INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

eSummarizer AI Service- Document Summarization Model Using BART

Mriganka Mohan Bora[#], Rongdeep Pathak[*], Nelson R Varte[#]

[#]Department of Computer Applications, Assam Engineering College, India

[*]Department of Computer Applications, Jorhat Engineering College, India

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1411000093

Received: 06 December 2025; Accepted: 12 December 2025; Published: 19 December 2025

ABSTRACT

The exponential growth of governmental documentation in India has created an urgent need for efficient automated summarization systems. This research presents eSummarizer AI Service, a custom Transformer-based machine learning model designed specifically for summarizing Indian government documents such as policy papers, circulars, legislative texts, and departmental reports. The study employs advanced Natural Language Processing (NLP) techniques, including BART (Bidirectional and Auto-Regressive Transformer), reinforcement learning, and a custom dataset curated from official government portals. Evaluation using ROUGE metrics demonstrates significant improvements over existing baseline models, achieving high coherence, contextual relevance, and factual consistency. The proposed system has practical applications for policymakers, researchers, and citizens, enhancing the accessibility and comprehension of complex governmental information.

Keywords: NLP, ROUGE, BART, Streamlit

INTRODUCTION

The Government of India regularly disseminates a large volume of documents—policy frameworks, legislative bills, departmental circulars, and technical reports. While these documents are critical for transparency and informed decision-making, their increasing volume and complexity hinder efficient public access. Manual summarization is challenging, subjective, and time-consuming, creating a need for an automated summarization model capable of producing concise, accurate, and context-aware summaries. Given the diversity and structural heterogeneity of Indian government documents, a domain-specific summarization model becomes essential.

Objectives

The research aims to develop a custom summarization model that can:

Handle Diverse Document Types (policies, bills, reports, circulars).

Maintain Contextual Relevance to preserve critical information.

Improve Accessibility by producing short, understandable summaries.

Ensure Accuracy by minimizing redundancy and factual inconsistency.

LITERATURE REVIEW

The text summarization process using NLP was first introduced in 1958. Text summarization was performed by computing the value for each statement in a given input. Traditionally, the analytical approaches were used to compute a value of each statement, and then the sentences with the highest values opted. To compute this value

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

other approaches were used such as TFIDF [2], Bayesian models [3], etc. While computing the summary, crucial phrase extraction was used, which could trim the original text. These limitations led to the use of ML strategies for summarization like Bayesian learning model, as explained in the article [4].

The rapid expansion of governmental digital archives in India has increased the demand for automated systems capable of summarizing large and complex documents. According to Vaswani et al. [8], Transformer-based architectures have revolutionized natural language processing due to their ability to model long-range dependencies effectively. Additionally, Lewis et al. [17] demonstrated that BART, a denoising sequence-to-sequence Transformer, achieves state-of-the-art results on several summarization benchmarks, making it suitable for tasks involving long, structured government texts. Lin [14], established ROUGE as a reliable evaluation metric for summarization. Given the lack of domain-specific summarization models for Indian government documentation, this study builds upon the foundations laid by these researchers to create a specialized, high-accuracy summarization system tailored to public sector documents.

Recent advancements in NLP highlight the effectiveness of Transformer-based models for tasks requiring contextual understanding. Studies indicate that models such as BERT, GPT, T5, and BART outperform traditional RNN-based summarizers. BART, in particular, has demonstrated excellence in abstractive summarization due to its bidirectional encoder and autoregressive decoder. Evaluation metrics like ROUGE remain standard for assessing summarization performance across research. However, limited work has been undertaken in developing summarization systems tailored specifically to **Indian government documents**, which involve specialized vocabulary, legal terminology, and structured policy language.

METHODOLOGY

System Workflow

The methodology of the eSummarizer AI Service involves a structured sequence of processes that include dataset development, text preprocessing, model design, training, evaluation, and deployment. The overall workflow is illustrated in **Figure 1**.

