# Transformer Based Approaches for Marathi Abstractive Summarization: A Comparative Analysis

Sonali Waje[1], C. M. Raut[2]
[1]*P.G. Student, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai 400 708 (India)*
[2]*Professor, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai 400 708 (India)*

***Abstract*—Automatic Text Summarization (ATS) is a significant task in the field of Natural Language Processing (NLP), aimed at generating concise and meaningful summaries from large textual content. In recent years, ATS has grabbed attention of many researchers because of large volumes of information which is readily available in multiple languages on a digital platform. Automatically generating precise summaries from large text has potential application in generation of news headlines, summary of research articles, financial research, chatbots etc. While English text summarization has achieved substantial progress due to abundant datasets and pretrained models, Marathi and other Indian regional languages remain underrepresented in research. Marathi language poses special difficulties for Natural Language Processing due to rich morphology and intricate syntactic patterns. This study focuses on comparison of different transformer-based model.**

***Index Terms*—Abstractive Text summarization (ATS), Transformer Models, IndicBART, mBART, mT5, Marathi language.**

## I. INTRODUCTION

The rapid growth of digital information has created a strong need for efficient techniques to process and understand large volumes of textual data. Automated text summarization has therefore become an important research area in the field of Natural Language Processing (NLP). The main objective of text summarization is to generate a concise and meaningful summary from lengthy documents while preserving the important information. Such systems help users save time and improve information retrieval efficiency in domains such as news analysis, education, healthcare, and digital media.

Text summarization techniques are generally categorized into extractive, abstractive, and hybrid approaches. Extractive summarization identifies and selects the most important sentences directly from the original document without modifying their structure. In contrast, abstractive summarization generates summaries in a human-like manner by understanding the context of the text and producing new sentences with improved coherence and readability. Hybrid methods combine the advantages of both extractive and abstractive techniques to produce informative and fluent summaries.

In Marathi language processing, earlier research focused mainly on extractive summarization methods. Traditional approaches such as TextRank and Latent Semantic Analysis (LSA) were widely used to identify important sentences from Marathi documents. Although these methods produced acceptable results, they often lacked semantic understanding and generated summaries that were less natural and less coherent. Since extractive methods only select existing sentences from the source text, they are unable to effectively paraphrase or restructure information.

Marathi is one of the major Indo Aryan languages spoken by millions of people in India. Despite its widespread usage, Marathi remains relatively underrepresented in NLP research compared to resource-rich languages such as English. Marathi possesses complex linguistic characteristics including rich morphology, flexible word order, and diverse

grammatical structures, which make automatic summarization a challenging task. Conventional rule-based and statistical techniques are often insufficient to capture the semantic relationships and contextual meaning required for high quality abstractive summarization.

Recent advancements in deep learning and transformer-based architectures have significantly improved the performance of various NLP applications. Pretrained models such as BERT, mBART, T5, and IndicBART have demonstrated remarkable capabilities in text generation and language understanding through transfer learning. Among these models, IndicBART is specifically designed for Indian languages and supports multilingual sequence-to-sequence text generation. Its architecture enables better contextual understanding and generation for Indic languages, making it suitable for Marathi abstractive text summarization.

This research focuses on improving Marathi abstractive text summarization using transformer-based models, particularly mBART and IndicBART. The proposed work utilizes Marathi datasets obtained from the L3Cube-MahaSum and XL-Sum corpora for training, validation, and testing purposes. The study fine-tunes the pretrained IndicBART model to generate concise, coherent, and informative summaries in Marathi. The performance of the generated summaries is evaluated using standard metrics such as ROUGE-1, ROUGE-2, ROUGE-L, and BLEU score. The research aims to analyze the effectiveness of transformer-based approaches in enhancing summary quality in terms of fluency, informativeness, and semantic accuracy.

The major contribution of this work is the comparative improvement and evaluation of mBART and IndicBART models for Marathi abstractive text summarization. By exploring advanced transformer architectures for low-resource Indian languages, this study contributes toward the development of more effective NLP solutions for Marathi and other Indic languages.

## II. LITERATURE REVIEW

Text summarization is an important task in Natural Language Processing (NLP) that aims to generate a concise and meaningful summary from a larger document while preserving its key information. With the increasing volume of digital textual information such as news articles, research papers, and government reports, automatic text summarization has become essential for efficient information management and knowledge extraction.

