International Journal of Inventive Engineering and Sciences (IJIES)
ISSN: 2319-9598 (Online), Volume-12 Issue-10, October 2025

Leveraging Transfer Learning for NLP in Extremely Low-Resource Language Settings

Abhay Bhatia, Anil Kumar

Abstract: ***Globally, individuals are increasingly gaining access to current technologies. They facilitate unprecedented access to knowledge, justice, and information. To ensure accommodation and universal accessibility, a globally spoken language must be considered, and it is essential to facilitate computers' comprehension of human language. The computational methodologies that exist have advanced significantly; yet these improvements can often require substantial resources, including data generation and processing. Such a reliance on resources hinders the efficient advancement of cross-lingual transfer techniques, especially for tasks like discourse analysis that require rigorous training and evaluation across more spoken languages and domains. The existing systems consume N resources, and the best language processing methods haven't been able to cover many languages and domains. Transfer learning seeks to solve this problem by leveraging pre-trained models trained on large datasets to work in resource-constrained environments. These strategies have gained popularity for their effectiveness in managing limited resources across diverse tasks, areas, and languages. This research focuses on the application of transfer learning approaches to Indian languages, notably Hindi and its code-mixed English-Hindi variety, which is commonly found on social networking sites. We examine cross-task and cross-lingual transfer strategies across several downstream tasks, demonstrating their effectiveness despite limited training data and computational resources. We also provide a syntactic-semantic curriculum-based learning architecture for English-Hindi code-mixed sentiment analysis, resulting in significant performance improvements.***

Keywords: Cross-Lingual Transfer, Discourse Analysis, Low-Resource Languages, Resource-Intensive Systems, Transfer Learning, Pre-Trained Models

Nomenclature:
NER: Named Entity Recognition
ICL: In-Context Learning
NMT: Neural Machine Translation
NLP: Natural Language Processing

I. INTRODUCTION

Transfer learning has changed the game for natural Language processing (NLP) in low-resource languages, especially in India. By sharing information from languages with a lot of data, this strategy makes languages with less data work much better. Below, we discuss some essential ways transfer learning can help you deal with Indian languages that lack many resources. Named Entity Recognition (NER): NER is a valuable tool for extracting crucial information from text. You can use it for things like managing content and helping customers. Research indicates that multilingual models, particularly those utilising versions of BERT, outperform monolingual models for Indian languages such as Hindi and Marathi. This shows how helpful it can be to leverage what you know about comparable languages.

NMT (Neural Machine Translation) faces significant challenges in low-resource languages such as Konkan due to insufficient training data. Innovative techniques, such as back translation, have been employed to generate synthetic data, significantly enhancing translation accuracy. Research demonstrates that improvements in BLEU scores are clearly visible across several places, highlighting the efficiency of these techniques, such as mitigating data scarcity.

Classification of Relationships: To facilitate classification, the IndoRE dataset was developed across various Indian languages. It was executed because of issues with reliable datasets. Transfer learning reduces the need for extensive human annotation, thereby enhancing the efficiency of relation extraction and improving performance across several languages.

Learning in Context (ICL): In-context learning (ICL) employs large language models to perform tasks in resource-scarce languages with limited contextual information. Research highlights the need to align semantics across languages to mitigate disparities between high- and low-resource languages, thereby improving performance in resource-limited contexts.

Transfer learning has demonstrated considerable potential for enhancing NLP in resource-scarce languages; however, some challenges remain. This entails acquiring high-quality datasets and addressing the complexities associated with multilingualism. We must rectify these deficiencies to advance in this field.

II. PROBLEMS AND PLANS FOR THE FUTURE

Transfer learning has made significant progress in improving NLP for Indian languages with limited resources; however, some problems remain to be addressed.

Availability of high-quality data: The scarcity of well-annotated datasets for many Indian languages undermines the effectiveness of fine-tuning and transfer learning.

Diversity in Language: The existing languages possess a diverse syntax, scripts and morphology. The complication had developed universally in all applicable models.

