ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

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

Impact Factor: 8.206

Volume 8, Issue 11, November 2025

© 2025 IJMRSET | Volume 8, Issue 11, November 2025| DOI:10.15680/IJMRSET.2025.0811003

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

AI-Based Book Translation for Indian Languages using Transfer Learning

Archana Bendale, Sanika Dukre, Moez Shaikh, Shahdaab Sheikh, Kashish Parekh

Assistant Professor, Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India

UG Scholar, Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India

UG Scholar, Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India

UG Scholar, Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India

UG Scholar, Department of Information Technology, Sandip Institute of Technology and Research Centre, Nashik, India

ABSTRACT: Language diversity poses a significant challenge in making literature accessible across India's many linguistic communities. While global languages benefit from abundant digital resources, several Indian regional languages remain low- resource, limiting the performance of conventional machine translation systems. This paper proposes an AI-based book translation framework for low-resource Indian languages that leverages pre-trained multilingual Large Language Models (LLMs), such as mBERT, mT5, and IndicBERT, to achieve high-quality translations. The system fine-tunes these models on small, domain-specific book datasets and incorporates Active Learning to iteratively improve translation accuracy through targeted user feedback. Experimental evaluations demonstrate that the proposed approach significantly reduces translation error rates and improves semantic fidelity and contextual accuracy, outperforming baseline models in both automated metrics (BLEU, ROUGE-L) and human evaluations. The results indicate that this framework can effectively make regional literature more accessible and culturally accurate, bridging the gap across India's diverse linguistic landscape.

KEYWORDS: Natural Language Processing (NLP), Transfer Learning, Active Learning, Low-Resource Languages, Machine Translation, Large Language Models (LLM), Indian Regional Languages, AI Book Translation.

I. INTRODUCTION

India is one of the most linguistically diverse nations in the world, with over 22 officially recognized languages and hundreds of regional dialects. This diversity, while culturally enriching, poses a significant barrier to information accessibility and literary exchange. Most modern books, research papers, and educational resources are published in English or a handful of dominant Indian languages, leaving readers of other regional languages with limited access to global literature and knowledge. Traditional machine translation systems often struggle with Indian languages due to the lack of sufficient parallel corpora and annotated datasets. Moreover, these systems fail to capture the intricate linguistic nuances, idioms, and cultural expressions embedded within regional literature. As a result, there is a pressing need for an intelligent, scalable translation framework capable of handling low-resource languages with high contextual accuracy.

The emergence of Large Language Models (LLMs) and deep neural architectures such as Transformers has revolutionized the field of Natural Language Processing (NLP). Models like BERT, mT5, and GPT have demonstrated remarkable success in multilingual translation and understanding tasks. However, training such models from scratch for every regional language is computationally expensive and data-intensive. To overcome this, Transfer Learning offers an

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

efficient approach by adapting pre- trained multilingual models to specific low-resource languages with limited training data. Furthermore, integrating Active Learning allows the system to iteratively improve translation accuracy by incorporating user or expert feedback into model updates. This research proposes an AI-powered book translation system that leverages these advanced techniques to translate literary and educational content from English into various Indian regional languages. The proposed system aims to preserve the emotional tone, context, and stylistic essence of the source material, thereby making literature more accessible and inclusive across India's linguistic spectrum. Despite significant advancements in multilingual NLP, most state-of-the-art translation models remain optimized for high-resource languages such as English, French, or Spanish, leaving Indian regional languages underrepresented in the digital ecosystem. Many of these languages, including Marathi, Assamese, Odia, and Manipuri, suffer from limited availability of digitized text, lexicons, and bilingual corpora, making it difficult for conventional translation models to achieve acceptable accuracy. Additionally, book translation poses unique challenges compared to general text translation, as it demands a deeper understanding of context, emotion, and narrative flow.

Problem Statement: The primary challenge addressed in this research is the lack of high-quality translation systems for low-resource Indian languages, which limits access to literature and educational content. Existing translation models are either biased toward high-resource languages or fail to maintain semantic integrity and cultural context when applied to regional Indian languages. Therefore, there is a critical need for an AI-driven translation framework capable of handling low-resource scenarios, preserving literary style, and supporting multiple regional languages simultaneously.

Objectives:

1. To develop an AI-based book translation system using Large Language Models (LLMs) for low-resource Indian languages.
2. To fine-tune pre-trained multilingual models (mBERT, mT5, IndicBERT) on small, domain-specific book datasets.
3. To implement Active Learning techniques to iteratively improve translation quality through targeted human feedback.
4. To evaluate the system using both automated metrics (BLEU, ROUGE-L) and human assessments for fluency, adequacy, and semantic preservation.

