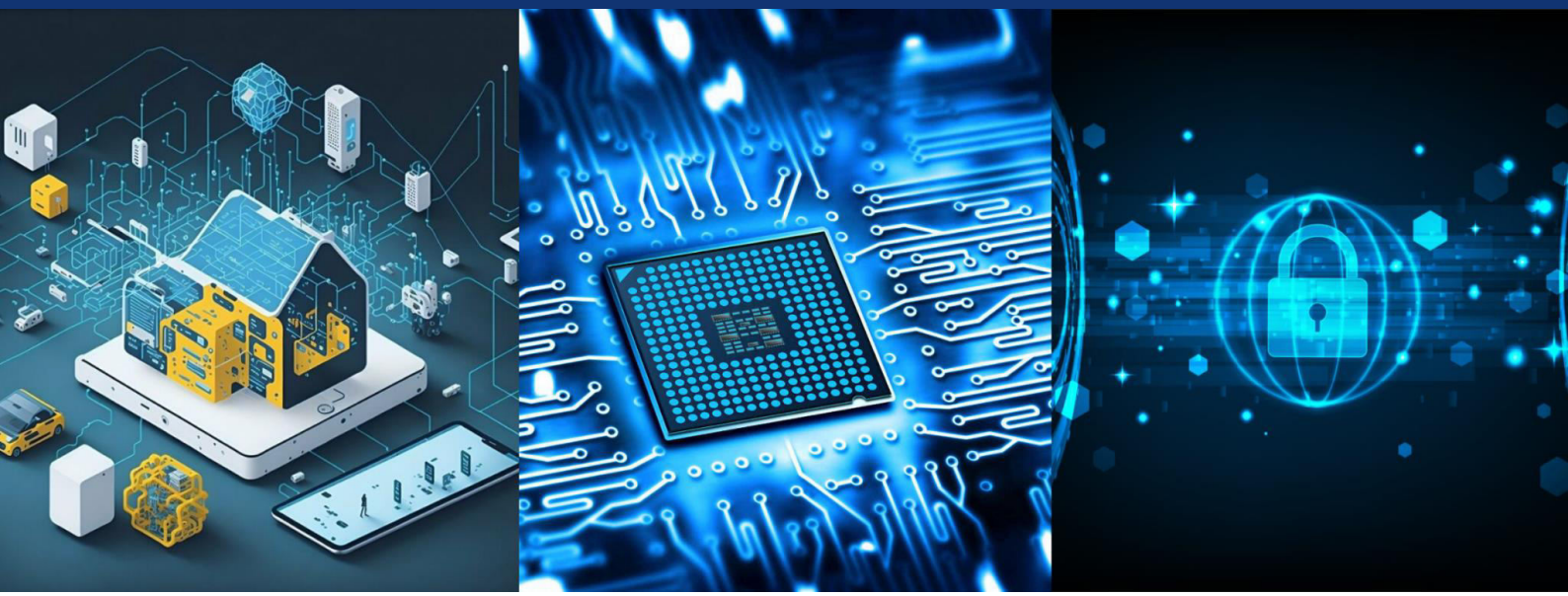
IJIRCCE

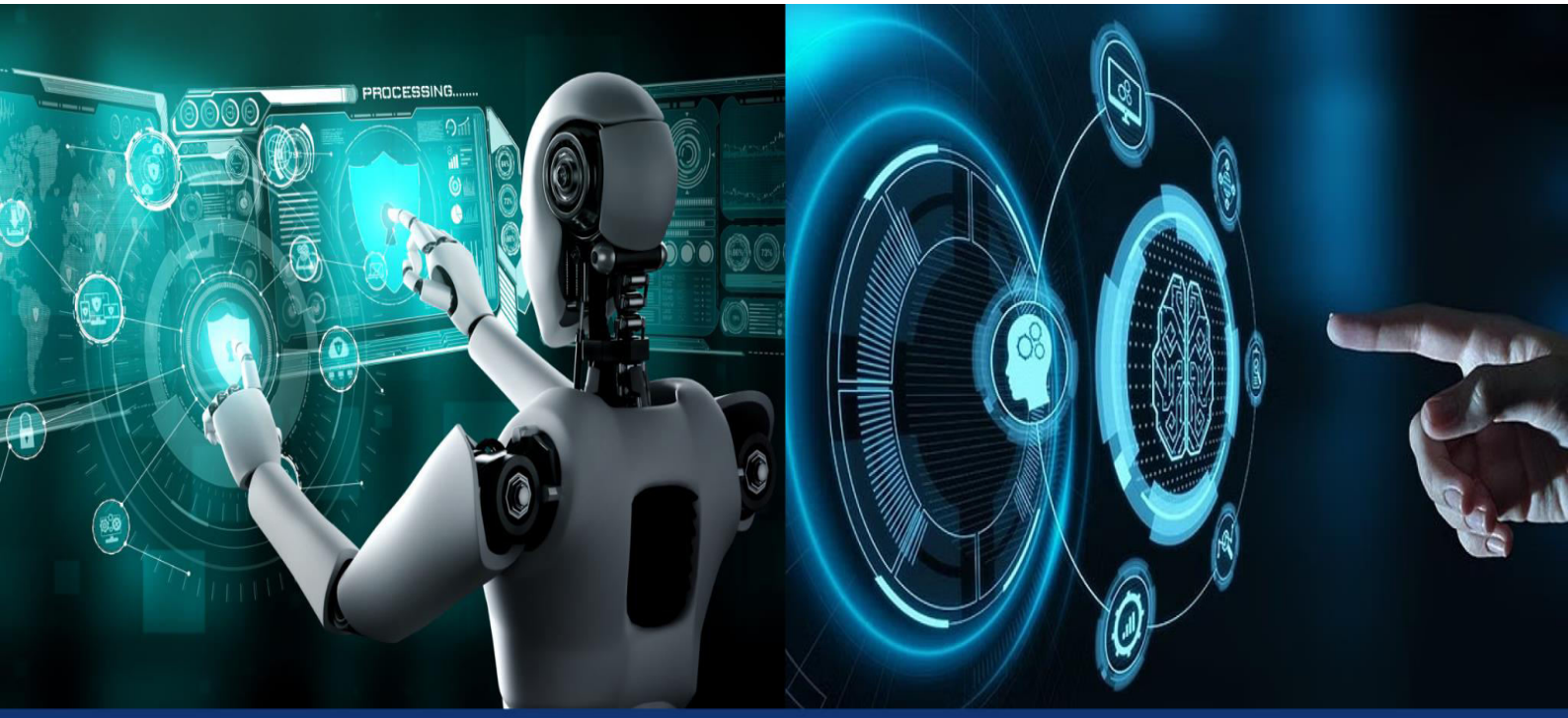
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407038

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Abstractive Kannada Text Summarization Using Multilingual T5 (mT5)

Prateek Malipatil, Nagamani D R

Department of CSE, Bangalore Institute of Technology, Bengaluru, India

Department of CSE, Bangalore Institute of Technology, Bengaluru, India

ABSTRACT: With the increasing availability of digital information, there is now a greater demand for text summarization systems that are capable of providing meaningful summaries of the information. The task of producing abstractive summaries, especially in the case of low resource languages like Kannada, is difficult because of insufficient data resources. In this paper, a new framework for abstractive text summarization in Kannada using mT5 transformer-based architecture has been proposed. Domain specific data was collected from Kannada news articles and preprocessed to make the dataset consistent. The model is trained to provide compact summaries containing two to three sentences. The performance of the model based on ROUGE scores suggests that the proposed framework provides reliable results even in a low-resource environment.

KEYWORDS: Kannada Language Processing, Abstractive Summarization, Transformer-Based Models, mT5, Low Resource NLP

I. INTRODUCTION

The rising amount of textual data available in the online medium necessitates the development of summarization methods for efficient consumption of information. In general, summarization is an approach that involves reducing long pieces of texts into smaller summaries by preserving their main ideas. Two common approaches are used for accomplishing this task; one is extractive summarization, and the other is abstractive summarization.

Recent advancements in transformer-based models such as T5 make it easier to perform abstractive summarization in high-resource languages. But performing abstractive summarization for a language like Kannada is not a trivial task owing to lack of annotation.

To deal with this problem, we propose a model utilizing the properties of multilingual transformers such as mT5 and thus facilitate knowledge transfer between languages. In this paper, we focus on developing an abstractive summarizer for shortening news articles written in Kannada language.

II. LITERATURE SURVEY

2.1 WCD-YOLO: Waste Classification Detection Model

Text summarization is one of the major tasks of Natural Language Processing (NLP), where the objective is to compress a vast amount of text content into meaningful summary representations. Previous methods of summarization used the approach of extractive summarization in order to achieve the objective. Recently, with the use of deep learning, abstractive summarization has become more prevalent.