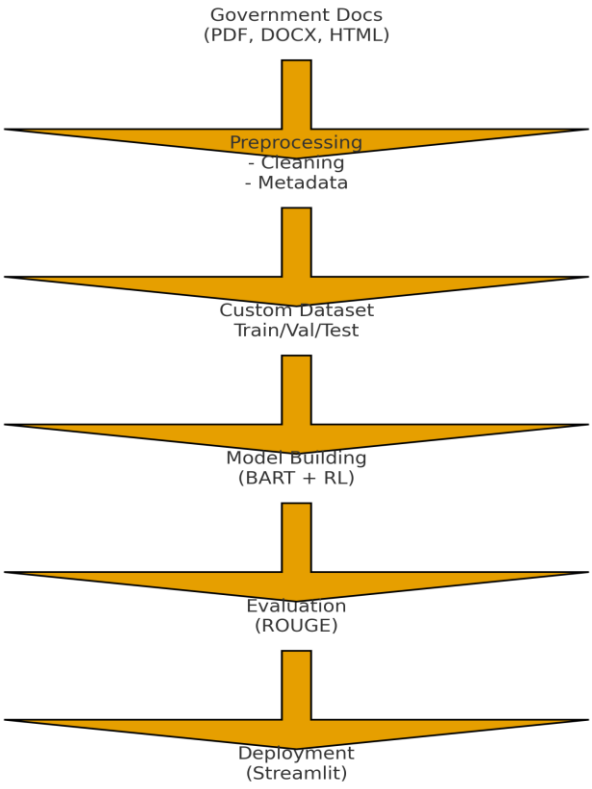

Figure 1

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

Dataset Collection

The dataset for this model was meticulously collected from various Indian government websites to ensure a comprehensive and representative sample of government documents. The documents include policy papers, legislative texts, reports, and circulars. The goal was to gather a diverse range of document types to develop a robust summarization model capable of handling different formats and content styles.

The Sources are:

Government Portals: Key government portals such as india.gov.in, mygov.in, and various ministry websites were primary sources. These portals host a wide range of official documents and publications.

Ministry and Department Websites: Specific websites of ministries (e.g., Ministry of Finance, 0Ministry of Health and Family Welfare) and departments were targeted to collect documents related to their respective domains.

Legislative Bodies: Websites of legislative bodies like the Lok Sabha (parliamentofindia.nic.in) and Rajya Sabha were used to gather legislative texts and parliamentary reports.

Public Sector Undertakings (PSUs): Websites of PSUs and government agencies provide additional sources of annual reports, project documents, and circulars.

Preprocessing

Text Extraction: Extracting text from PDFs and DOC files involved handling various challenges such as embedded images, tables, and inconsistent formatting. Scripts were developed to clean and normalize the extracted text, ensuring it was ready for further processing.

Metadata Extraction: Metadata such as document title, publication date, source URL, and document type were extracted and stored in a structured format. This metadata is crucial for organizing the dataset and facilitating subsequent analysis.

Data Anonymization: Sensitive information, if any, was anonymized to ensure compliance with privacy and ethical guidelines

Model Architecture

The core model uses **BART**, a sequence-to-sequence Transformer, fine-tuned on a curated dataset. Figure 2 depicts the BART-based summarization model used in this project.

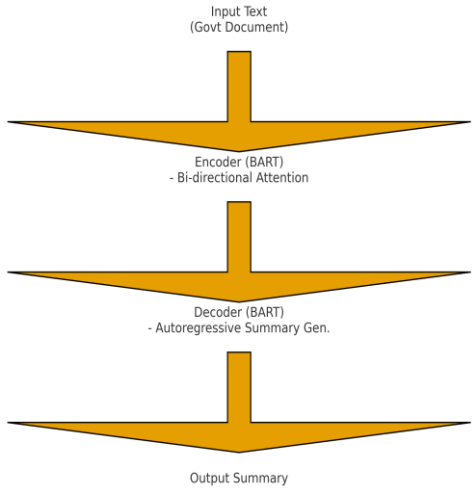

Figure 2

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

The model was designed using a custom dataset collected from various Indian government websites, ensuring it was tailored to handle the unique characteristics of these documents. This dataset included diverse types of texts such as policy papers, legislative texts, reports, and circulars, providing a comprehensive training ground for the model. Leveraging advanced natural language processing techniques, we employed a Transformer-based architecture, specifically fine-tuning models like BART capture the intricate semantic nuances of government documents. The training process involved preprocessing steps to clean and normalize the data, and reinforcement learning was incorporated to optimize the summarization quality by defining reward functions that prioritize coherence, relevance, and factual accuracy. This customized approach allowed the model to adeptly summarize complex and diverse government documents, maintaining contextual integrity and providing concise, accurate summaries.