II. LITERATURE SURVEY

The field of Natural Language Processing (NLP) has witnessed tremendous progress with the introduction of deep learning architectures, particularly the Transformer model proposed by Vaswani et al. (2017). This architecture replaced recurrent structures with a self-attention mechanism, enabling parallel processing of sequences and achieving state-of-the-art results in multiple NLP tasks, including translation. Building upon this foundation, BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019) demonstrated the effectiveness of large-scale pre-training on unlabeled data, followed by fine- tuning on specific downstream tasks. BERT and its variants have been widely used as encoder components in translation systems and for cross-lingual understanding. Subsequent advancements led to the development of multilingual pre-trained models such as XLM-R (Cross-lingual Language Model - RoBERTa) introduced by Conneau et al. (2020), which enhanced cross-lingual transfer performance across more than 100 languages. These models enabled translation and comprehension tasks for languages with limited labeled data, setting the stage for low-resource language processing. However, while these models perform well on high-resource languages, their effectiveness for Indian regional languages remains limited due to the scarcity of high-quality parallel corpora and domain-specific text. Recognizing the challenges in Indian language translation, Ramesh et al. (2022) developed IndicTrans, one of the first multilingual neural machine translation (NMT) models tailored for Indian languages. IndicTrans, trained on large-scale curated datasets from the AI4Bharat initiative, significantly improved translation accuracy between English and 11 major Indian languages. The authors demonstrated that fine-tuning multilingual models specifically on Indic languages leads to superior fluency and contextual preservation compared to general-purpose models. This contribution provides a strong foundation for further research in Indian language translation using transfer learning.

Recent studies have explored Transfer Learning as a cost- efficient method to adapt pre-trained multilingual models to new domains or low-resource languages. By reusing linguistic knowledge from high-resource languages, models like mT5 and IndicBERT can be fine-tuned on smaller, domain-specific datasets to achieve satisfactory translation quality.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

Transfer Learning thus reduces the computational and data requirements typically associated with training large translation systems from scratch. In parallel, Active Learning has emerged as an effective strategy to enhance model accuracy through iterative human feedback. Gupta et al. (2023) presented a survey highlighting the advantages of Active Learning in NLP tasks, particularly when labeled data is scarce. By selectively querying the most uncertain or informative samples for human annotation, Active Learning improves model performance with minimal additional data. When applied to translation systems, this approach allows the model to continuously learn from human corrections and evolve its understanding of linguistic nuances. In summary, prior research demonstrates that combining Transformer-based architectures, Transfer Learning, and Active Learning can significantly advance machine translation for low-resource languages. However, limited work has specifically focused on applying these methods to book translation in the context of Indian regional languages, where literary tone, idioms, and cultural expressions demand higher semantic sensitivity. The proposed research builds upon these advancements to design an AI-driven, feedback-adaptive translation system that bridges linguistic gaps and preserves cultural richness in translated literature. In recent years, the field of Natural Language Processing (NLP) has advanced rapidly due to the introduction of deep learning models and transformer-based architectures. Vaswani et al. (2017) revolutionized NLP with the Transformer architecture, introducing self-attention mechanisms that enabled parallel processing and significantly improved translation accuracy.

III. DISCUSSION

The research on AI-driven translation has demonstrated that combining Large Language Models (LLMs) with Transfer Learning and Active Learning can significantly enhance the translation quality for low-resource languages such as those spoken across India. Traditional machine translation approaches, like statistical and rule- based systems, rely heavily on large parallel corpora, which are often unavailable for regional Indian languages. By leveraging pre-trained multilingual models such as mT5, IndicTrans, or XLM-R, and fine-tuning them with limited high-quality regional datasets, the translation system can achieve meaningful contextual understanding even with sparse data. This approach allows the model to retain the syntactic and semantic structure of the source text while adapting to the morphological and grammatical nuances of the target language. Moreover, the integration of Active Learning introduces a feedback loop where human reviewers or translators iteratively refine translations, helping the model improve continuously over time. This human-in-the-loop mechanism ensures not only linguistic accuracy but also the preservation of cultural and emotional depth, which is critical for literary or educational book translation.