The use of Transformer-based systems has led to better results in abstractive summarization. T5 is one of the architectures where researchers have come up with a single architecture for many NLP tasks including summarization. It has been proven that the T5 system yields coherent and contextually accurate summaries, which perform better in ROUGE evaluation than conventional sequence-to-sequence systems. Additionally, multilingual versions of T5 known as mT5, are capable of doing multiple language tasks. They include low-resource language tasks [1].

Many researchers have applied their knowledge of summarization in languages used in India. For instance, the application of an ensemble of mT5 and IndicBART in summarizing Indian languages has led to more accurate results,

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

measured in terms of ROUGE scores, compared to other transformer models. Similar results have been reported on summarizing Telugu text using various multi-lingual transformer models such as mT5, mBART, and IndicBART. It has been reported that these models outperform the conventional methods of extractive text summarization [2][3].

Similarly, the use of Indian languages is not the only challenge in the domain of summarization as issues are also seen with morphologically rich languages like Arabic. A recent paper suggested the use of a transformer-based model, namely mT5 and AraBART, coupled with new large-scale datasets. It was also shown that transformer based models can provide competitive results in such complicated languages [4].

However, there have been no many advancements in the summarization of Kannada text. Existing literature mostly deals with high resource languages or limited Indian languages such as Hindi and Telugu. Lack of large-scale datasets and complicated structure of Kannada makes it difficult for researchers to develop efficient approaches. Inspired by the success of the multilingual transformer architecture, this work will leverage mT5 for Kannada abstractive text summarization [5].

III. METHODOLOGY

Fig 1 illustrates propose system flow diagram which describes the mT5 Transformer model's Kannada abstractive text summarization process. A text summarization process follows four crucial stages including data gathering and preparation and model optimization and summary evaluation.

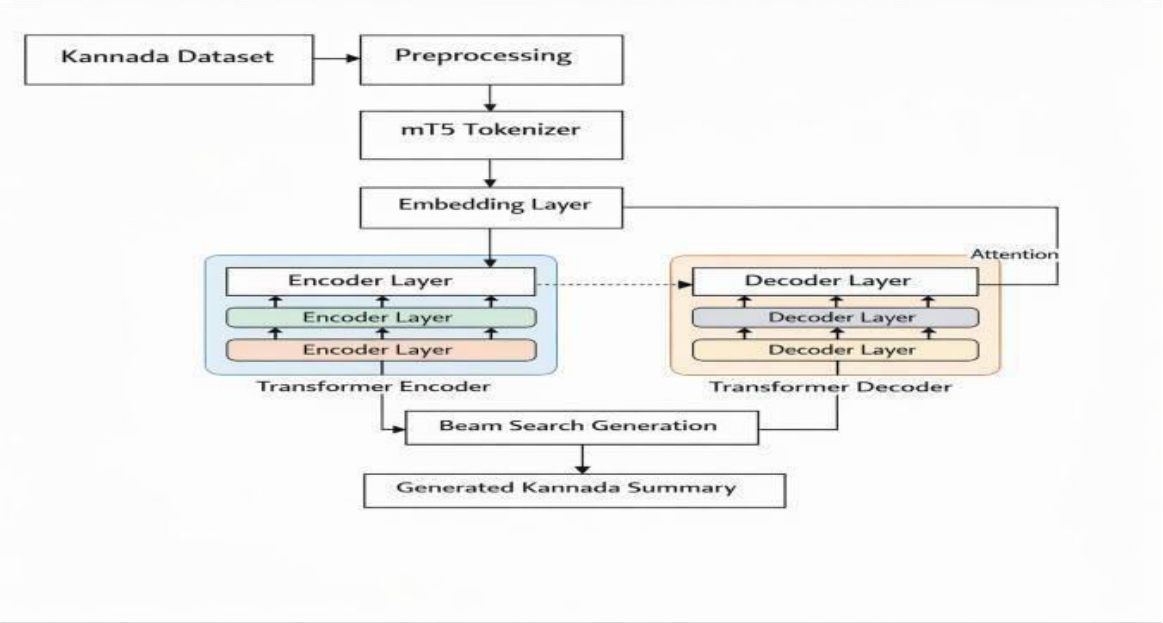


Fig. 1. Proposed System Flow Diagram

3.1 Description of Dataset

Given that there is no existing public large dataset for abstractive summarization of Kannada, a dataset has been prepared specifically for this research. Articles in Kannada have been extracted from various web sources, namely Prajavani, Vijay Karnataka, and Public TV, which offer a wide range of topics like politics, education, sports, and general news [6].

