ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

TELUGU TEXT SUMMARIZATION USING EXTRACTIVE METHOD

[1]PILLI VENKATA RATNAM, [2]Y SRINIVAS RAJU

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Assistant Professor**,** Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

Text summarization plays a crucial role in managing and understanding large volumes of textual data, especially in regional languages like Telugu, where limited resources and tools are available. This project proposes an extractive text summarization system for Telugu language using deep learning techniques such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM). The system is designed to process Telugu text documents and generate concise summaries while preserving the original context and meaning. The dataset used for training consists of Telugu text and corresponding summaries, which are preprocessed using techniques such as tokenization, normalization, and TF-IDF vectorization to convert textual data into numerical form suitable for model training. The LSTM and Bi-LSTM models are trained using an encoder-decoder architecture, where the encoder learns the representation of input text and the decoder generates the corresponding summary. LSTM is particularly effective in capturing sequential dependencies and contextual information, making it suitable for summarization tasks. Experimental results show that the LSTM model achieves better performance with an accuracy of 71%, while the Bi-LSTM model achieves comparatively lower accuracy of 31%. The system also provides a user interface where users can upload Telugu text documents and obtain summarized outputs in real time. The proposed system demonstrates the effectiveness of deep learning approaches for regional language processing and provides a scalable solution for automatic text summarization. It can be applied in various domains such as news summarization, educational content analysis, and document management, contributing to advancements in natural language processing for Indian languages.

Keywords: LSTM and Bi-LSTM , TF-IDF, NLP

I. INTRODUCTION

With the rapid growth of digital content, the need for efficient text processing techniques has become increasingly important. Text

summarization is one such technique that helps in extracting key information from large documents and presenting it in a concise form. While significant progress has been made in text summarization for widely used languages such as English, limited research has been conducted for regional languages like Telugu. Telugu, being one of the most spoken languages in India, lacks sufficient tools and resources for automated text processing, making it challenging to manage and analyze large volumes of textual data effectively.

Traditional text summarization methods rely on statistical and rule-based approaches, which often fail to capture the contextual meaning of sentences. With the advancement of deep learning and natural language processing, more effective techniques have been developed to address these limitations. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models, have shown strong performance in handling sequential data and capturing long-term dependencies in text. Bidirectional LSTM (Bi-LSTM) further enhances this capability by processing text in both forward and backward directions, improving context understanding.

In this project, a Telugu text summarization system is proposed using extractive methods combined with deep learning techniques. The system utilizes LSTM and Bi-LSTM models in an encoder-decoder architecture to generate summaries from Telugu text documents. The dataset is preprocessed using TF-IDF vectorization, and the models are trained to identify important sentences for summarization. The system also provides a user interface for uploading documents and generating summaries in real time. This approach aims to improve accessibility and efficiency in processing Telugu textual data, contributing to advancements in regional language natural language processing.

II SURVEY OF RESEARCH

1. Traditional Text Summarization Techniques

Early text summarization methods were primarily based on statistical and rule-based approaches. Techniques such as frequency-based methods, sentence ranking, and keyword extraction were used to identify important sentences in a document. Algorithms like TextRank and TF-IDF played a significant role in extractive summarization. While these methods are simple and computationally efficient, they often fail to capture the semantic meaning and contextual relationships between sentences. Research indicates that such approaches are less effective for complex and large-scale datasets, especially in regional languages like Telugu, where linguistic structures are more diverse.

2. Machine Learning-Based Approaches

Machine learning techniques have been

introduced to improve summarization performance by learning patterns from data. Algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Decision Trees have been used to classify sentences based on their importance. These methods rely on manually engineered features such as sentence length, position, and keyword presence. Although machine learning improves accuracy compared to traditional methods, it still depends heavily on feature selection and does not fully capture deep contextual information in text.

3. Deep Learning for Text Summarization

Deep learning has significantly enhanced text summarization by enabling automatic feature extraction and better context understanding. Models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) are widely used for sequence modeling tasks. LSTM models are particularly effective in capturing long-term dependencies in text, making them suitable for summarization. Research shows that deep learning models outperform traditional and machine learning approaches in generating coherent and meaningful summaries.

4. Bidirectional LSTM (Bi-LSTM) Models

Bidirectional LSTM extends the capability of standard LSTM by processing input sequences in both forward and backward directions. This allows the model to capture context from both past and future tokens, improving understanding of sentence structure. Studies indicate that Bi-LSTM models provide better performance in tasks such as text classification and summarization. However, they require more computational resources and training time. This project compares LSTM and Bi-LSTM models to evaluate their effectiveness in Telugu text summarization.

5. Natural Language Processing for Regional Languages

Research in Natural Language Processing (NLP) for regional languages like Telugu is still developing due to limited datasets and resources. Challenges include complex grammar, lack of standardized datasets, and limited pre-trained models. Recent studies have focused on adapting deep learning techniques for regional languages to improve text processing tasks such as translation and summarization. This project contributes to this area by applying deep learning models to Telugu text summarization.

6. Evaluation Metrics and Performance Analysis

Evaluating summarization models is essential to measure their effectiveness. Common metrics include accuracy, precision, recall, and F1-score, along with text similarity measures such as ROUGE scores. Visualization techniques such as performance graphs help in comparing different models. Research

emphasizes the importance of using multiple evaluation metrics for a comprehensive analysis. In this project, LSTM and Bi-LSTM models are evaluated based on accuracy and performance comparison graphs, with LSTM showing better results.