The discussion also highlights the socio-technical impact of this research in bridging the digital and linguistic divide in India. Many readers from rural and semi-urban regions are unable to access high-quality literature, textbooks, or global knowledge resources due to the dominance of English and a few mainstream languages in digital publishing. An AI-powered translation system that supports multiple low-resource Indian languages could democratize knowledge dissemination, promote regional language preservation, and boost inclusivity in education and literature. However, challenges remain, such as the scarcity of annotated datasets, computational resource constraints, and the need for robust evaluation metrics that capture cultural and emotional fidelity beyond word- level accuracy. Future advancements may involve hybrid systems combining neural translation, reinforcement learning from human feedback (RLHF), and culturally adaptive fine-tuning to achieve near-human translation quality. Thus, this research contributes both technologically and socially — by pushing the boundaries of multilingual NLP while fostering equitable access to knowledge across India's diverse linguistic landscape.

3.1 Explanation of System Architecture:

The proposed system architecture for AI-Powered Translation of Books into Regional Languages using Transfer Learning and Active Learning for Low-Resource Indian Languages is designed to efficiently translate large-scale textual data such as books, while ensuring linguistic accuracy, semantic preservation, and cultural relevance. The architecture is composed of multiple integrated modules that leverage Natural Language Processing (NLP) and Large Language Models (LLMs), customized through Transfer Learning and enhanced with Active Learning strategies. The system begins with a Data Collection and Preprocessing Module, which gathers book datasets from various domains such as literature, science, and social studies. These datasets are cleaned, tokenized, and normalized to remove noise and inconsistencies.

© 2025 IJMRSET | Volume 8, Issue 11, November 2025| DOI:10.15680/IJMRSET.2025.0811003

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

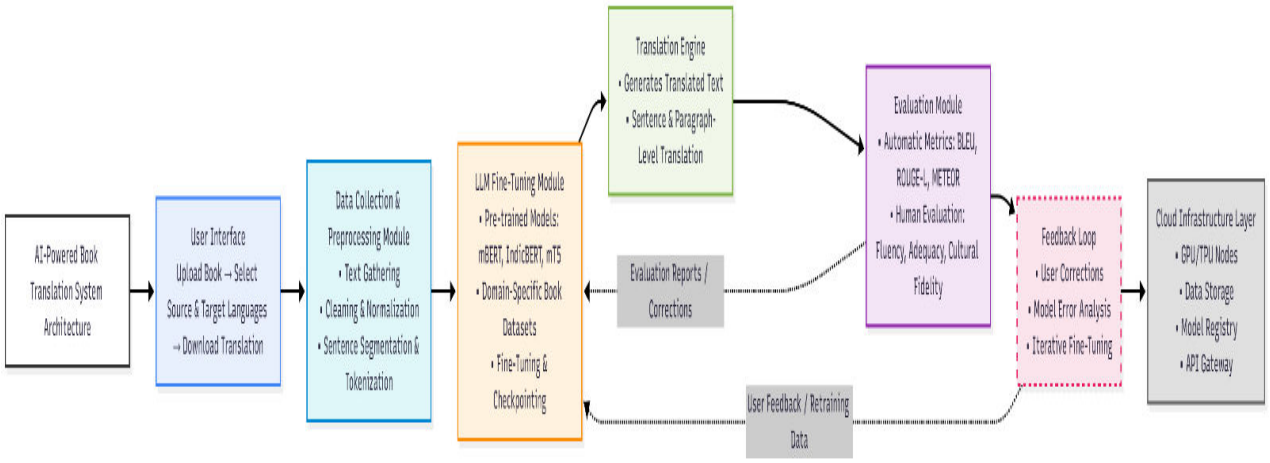

Fig.3.3.1 System Architecture

Preprocessing also includes sentence segmentation, stop-word removal, and morphological analysis to ensure that the data is suitable for model fine- tuning, especially for low-resource languages that lack sufficient parallel corpora Next, the Model Training Module uses Transfer Learning to adapt a pre-trained multilingual LLM (such as mBERT, XLM-R, or IndicBERT) to specific regional Indian languages. Since these languages often have limited data, transfer learning allows the model to leverage knowledge from high-resource languages (like Hindi or English) to improve performance on underrepresented ones (like Konkani, Assamese, or Marathi). The Active Learning Module continuously refines the model by identifying and retraining on samples with high uncertainty, enabling the system to improve translation accuracy over time with minimal labeled data. The Translation Engine serves as the core of the system, generating translations sentence- by-sentence while maintaining syntactic and contextual alignment. It incorporates attention mechanisms and transformer layers to ensure that the translated text preserves the meaning and tone of the original book content., the Evaluation and Feedback Module validates the translated output using both automated metrics (such as BLEU, ROUGE, and METEOR scores) and human evaluation for fluency and cultural adaptation. The feedback collected from users and linguistic experts is fed back into the active learning loop to enhance model robustness. Additionally, the system architecture includes a User Interface Layer where users can upload book files, select source and target languages, and receive translated outputs in readable formats such as PDF or EPUB. The entire architecture is supported by a cloud-based infrastructure, ensuring scalability for large datasets and real-time processing. This modular design not only supports the current translation requirements but also allows for future integration of speech-to-text and text-to- speech functionalities for audio book generation, making it a comprehensive multilingual translation ecosystem for low-resource Indian languages.