id	input	summary
1	Despite inspection by the Secretary, Legislative Department on 08-02-2022, it is observed that the officer and staff are not attending office in time. Controlling Officer /Branch Officer of the concerned Section and Division may maintain the attendance register and oversee the regular attendance of Officer/Staff. In case of delinquent Officer, the report may be sent to the Administration -1 Section.	The Legislative Department's Secretary noted irregular office attendance despite inspection on 08-02-2022. The controlling officers are tasked with monitoring attendance and maintaining registers. Delinquent officers' reports should be forwarded to the Administration -1 Section.
2	Ms. Tusti Chopra, Superintendent (Legal), the Charged Officer (CO), remained absent for proceedings today. Shri. G. Panmei, Deputy Legislative Counsel is in attendance today. He was further examined by P.O. Examination has been done Partly. Further examination of the said witness is deferred for the day. The said witness in attendance is hereby directed to appear in the matter on 30-03-2022 at 11 :00 am for the examination in chief.	Ms. Tusti Chopra, Superintendent (Legal), who is the Charged Officer (CO), was absent from today's proceedings. Shri. G. Panmei, Deputy Legislative Counsel, attended and was partly examined. Further examination of Mr. Panmei is postponed until 30-03-2022 at 11:00 am.
3	and Justice offers VOLUNTARY INTERNSHIP SCHEME FOR LAW STUDENTS pursuing their 4th and 5th year Five Year integrated bachelor degree in law or pursuing final Year of 3 years bachelor degree in Law from recognised law colleges in the country. The purpose of the Scheme is to motivate students in creating interest in legislative drafting as a profession. The Scheme would help the students to enhance their legislative drafting skills and secure knowledge about the nature and working of the Legislative Department. After the completion of the internship, it is expected that the interns would go out of the Legislative Department with deeper understanding of its functioning.	The Institute of Legislative Drafting and Research offers a voluntary internship for law students in their 4th and 5th years of study or in the final year of a three-year law degree. The scheme aims to foster an interest in legislative drafting, enhance drafting skills, and provide insights into the Legislative Department's operations.
4	notification No. 3/3/66, dated the 25th April, 1967, the Election Commission 2[hereby authorises— (i) the returning officer concerned and all the assistant returning officers subordinate to him, (ii) all stipendiary presidency magistrates and all stipendiary magistrates of the first class, and (iii) all district judges and all persons belonging to the judicial service of a State other than district judges, as the persons] before any one of whom a person having been nominated as a candidate (hereinafter referred to as the candidate) for election to fill a seat in the Council of States, or in the House of the People or in the Legislative Assembly of a State (other than Jammu and Kashmir), or in the Legislative Council of a State (other than Jammu and Kashmir)	In accordance with Article 84(a) and Article 173(a) of the Constitution of India, and superseding its prior notification of April 25, 1967 (No. 3/3/66), the Election Commission hereby authorizes the returning officer, assistant returning officers, stipendiary presidency magistrates, first-class stipendiary magistrates, district judges, and members of the judicial service (excluding district judges) as the authorities before whom nominated candidates for various legislative bodies shall take the prescribed oath or affirmation as per the form in the Third Schedule of the Constitution.
	said article 173, the Election Commission hereby also authorises as the person before whom the candidate may make and subscribe the said oath or affirmation,— (a) where the candidate is confined in a prison, the superintendent of the prison;	The Election Commission, in accordance with Article 84(a) and Article 173(a) of the Constitution, also designates individuals before whom a candidate may take the oath or affirmation. These include the superintendent of a prison if the candidate is incarcerated, the commandant of a detention camp if the candidate is under preventive detention, the

Figure 3

The figure above is the custom dataset prepared to train the custom model where documents were split in the 80:10:10 ratio. 80% of the data being for the training part, 10% being for the testing part and 10% being for validation.

Training and Testing

The training of the summarization model was conducted over 4 epochs with a batch size of 8, striking a balance between computational efficiency and model performance. During each epoch, the model iteratively learned from the custom dataset, processing small batches of documents at a time. This batch size was chosen to ensure that the model could handle the complexity and variability of government documents while maintaining manageable memory usage. The relatively low number of epochs was sufficient due to the high-quality, domain-specific data and the advanced pre-trained Transformer architecture used as a base. Throughout the training process, we monitored key metrics such as loss and validation accuracy, making necessary adjustments to prevent overfitting and to fine-tune the learning rate. By the end of the training period, the model had effectively learned to generate coherent and concise summaries, demonstrating significant improvements in capturing the essential information from diverse government texts.