Text summarization techniques are broadly categorized into extractive summarization and abstractive summarization. Extractive summarization selects important sentences directly from the original document based on statistical or linguistic features. Early methods relied on techniques such as TF–IDF weighting, sentence position scoring, and similarity measures to identify relevant sentences. These methods often used clustering techniques such as K-means to remove redundancy and improve summary diversity.

Later research introduced machine learning approaches for summarization. Algorithms such as Support Vector Machines, Naïve Bayes, and neural networks were applied to determine sentence importance using various linguistic and statistical features. However, these methods required extensive feature engineering and struggled to capture contextual relationships between sentences.

With the advancement of deep learning, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) were introduced for abstractive summarization tasks [2]. These models use encoder–decoder architectures with attention mechanisms to generate summaries that better capture the semantic meaning of the original text. Studies conducted on languages such as Urdu and Bengali demonstrated that attention-based neural architectures improve summarization quality compared to traditional methods.

In recent years, Transformer-based architectures have significantly improved the performance of text summarization systems. Transformer models rely on self-attention mechanisms that allow the model to understand relationships between words across long text sequences. Popular transformer models include BERT, BART, T5, PEGASUS, mBART, and mT5, which have achieved state-of-the-art performance in various NLP tasks including summarization.

For low-resource languages such as Marathi, specialized transformer models have been developed to address linguistic challenges such as complex morphology and limited training datasets. Research indicates that Indic-specific transformer models such as IndicBART perform better than general multilingual models because they are trained on large corpora of Indian languages.

Another important concept in summarization research is the evaluation of generated summaries. Traditional evaluation metrics such as ROUGE and BLEU measure lexical overlap between the generated summary and reference summary. However, these metrics sometimes fail to capture semantic similarity. Therefore, newer evaluation methods such as BERTScore and semantic similarity metrics have been proposed to better evaluate summary quality and contextual relevance.

The availability of large datasets is also critical for training summarization models. Recent research introduced datasets such as MahaSUM and XL-Sum, which provide large collections of Marathi news articles paired with summaries for training and evaluation of abstractive summarization systems.

Thus, the evolution of summarization techniques from statistical methods to deep learning and transformer-based architectures has significantly improved summarization performance, particularly in multilingual and low resource language scenarios.

## III. PROPOSED SYSTEM

Fig.1 depicts the proposed methodology used for developing an ATS system for Marathi language using different transformer models. It consists of steps like dataset selection, data preprocessing, model selection & fine tuning, summary generation and evaluation.

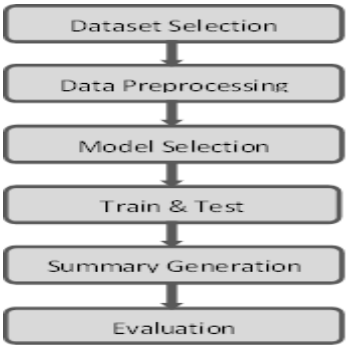


Figure 1: Proposed Methodology

(a) Dataset Selection

BBC Research developed large-scale multilingual summarization dataset XL-Sum that represents information across 44 languages with Marathi among them. The system provides human-written abstract text summaries that stay connected to the original context. This study utilized the Marathi subset from the XL-Sum dataset for benchmark performance evaluation.

The BBC Marathi news article collection provides 5,855 records for Marathi language which contains the fields like ID (A unique identifier for each news article), URL (The link to the original BBC Marathi news article), Title (Headline of the article), Article (Full text of the news article), Summary (Manually generated abstractive summary). The dataset divides its contents into 3 sets as Training Set: 4,684 records, Validation Set: 732 records and Test Set: 439 records.

(b) Data Preprocessing

The method of converting raw data into clean and structured arrangement before feeding it into a machine learning model. It involves the techniques to improve model accuracy and efficiency like Cleaning, Stop word Removal, Stemming & Lemmatization, Handling Encoding Issues, Text Normalization.

(C) Model Selection

The proposed system uses mT5, mBART, and IndicBART models for Marathi abstractive text summarization. These transformer-based models are selected because of their strong capability in multilingual text generation and contextual understanding. Among them, IndicBART is specifically designed for Indian languages, making it highly suitable for Marathi summarization tasks. The models are fine-tuned using XL-Sum dataset, and their performance is evaluated using ROUGE and BLEU scores.