3.2 Dataset Statistics

The final dataset contains around 6,500 pairs of articles and summaries and can be used to train transformer models in resource-poor settings.
Average Length of Articles: 500 words Average Length of Short Summaries: 50 words
Average Length of Large Summaries: 120 words

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.3 Data Preprocessing

dataset to be prepared for efficient training of the summary generation model. Taking into consideration the complexities of the language, such as its rich morphology and sentence structures, the need for meticulous preprocessing was imperative to ensure that the input data had consistent and high quality. To start with, the dataset was initially loaded from structured files (CSV file format). It was ensured that the input data had a complete text-summary pairing. Following the initial loading, the text data was further preprocessed by removing HTML tags, punctuation marks, and other extraneous characters. The dataset underwent normalization in terms of Unicode encoding in order to ensure that there was no inconsistency with regard to the Kannada text. In the next step, the data was tokenized using the tokenizer for the mT5 model. Tokenization is an important step as it enables the model to understand and process rare words, among others, by converting the text data into subword units.

Afterward, the inputs were encoded using:

- Input IDs – numerical encoding of tokens
- Attention masks – distinction between actual tokens and padding.

3.4 Tokenization and Embedding

After preprocessing, the text goes through mT5 tokenization, where the Kannada text is converted to subword tokens. Tokenization works well for morphological variants and rare words typically seen in Kannada. These tokenized sequences are fed into an embedding layer to create numeric representations. This layer captures the semantics and syntax of the text.

3.5 Transformer Encoder

The tokenized sequences undergo transformation using the transformer encoder layer, where there are several encoder layers in the network. Each layer performs the process of self-attention and feedforward neural networks. Self-attention makes the model focus on relevant areas of the text sequence, thus allowing the model to detect dependencies and sentence structure.

3.6 Transformer Decoder

The encoded text sequences are then passed through the transformer decoder layer. It decodes the sequences and produces the output summarization sequence. The transformer decoder has several layers.

- The self-attention layer understands the generated tokens in the output sequence
- The cross-attention layer focuses on relevant encoded sequences

This process helps the model create an accurate summary output sequence.

3.7 Summary Generation using Beam Search

Beam search is utilized for generating summaries compared to greedy decoding, as beam search explores various combinations of potential output sentences and produces an optimal summary sentence sequence.

3.8 Evaluation

The performance of summarization was evaluated using ROUGE measures by checking the level of similarity with respect to the reference sentences [3].

1) ROUGE Recall: The ratio of overlapping units (e.g., unigrams, bigrams, LCS) to the total number of units in the reference summary as shown in Equation

$$Recall = \frac{\text{Overlapping units}}{\text{Total units in reference}} \quad (1)$$

2) ROUGE Precision: The ratio of overlapping units to the total number of units in the generated (candidate) summary as shown in Equation (2).

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

$$Precision = \frac{\text{Overlapping units}}{\text{Total units in candidate}} \quad (2)$$

3) ROUGE F1-score: The harmonic mean of ROUGE precision and recall, providing a balanced measure as shown in Equation (3).

$$\text{F1-Score} = 2\,x\,\frac{Precision\ x\ Recall}{Recall+Precision} \quad (3)$$

IV. RESULTS

The proposed architecture was then trained and tested on the created Kannada text summarization dataset. In this case, the dataset was split into training and testing datasets to make sure that the test results were unbiased. Implementation of this system was done using Python programming language and the HuggingFace transformers library. Training was done in a GPU-enabling environment.

Table I presents outlines the ROUGE scores achieved by the mT-5 transformer model.

TABLE I. ROUGE SCORES FOR MT5 TRANSFORMER MODEL

	Rouge-1	Rouge-2	Rouge-L
Recall	0.35	0.28	0.33
Precision	0.32	0.19	0.29
F- measure	0.36	0.21	0.31

Performance Evaluation of Model Using ROUGE Metrics. The proposed mT5 model was evaluated using ROUGE metrics. Table I shows recall, precision, and F1-scores of ROUGE-1, ROUGE-2, and ROUGE-L.