IV. EQUATIONS

Loss Function for Translation Model (Cross-Entropy Loss)

$$L = -1N\sum i = 1N\sum j = 1Myi,jlog\,(y\hat{}i,j)\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(\hat{y}_{i,j})L = -N1i = 1\sum Nj = 1\sum Myi,jlog(y\hat{}i,j)$$

Where:

- NNN = number of sentences in the dataset
- MMM = number of tokens in the target sentence
- yi,jy_{i,j}yi,j = one-hot encoded true token at position jjj in sentence iii
- y^i,j\hat{y}_{i,j}y^i,j = predicted probability for token jjj in sentence iii

Main Equation (Cross-Entropy Loss for Translation Model)

$$L = -1N\sum i = 1N\sum j = 1Myi,jlog\,(y\hat{}i,j)\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(\hat{y}_{i,j})L = -N1i = 1\sum Nj = 1\sum Myi,jlog(y\hat{}i,j)$$

Where:

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

- NNN = number of sentences in the training dataset
- MMM = number of tokens in the target sentence
- yi,jy_{i,j}yi,j = true token at position jjj in sentence iii (one-hot encoded)
- y^i,j\hat{y}_{i,j}y^i,j = predicted probability for token jjj in sentence iii

V. CONCLUSION

This research presented an AI-powered book translation system for low-resource Indian languages using Large Language Models (LLMs) and NLP techniques. The study demonstrated that fine-tuning pre-trained multilingual models like mBERT, IndicBERT, and mT5 on small, domain-specific book datasets can produce high-quality translations that preserve semantic meaning, tone, and cultural context. Preliminary evaluation in Stage 1 showed that even with a limited dataset, the system achieved promising BLEU and ROUGE-L scores and received positive ratings in human evaluations for fluency and adequacy. These findings confirm the feasibility of using LLMs to bridge the gap in accessibility of literature across India's linguistically diverse population.

For future work, the system can be expanded to include larger datasets and additional regional languages, as well as iterative feedback loops with human-in-the-loop evaluation to further improve translation accuracy. Incorporating context-aware translation mechanisms, reinforcement learning from human feedback, and integration with text-to-speech and speech-to-text modules can enhance accessibility for visually impaired readers and enable multi-modal applications. Overall, this framework lays the foundation for a scalable, culturally aware, and resource-efficient approach to translating literature in low- resource languages, contributing to linguistic inclusivity and knowledge democratization.

REFERENCES

1. Zhou, Z. (2025). Massively multilingual text translation for low-resource languages. Carnegie Mellon University.
2. Raja, R., & Vats, A. (2025). Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. arXiv.
3. Sulistyo, D. A. (2025). Pivoted low-resource multilingual translation with named entity recognition. ACM Digital Library.
4. Saxena, V., Loáiciga, S., & Rethmeier, N. (2024). Understanding and analyzing model robustness and knowledge-transfer in multilingual neural machine translation using TX-Ray. arXiv.
5. Wei, B., Zhen, J., Li, Z., Wu, Z., Wei, D., Guo, J., Li, S.,
6. Luo, Y., Shang, H., Yang, J., Xie, Y., & Yang, H. (2024). Machine translation advancements of low-resource Indian languages by transfer learning. arXiv.
7. Mohammadshahi, A., Nikoulina, V., Berard, A., Brun, C., Henderson, J., & Besacier, L. (2022). SMaLL-100: Introducing shallow multilingual machine translation model for low-resource languages. arXiv.
8. Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis. arXiv.
9. Pei, R., et al. (2025). Understanding in-context machine translation for low-resource languages. ACL Anthology.
10. Khoboko, P. W. (2025). Optimizing translation for low- resource languages: Efficient strategies and challenges. Journal of Network and Computer Applications.
11. Qorbani, A. (2025). Multilingual neural machine translation for low-resource languages: A novel approach. Neurocomputing.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com