(d) Evaluation

Evaluation: Most extensively used method for evaluating text summarization is ROUGE score. It computes the overlap of produced summary with the reference summary using different n-gram and sequence-based approaches. The performance of summarization is assessed using ROUGE-1, ROUGE-2, ROUGE-L and BLUE score.

ROUGE-1: This method determines the extent of shared unigram content between the produced summary and the benchmark summary. It offers a simple way to evaluate the level of overlap between specific terms in the benchmark summary and the produced summary. A higher score indicates better content coverage.
Equation (1) shows ROUGE-1 calculation formula.
The formula for ROUGE-1 Recall is:

$$\text{ROUGE-1} = \frac{\sum \text{Number of overlapping unigrams}}{\sum \text{Total unigrams in reference summary}} \quad (1)$$

Where:
Overlapping unigrams = common words between generated and reference summaries
Reference summary = actual human-written summary

ROUGE-2: It evaluates how much the produced and benchmark summaries coincide in terms of bigrams, which are two-word sequences. By examining word sequences, it captures important elements of the summary generating process & offers insights regarding the coherence and fluency of the summary. By considering word pairs, it mainly captures contextual coherence.
Equation (2) shows ROUGE-2 calculation formula.
The formula for ROUGE-2 Recall is:

$$\text{ROUGE-2} = \frac{\sum \text{Number of overlapping bigrams}}{\sum \text{Total bigrams in reference summary}} \quad (2)$$

Where:
Overlapping bigrams = matching two-word sequences between generated and reference summaries
Reference summary = original human-written summary

ROUGE-L: The produced summary & benchmark summary's longest common subsequence (LCS) match is measured. It assesses the summary's capability to maintain the structure and sequence of content as presented in the reference summary. Higher ROUGE-L indicates that produced summary follows the structure of the benchmark summary more closely.

Equation (3) shows ROUGE-L calculation formula.
The formula for ROUGE-L Recall is:

$$\text{ROUGE-L} = \frac{\text{LCS(Generated Summary, Reference Summary)}}{\text{Length of Reference Summary}} \quad (3)$$

Where:
LCS = Longest Common Subsequence between generated and reference summaries
Reference Summary Length = total number of words in the reference summary

4) BLEU Score: BLEU (Bilingual Evaluation Understudy) score measures the similarity between the generated summary and the reference summary by calculating the precision of matching n-grams.
The BLEU score formula is:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:
BP = Brevity Penalty
$p_n$= Precision of n-grams
$w_n$= Weight assigned to each n-gram
N = Maximum n-gram length

BLEU score evaluates the fluency and accuracy of the generated summary by comparing it with the reference summary. Higher BLEU values indicate better quality summaries.

Table 1: Evaluation Metrics

| Evaluation Metrics | IndicBART | mT5 | mBART |
|---|---|---|---|
| ROUGE-1 | 17.7 | 15.3 | 11.8 |
| ROUGE-2 | 8 | 9.4 | 9.4 |
| ROUGE-L | 19.8 | 14.8 | 13.1 |
| BLEU Score | 12.6 | 10.2 | 8.7 |

## IV. CONCLUSION

- The study successfully demonstrated the use of transformer-based deep learning models for Marathi abstractive text summarization.
- Comparative analysis showed that fine-tuned multilingual transformer models generate more coherent, meaningful, and contextually accurate summaries.
- The results confirmed the effectiveness of transformer architectures in handling low-resource regional languages like Marathi.
- The research highlighted the importance of transfer learning, attention mechanisms, and multilingual NLP techniques in improving summarization quality.

- Challenges such as limited datasets, linguistic complexity, and evaluation difficulties were also identified.
- The proposed approach can be applied in areas such as news summarization, education, e-governance, and document analysis.
- Future work may focus on larger datasets, advanced language models, multilingual systems, and improved evaluation techniques.
- Overall, the work provides a strong foundation for future research in Marathi Natural Language Processing and low-resource language technologies.