III. WORKING METHODOLOGY

The proposed Telugu text summarization system begins with data collection and preprocessing using a Telugu summary dataset. The dataset consists of Telugu text documents along with their corresponding summaries. Preprocessing steps such as text cleaning, tokenization, stop-word removal, and normalization are applied to improve data quality. After preprocessing, TF-IDF vectorization is used to convert textual data into numerical representations suitable for deep learning models. The dataset is then divided into training and testing sets in an 80:20 ratio to ensure proper model evaluation. This stage prepares the data for effective learning and improves the overall performance of the models.

In the next phase, deep learning models such as LSTM and Bidirectional LSTM (Bi-LSTM) are implemented using an encoder-decoder architecture. The encoder processes the input Telugu text and captures contextual information, while the decoder generates the corresponding summary. The LSTM model learns sequential dependencies in the text, making it effective for summarization tasks. The Bi-LSTM model processes the sequence in both forward and backward directions, aiming to capture more contextual information. Both models are trained using the processed dataset, and their performance is evaluated using accuracy metrics. Hyperparameters such as learning rate, batch size, and epochs are tuned to achieve optimal results.

Finally, the trained models are used for generating summaries of new Telugu text documents. Users can upload text files through the system interface, and the best-performing model is applied to generate concise summaries. The system displays both the original text and the generated summary for comparison. Performance visualization tools such as accuracy graphs are used to analyze model behavior. The results show that the LSTM model outperforms the Bi-LSTM model in this case. This methodology provides an efficient and scalable solution for Telugu text summarization using deep learning techniques.

IV RESULTS EXPLANATIONS

In propose work we have used Telugu Summary dataset to train different deep learning algorithms such as LSTM and BI-LSTM to predict summary. LSTM algorithm is suitable for textual analysis and for summary generation.

In summary generation, LSTMs (Long Short-Term Memory) are a type of recurrent neural

network (RNN) that excel at processing sequential data, like text, and capturing long-term dependencies, making them suitable for tasks like understanding context and generating coherent summaries.

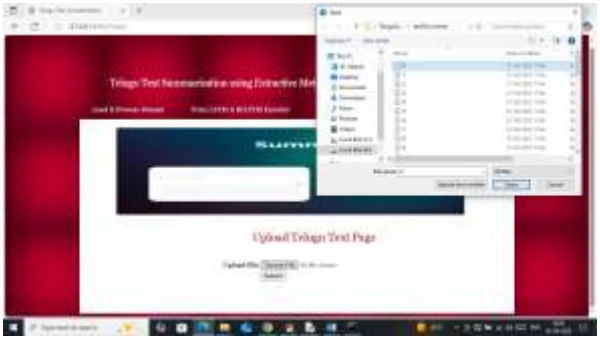
To implement this project we have designed following modules

1) User Login: user can login to system using username and password as 'admin and admin'.
2) Load & Process Dataset: after login user will run this module to load and process dataset and then generate TFIDF vector which will be input to LSTM algorithm to process sequential data and tp describe summary
3) Train LSTM & BILSTM Encoder: processed data will be input to LSTM and BILSTM to algorithm to train encoder and decoder where LSTM encoder will get trained on Telugu data and then Decoder will get trained on Telugu Summary and both this model will be utilize to generate summary from new test data and then calculate accuracy on test data
4) Text Summarization: using this module user can upload test document and then best performing encoder and decoder model will be applied to generate summar



In above screen dataset loaded and processed and now click on 'Train LSTM & BILSTM Encoder' link to train algorithms and then will get below page

In above screen in table format can see LSTM got 71% accuracy and BILSTM got 31% accuracy and can see performance of both algorithms in graph format where x-axis represents algorithm names and y-axis represents accuracy. Now click on 'Text Summarization' link to get below page



In above screen selecting and uploading 'Telugu Text Document' and then click on 'Submit' button to load document and then will get below summary output

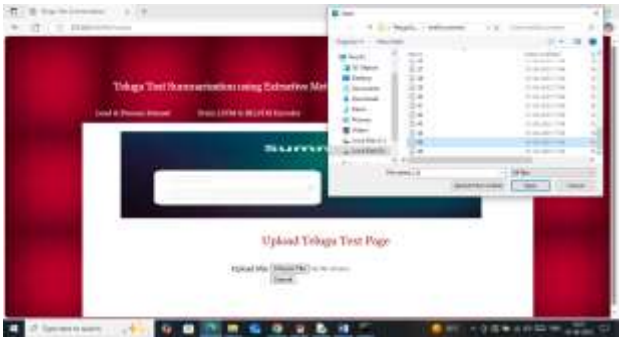




In above screen can see original uploaded text and predicted summary.



In above screen in first paragraph can see uploaded original TEXT and then in second paragraph can see generated summary. Similarly you can upload any document and generate summary. Below is another example



In above screen uploading another document and below is the summary output

V. CONCLUSION

The proposed Telugu Text Summarization system using extractive methods and deep learning techniques provides an effective solution for generating concise summaries from large Telugu text documents. By utilizing LSTM and Bi-LSTM models, the system is able to capture sequential dependencies and contextual information within the text. Among the implemented models, LSTM demonstrated superior performance with higher accuracy compared to Bi-LSTM, making it more suitable for this task. The use of preprocessing techniques and TF-IDF vectorization further enhanced the model's ability to understand and process textual data. The system successfully generates meaningful summaries that retain the core information of the original content. Additionally, the user-friendly interface allows users to easily upload documents and obtain summaries in real time. Overall, this project highlights the potential of deep learning in regional language processing and contributes to

advancements in natural language processing for Telugu.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder," in *Proc. EMNLP*, 2014.

[4] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017.

[5] T. Mikolov *et al.*, "Efficient estimation of word representations," 2013.

[6] C. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, 1988.

[8] F. Chollet, "Keras," 2015.

[9] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2016.

[10] Scikit-learn Developers, "Scikit-learn: Machine learning in Python," 2020.