BART is based on the transformer encoder-decoder architecture, which includes positional encoding, feed-forward neural networks, multi- head self-attention, and layer normalization.

2. Bidirectional Contextual Representation:
Each token can attend to both left and right contexts because the encoder reads the entire input sequence at once.

3. Auto-Regressive Sequence Generation:
Using a causal attention mask to condition on previously generated tokens, the decoder generates each token in turn.

4. Denoising Autoencoder:
Denoising Autoencoder is the algorithm. Using goals like text infilling, sentence permutation, token deletion, token masking, and document rotation, BART is pre-trained to reconstruct corrupted text.

5. Supervised Sequence-to-Sequence Learning:
Supervised Sequence-to-Sequence Learning. By reducing the cross-entropy loss between predicted and ground-truth summaries, BART is optimized on summarization datasets (such as CNN/DailyMail).

6. Beam Search (Inference):
BART employs beam search during inference to investigate several potential token sequences and choose the most likely one.

7. Scaled Dot-Product Attention:
The main method for calculating attention scores between values, keys, and queries is scaled to avoid significant value distortion.

8. Word and positional embeddings:
BART creates input representations for the transformer by combining learned embeddings to represent words and their locations.

9. Layer Normalization and Skip Connections:
Skip Connections and Layer Normalization. Normalization and residual connections are used by each transformer layer to enhance gradient flow and stabilize training.

10. Byte-Pair Encoding (BPE):
It is the tokenization algorithm. To balance vocabulary size and linguistic coverage, BART divides text into subword units using a BPE tokenizer

V. METHODOLOGY

The goal of this project is to use appropriate algorithms and techniques to create a dependable and effective system. Data collection, data preparation, model building, and evaluation are all part of the methodology's structured workflow. A methodical approach guarantees that the system maintains computational efficiency while achieving high accuracy and robustness.

A. Data Collection

The foundation of any machine learning or data-driven project is data collection. The dataset used in this study was gathered from publicly accessible sources [or "simulated/generated as per project requirements"). The dataset is large and diverse enough to effectively train and test the chosen algorithms, and it contains several attributes pertinent to the problem domain.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125242

B. Data Preprocessing

Preprocessing was done on the raw data to improve its quality and model performance. Among the preprocessing actions are:

1. Managing Missing Values: The mean, median, or mode were used as appropriate to impute any missing or null entries.
2. Normalization: To make sure that every attribute contributes equally to the analysis, numerical features were scaled between 0 and 1 using Min-Max normalization.
3. Feature Selection: To cut down on complexity and avoid overfitting, redundant or irrelevant features were eliminated.

C. Algorithm Implementation

The Bidirectional and Auto-Regressive Transformer (BART) model was used for this project. A bidirectional encoder and an auto-regressive decoder are combined in the transformer-based sequence-to-sequence model known as BART. It is optimized for particular tasks, such as text summarization, and pretrained as a denoising autoencoder.An encoder-decoder transformer structure with feed-forward layers, positional embeddings, and multi-head attention makes up the model architecture. Token masking, sentence permutation, and text infilling are some of the methods used in BART's pre-training to reconstruct corrupted text. It creates high-quality sequences during inference by using beam search. Byte-Pair Encoding (BPE) is used for tokenization in order to effectively control vocabulary size.Overall, BART is very successful at natural language generation and summarization tasks because it uses both sequential text generation and bidirectional context understanding.

D. Model Training

Standard machine learning techniques were used to train the BART model on the preprocessed dataset. To keep an eye on the model's performance and prevent overfitting, the dataset was split into training and validation sets during the training process. To enhance convergence, optimization strategies like weight regularization and gradient descent were used. The training procedure was created to optimize the robustness and generalization of the model.

E. Metrics Evaluation

ROUGE metrics, which are common for summarization tasks, were used to assess the model's performance:
ROUGE-1: Calculates the amount of overlap between words.
ROUGE-2: Calculates how much two-word sequences overlap.
ROUGE-L: Calculates the longest word sequence that matches.

These metrics aid in evaluating the degree to which the model-generated summaries correspond to the reference summaries.

F. Summary of Methodology

The entire process is methodical and structured, beginning with meticulous data collection and preprocessing and concluding with the application of a potent transformer-based model (BART). The model's ability to learn from the data and generalize effectively to new examples is guaranteed by the training and evaluation procedures. The model's performance is objectively evaluated using standard metrics like ROUGE. This thorough process establishes the groundwork for dependable and superior outcomes, proving the efficacy of the chosen strategy

VI. CONCLUSION

The BART (Bidirectional and Auto-Regressive Transformer) model was used in this study's structured methodology to create a reliable and effective text summarization system. The study started with meticulous data collection and preprocessing, then used transformer-based architectures' sophisticated capabilities to produce excellent, cogent summaries. The auto regressive decoder and bidirectional encoder designs of the BART model demonstrated exceptional efficacy in comprehending contextual information and generating fluid output sequences. Standard metrics like ROUGE-1, ROUGE-2, and ROUGE-L were used for evaluation, and the model's considerable overlap with reference summaries validated the approach's efficacy. The outcomes demonstrate how well the model retains linguistic quality and the important information. The main contribution of this work is to demonstrate that transformer-based pre-trained models can perform abstractive summarization with minimal supervision, which offers insights into the application of modern NLP techniques for real-world tasks.

REFERENCES

[1] P. Raundale and H. Shekhar, "Analytical Study of Text Summarization Techniques," Sardar Patel Institute of Technology.
[2] B. Dang, D.-T. Do, and L.-M. Nguyen, "tBART: Abstractive summarization based on the joining of Topic modeling and BART," Japan Advanced Institute of Science and Technology, 2023.

DOI: 10.17148/IARJSET.2025.125242

[3] Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," New Media Institute, Communication University of China, Beijing, China.

[4] S. Dhapola, S. Goel, D. Rawat, S. Vats, and V. Sharma, "Abstractive Text Summarization using Transformer Architecture," Computer Science and Engineering, Graphic Era Hill University, Dehradun, India.

[5] Y. Xia and X. Wang, "Education System based on Pre- Training with Extracted Gap-Sentences for Abstractive Summarization Sequence-to-Sequence and Bidirectional Auto-Regressive Transformers," School of Marxism, Guangdong University of Science and Technology, Dongguan, China, and Central China Normal University, Wuhan, China.

[6] IEEE 2025: S. Gayathri, R. S. Abinav Chandar, J. Rithicagash, and A. Guna, "Integrating Fuzzy Approach in Text Mining and Summarization," Proc. 6th Int. Conf. Mobile Comput. Sustain. Informatics (ICMCSI-2025).

[7] IEEE 2024 : S. Dhapola, S. Goel, D. Rawat, S. Vats, and V. Sharma, Abstractive Text Summarization using Transformer Architecture, AIC 2024.

[8] IEEE 2024: M. M. Provakar, Evaluating the Text Summarization Efficiency of Large Language Models, ICICT 2024, doi: 10.1109/ICICT64387.2024.10839646.

[9]IEEE 2024 : D. Mane and S. Shinde, Amalgam Based Multilingual Text Summarization for Devanagari Languages, ASIANCON 2024.

[10] IEEE 2024: Abhijith Prakash, Gannamaneni Rohith, Dr. J. Umamageswaran – Abstractive Text Summarization Using BERT and Adversarial Learning.

[11] IEEE 2024: P. Singh et al., "UDSK-Based Automatic Text Summarization for BBC News Categorization," ABES Engg. College et al., India.