REFERENCES

[1] Yang, Y., Tan, Y., Min, J., and Huang, Z., "Automatic Text Summarization for Government News Reports Based on Multiple Features," The Journal of Supercomputing, vol. 80, no. 3, pp. 3212-3228, 2024.

[2] Awais, M., and Nawab, R. M. A., "Abstractive Text Summarization for the Urdu Language: Data and Methods," IEEE Access, vol. 12, pp. 61198-61210, 2024.

[3] S. S. Rifat, M. A. Rahman, and M. R. Islam, "Abstractive Text Summarization for Bangla Language Using NLP and Machine Learning Approaches," in Proceedings of the International Conference on Electrical, Computer and Communication Engg (ECCE), Feb. 13-15, 2025.

[4] K. Joshi, A. Kunchukuttan, P. Bhattacharyya, and S. Varma, "L3Cube-MahaSum: A Comprehensive Dataset and BART Models for Abstractive Text Summarization in Marathi," arXiv preprint arXiv:2410.09184, Oct. 2024.

[5] S. Patil, A. Kulkarni, and P. Deshpande, "Enhancing Marathi Abstractive Text Summarization with IndicBART," in Proceedings of the International Conference on Inventive Computation Technologies (ICICT), IEEE, 2025.

[6] S. Kulkarni, P. Patil, and R. Joshi, "Multilingual and Cross-Lingual Text Summarization of Marathi and English Using Transformer-Based Models and Their Systematic Evaluation," International Journal of Computer Applications, vol. 186, no. 26, pp. 15-22, Jun. 2024.

[7] S. Deshmukh and R. Kulkarni, "Experimental Evaluation and Enhancement of Automatic Unsupervised Extractive Text Summarization of Marathi Text Using Machine Learning Algorithm," Journal of Machine and Computing, vol. 2, no. 1, pp. 45-52, 2022. [8] M. Hosseini, A. Shakery, and H. Zamani, "Optimizing Question-Answering Framework Through Integration of Text Summarization Model and Third-Generation Generative Pre-Trained Transformer," in Proceedings of the 14th Int. Conference on Computer and Knowledge Engg (ICCKE), 2024.

[8] A. Kumar, S. Sharma, and P. Gupta, "Hybrid Extractive-Abstractive Model for Indic News Summarization," in Proceedings of the International Conference on Natural Language Processing (ICON), 2024.

[9] A. Gupta, R. Sharma, and S. Verma, "Transformer-Based Low-Resource Summarization for Indian Languages," in Proceedings of the IEEE International Conference on Big Data, 2024.

[10] A. Kunchukuttan, P. Bhattacharyya, and S. Varma, "Marathi XL-Sum Evaluation Study for Abstractive Text Summarization," Computer Science Series, Springer, 2024.

[11] J. Zhang, X. Liu, and Y. Wang, "Fine-Tuning PEGASUS for Indic Language Abstractive Summarization," in Proceedings of the IEEE International Conference on Computing and Communication, 2024.

[12] T. Liu, H. Chen, and Y. Zhang, "Comparative Analysis of TextRank and BERT-Based Extractive Models for News Summarization," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2024.

[13] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, et al., "Cross-Lingual Abstractive Text Summarization Using M2M100 Transformer," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

[14] Y. Zhang, T. Liu, and H. Chen, "Semantic-Aware Evaluation Metrics for Abstractive Text Summarization," Information Processing & Management, Elsevier, vol. 61, no. 3, 2024.

[15] A. Kumar, S. Sharma, and R. Patel, "Deep Learning Techniques for Regional Language Natural Language Processing: A Comprehensive

Review," Neural Networks, Elsevier, vol. 168, pp. 45-62, 2024.

[16] A. Kumar, S. Sharma, and R. Patel, "Deep Learning Techniques for Regional Language Natural Language Processing: A Comprehensive Review," Neural Networks, Elsevier, vol. 168, pp. 45-62, 2024.

[17] P. Deshmukh, A. Kulkarni, and S. Patil, "Comparative Analysis of Transformer Architectures for Marathi Natural Language Processing," IEEE Access, vol. 13, pp. 11245-11260, 2025.

[18] R. Sharma, A. Singh, and K. Gupta, "Benchmark Evaluation of Transformer Models for Indic Natural Language Processing Tasks," IEEE Access, vol. 13, pp. 15420-15435, 2025