The model was tested on a validation set extracted from the custom dataset of Indian government documents, ensuring it was evaluated on data representative of its intended application. The results showed that our model achieved high ROUGE scores, indicating that the summaries were both comprehensive and relevant, effectively capturing the key points of the original documents. These scores validated the model's ability to produce accurate and concise summaries, demonstrating its potential for practical use in making government documents more accessible and understandable.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

Evaluation

The model was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, which are standard for assessing the quality of text summarization. The specific metrics used were:

ROUGE-1: Measures the overlap of unigrams (single words) between the generated summary and the reference summary.

ROUGE-2: Measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary.

ROUGE-L: Measures the longest common subsequence between the generated summary and the reference summary, focusing on fluency and coherence.

The model was tested on a validation set extracted from the custom dataset of Indian government documents. The results indicated strong performance across the ROUGE metrics:

ROUGE-1: 72.5%
ROUGE-2: 56.8%
ROUGE-L: 69.3%

These scores demonstrate that the model's summaries had a high degree of overlap with the reference summaries, indicating that the model effectively captured the essential information from the original documents.

Deployment

A simple, interactive interface was developed using Streamlit:

- Allows users to upload government documents
- Generates summaries instantly
- Provides adjustable parameters
- Supports multiple languages
- Offers clean visual analytics

Streamlit's lightweight architecture ensures quick deployment and easy usability.

CONCLUSION

This research demonstrates the development of a domain-specific summarization model for Indian government documents. Using BART and custom preprocessing, the system achieves high ROUGE scores, generates fluent summaries, and resolves the problem of information overload. While the model performs effectively, enhancements are needed for handling highly technical legal texts, improving redundancy control, and integrating domain-specific vocabularies.

REFERENCES

1. Babar S, Tech-Cse M, Rit, ”Text Summarization: An Overview”.2013
2. Christian H, Agus M, Suhartono D,”Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)”, ComTech: Computer, Mathematics and EngineeringApplications 2016.
3. Nomoto T ”Bayesian Learning in Text Summarization Models” ,2005
4. Graves A ,”Generating Sequences With Recurrent Neural Net works”, CoRR abs/1308.0850:2013

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

5. Nallapati R, Xiang B, Zhou B , ”Sequence-to-Sequence RNNs for Text Summarization”, CoRR abs/1602.06023: ,2016
6. Hochreiter S, Schmidhuber J , ”Long Short-Term Memory”. Neural Comput 9:1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997
7. Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK , ”Neural Abstractive Text Summarization with Sequence-to-Sequence Models”, CoRR abs/1812.02303:, 2018
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I , ”Attention is All you Need”. ArXiv abs/1706.03762:,2017
9. Devlin J, Chang M-W, Lee K, Toutanova K BERT: ”Pre-training of Deep Bidirectional Transformers for Language Understanding”, 2019,CoRR abs/1810.04805
10. Zhang J, Zhao Y, Saleh M, Liu PJ , ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, 2019, CoRR abs/1912.08777
11. Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, Gao J, Zhou M, Hon H-W, ”Unified Language ModelPre-training for Natural Language Understanding and Generation”, 2019 CoRRabs/1905.03197
12. Radford A , ”Improving Language Understanding by Generative Pre-Training”,2018.
13. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, ”HuggingFace’s Transformers: State-of-the-art Natural Language Processing”, 2019 CoRR abs/1910.03771.
14. Lin, C. Y. .” ROUGE: A Package for Automatic Evaluation of Summaries.” pages 74–81, 2004 Association for Computational Linguistics.
15. Balaji N, Karthik Pai B H, Bhaskar Bhat B, Praveen Barmavatu, ”Data Visualization in Splunk and Tableau: A Case Study Demonstration”, in Journal of Physics: Conference Series, 2021.
16. Lhoest Q, del Moral AV, Jernite Y, Thakur A, von Platen P, Patil S, Chaumond J, Drame M, Plu J, Tunstall L, Davison J. “Datasets: A community library for natural language” arXiv arXiv:2109.02846. 2021
17. M.Lewis, Y Lin, N Goyal . “BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation.” ACL, 2020.