From the table, the model was able to achieve an F1-score of 0.36 for ROUGE-1, 0.21 for ROUGE-2, and 0.31 for ROUGE-L. These results show that the model is able to capture essential information from the Kannada texts. However, its performance is relatively poor because the model is trained in a low-resource environment.

TABLE II. COMPARISON BETWEEN REFERENCE SUMMARY AND M-T5 MODEL SUMMARY

SI No	Reference Summary	Generated Summary (mT5 model summary)
1	**Kannada:** ಮೊಬೈಲ್ ಅಂಗಡಿ ಮಾಲೀಕನ ಮೀಲೆ ಹಲೆೆ ನಡೆಸಿದ ಪ್ರಕರಣದಲೆ ಇನ್ನಿ ಬ್ಬರ ಬ್ಂಧನ. **EnglishTransliteration:**	**Kannada:** ಮೊಬೈಲ್ ಅಂಗಡಿ ಮಾಲೀಕನ ಮೀಲೆ ಹಲೆೆ ಪ್ರಕರಣ: ಪೊಲೀಸರು ಮತ್ತೆ ಇಬ್ಬ ರನ್ನಿ ಬ್ಂಧಿಸಿದ್ಾ ರೆ. **EnglishTransliteration:**

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407038

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	Mobail aṅgaḍi mālīkana mēle halle prakaraṇadalli mattibbara bandhana **Translated in English:** Two more people arrested in the case of assault on a mobile shop owner.	Mobail aṅgaḍi mālīkana mēle halle prakaraṇa: Polīsaru matte ibbarannu bandhisiddāre. **Translated in English:** Assault case on a mobile shop owner: Police arrested two more persons.
2	**Kannada:** ನಾಡಪ್ರಭು ಕಂಪೇಗೌಡರಿಗೆ ಅವಮಾನ ಮಾಡಿದ ಆರೋಪ: ನಟ ಚೇತನ್ ವಿರುದಧ ದೂರು ದ್ದಖಲು. **EnglishTransliteration:** Nāḍaprabhu kempēgauḍarige avamāna ārōpa: Naṭacētan virud'dha dūru **Translated in English:** Allegation of insulting Nadaprabhu Kempegowda: Complaint filed against actor Chetan.	**Kannada:** ನಾಡಪ್ರಭು ಕಂಪೇಗೌಡರ ಬ್ಗೆ ಅವಹೀಳನಕಾರಿ ಪೊೀಸ್ಟ ಹಂಚಿಕ: ನಟ ಚೇತನ್ ವಿರುದಧ ದೂರು ದ್ದಖಲಾಗಿದೆ. **EnglishTransliteration:** Nāḍaprabhu kempēgauḍa bagge avahēḷanakāri pōsṭ śēr: Naṭa cētan virud'dha dūru **Translated in English:** Sharing a derogatory post about Nadaprabhu Kempegowda: Complaint filed against actor Chetan.

Table II presents a sample comparative analysis between the reference summary and the summary generated by the mT5 model, highlighting key differences in content coverage, coherence, and informativeness.

4.1 Quality Results

The summaries were judgued on the basis of coherence, fluency, and relevance. The model able to generate coherent and fluent summaries, that were to the input article closely matching the reference headlines. Here is an example:

Input Article:

ಮೈಸೂರು ಜಿಲೆ ಯ ಕ.ಆರ್.ನಗರದ ಅಕೀೆೀಶ್ವ ರ ದೀವಸ್ಥಾ ನದ ಬ್ಳಿ ಕಾವೀರಿ ನೀರಿನಲೆ ಈಜಲು ತೆರಳಿದದ ಆರು ಮಂದಿ ಪ್ರವಾಸಿಗರು ನೀರಿನಲೆ ಮುಳುಗಿ ಮೃತಪಟ್ಟ ಬಗೆ ಮಾಹಿತಿ ಲಭ್ಯವಾಗಿದೆ. ಉರೂಸ್ಟ ಕಾಯೀಕರ ಮದ ಹಿನ್ನು ಲೆಯಲೆ ಬಂಗಳೂರು ಮತ್ತಿ ಉಳ್ಟಿಯಂದ ಆಗಮಿಸಿದದ ಎಂಟು ಜನರ ತಂಡವಂದು ನೀರಿನಲೆ ಸ್ನಾನ ಮಾಡಲು ಇಳಿದದ ಗ ಈ ದುರ್ಘಟನೆ ಸಂಭವಿಸಿದ.

Generated Summary:

ಮೈಸೂರು: ಕಾವೀರಿ ನೀರಿನಲೆ ಈಜಲು ತೆರಳಿದದ ಆರು ಪ್ರವಾಸಿಗರು ಸಾವು.

Fig 2 shows the GUI implemented to get summary for given entered text. To get summary for any text GUI has been designed in which user has to enter the text and get the summary.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407038

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

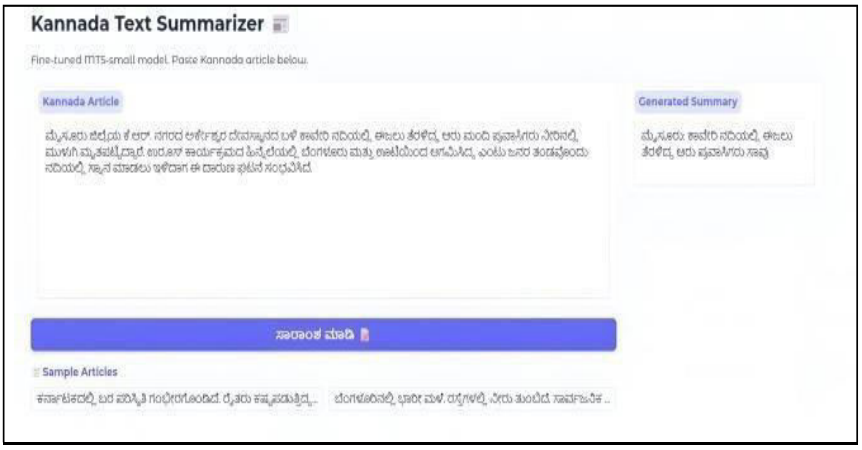


Fig. 2 Graphical User Interface of proposed system

V. CONCLUSION AND FUTURE SCOPE

This paper proposed an abstractive text summarization method for the Kannada language using the mT5 model. As there is no large publicly available corpus of Kannada language, a custom data set consisting of news articles was used after preprocessing. Using the fine-tuned model, abstractive summarization was achieved, thereby showing the applicability of transformer-based models in low-resource languages.

The analysis through various ROUGE scores indicates that the F1-scores achieved were 0.36 for ROUGE-1, 0.21 for ROUGE-2, and 0.31 for ROUGE-L. It indicates satisfactory performance of the model despite the small size of the dataset and complexities associated with the language. Although not up to expectations, the produced summaries are coherent and relevant. This highlights the efficacy of multilingual transformer-based models.

The next phase could include enhancing the model performance by increasing the size of the data, employing more powerful models like mT5-base, and ensembling. In addition, using other evaluation metrics such as BLEU scores and human evaluation, handling long documents, and adapting the model for different domains can improve its performance.

REFERENCES

[1] Aparna Madhukar Mete, Dr. M. L. Dhore, *"Abstractive Text Summarization for Hindi Language Using T5 Transformer"*, 2025 4th OPJU International Technology Conference (OTCON), IEEE, 2025.
[2] Yamini Niharika, Thirumalaraju Akhila, Priyanka C. Nair, "Summarization of Telugu Texts using versions of mT5, mBART50 and IndicBART," *2025 International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, 2025.
[3] Manisha Maharana, Meghna J., Ananya Kaushik, Anshika Ambavat, Amita Dev, Poonam Bansal, Aparna Kaushik, *"Hindi Text Summarization with Transformer-Based Ensembling and Refinement"*, 2025 5th International Conference on Intelligent Technologies (CONIT), IEEE, 2025.
[4] Manisha Maharana, Meghna J., Ananya Kaushik, Anshika Ambavat, Amita Dev, Poonam Bansal, Aparna Kaushik, *"Hindi Text Summarization with Transformer-Based Ensembling and Refinement"*, 2025 5th International Conference on Intelligent Technologies (CONIT), IEEE, 2025.
[5] Asmaa Elsaid, Ammar Mohammed, Lamiaa Fattouh, Mohammed Sakre,"Abstractive Arabic Text Summarization Based on MT5 and AraBART Transformers," *2023 Intelligent Methods, Systems, and Applications (IMSA)*, IEEE, 2023.
[6] Pushpa B. Patil, Dakshayani Ijeri and Reshma Pujari"Abstractive Text Summarization of Kannada using CNN-LSTM", *Proceedings of the 4th international conference on cognitive and intelligent Computing, Vol 2,* pp 21-29, 2026.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details