Manuscript received on 30 September 2025 | Revised Manuscript received on 07 October 2025 | Manuscript Accepted on 15 October 2025 | Manuscript published on 30 October 2025.
*Correspondence Author(s)
Dr. Abhay Bhatia*, Associate Professor, Department of Computer Science and Engineering, Roorkee Institute of Technology, Roorkee (Uttarakhand), India. Email ID: dhawan.abhay009@gmail.com, ORCID ID: 0000-0001-7220-692X
Dr. Anil Kumar, Associate Professor, Department of Computer Science and Engineering, Roorkee Institute of Technology, Roorkee (Uttarakhand), India. Email ID: chauhananil01@gmail.com

A. Computational: Training large models for transfer learning requires substantial computational resources, which may not be readily available to academics and organisations working with low-resource languages. Code-Mixing and Multilingualism: The widespread use of code-mixed texts (e.g., Hindi-English) on social media and other platforms complicates matters, necessitating specialised techniques to manage mixed-language data.

B. Conclude this Section: Transfer learning has completely changed how the mostly spoken Indian languages with few resources handle natural language. This method has made a lot of progress in tasks named as Named Entity Recognition (NER), relation categorization, Neural Machine Translation (NMT) [5], and sentiment analysis by using the best part of the languages with a lot of resources. Even though there are specific issues, the constant generation of ideas and a focus on low-resource contexts can eventually lead to the latest language technologies that are more open and accessible to everyone. Such activity gives the power to different spoken language groups throughout the country (India).

III. LITERATURE SURVEY

The research seeks to enhance Named Entity Recognition (NER) for low-resource Indian languages, specifically Hindi and Marathi, by applying transfer learning techniques utilising various BERT adaptations. This demonstrates that multilingual models outperform monolingual ones. The research presents IndoRE, a dataset for relation categorisation in Indian languages, utilising transfer learning with mBERT. It looks at the pros and cons of using gold and silver data, active learning for annotation, and an ensemble model to improve performance when resources are limited [Arijit Nag] [2]. The paper shows that using multilingual sentiment lexicons with models like XLM-R greatly improves zero-shot sentiment analysis in low-resource languages, such as Indian languages, without needing sentence-level sentiment data. This is better than models fine-tuned on English datasets. Neural transfer learning in NLP leverages pre-trained models like BERT and GPT-3 to improve performance on downstream tasks when little labelled data is available. Multi-tasking, domain adaptation and Fine-tuning are notable methodologies. These addressing solutions faced challenges such as dataset bias and generalisation. Nitin Sharma [4]. This research paper introduced a transfer-learning-based methodology for low-resource neural machine translation, markedly improving BLEU scores by first training on high-resource language pairs and then transferring the learned parameters to a low-resource pair, thereby achieving enhanced performance relative to baseline models. The research examines the use of TL methodologies that aim to improve English-Khasi neural machine translation in low-resource settings. It emphasises improving translation quality by using pre-trained models, even when substantial data is absent. Aiusha Vellintihun Hujon [6]. The research examines the trivial transfer learning by employing a pretrained model on available high-resource language pairs without any modifications. This demonstrates the efficiency of translation performance enhancement in low-resource languages and examines its effectiveness across varying training settings.

IV. COMPARATIVE STUDY

A. Comparative Study on Transfer Learning for Low-Resource Indian Languages

A comparative study of limited-resource Transfer Learning for Indian Languages is conducted. Day by day, transfer learning is becoming a key aspect of making NLP [8][9] better in regions like India, where there are many languages and not a lot of resources. A comparative review of approaches, problems, and outcomes elucidates its application in tasks such as Named Entity Recognition (NER), Neural Machine Translation (NMT) [7], Relation Classification, and Sentiment review. We shall examine and compare key works and methods in this field below.

i. Named Entity Recognition (NER)

- Monolingual Models: Older NER approaches employed monolingual models that were trained on tiny datasets that were just for one language. Even if they supplied baseline performance, they often struggled to generalise because there wasn't enough data.

ii. Ways to Learn from Other People

- *Multilingual Models:* Research, like that by Sabane et al. (2023) [1], demonstrates that multilingual pre-trained models, such as mBERT, significantly outperform monolingual models by leveraging shared linguistic features from related languages, including Hindi, Marathi, and Tamil.
- *Cross-Lingual Transfer:* Researchers achieved superior outcomes by refining models on a high-resource language such as Hindi before transitioning them to a low-resource language like Konkani, rather than training monolingual models from the beginning.

iii. Things to be Noted

- *Performance Gains:* Transfer learning makes F1 scores better for NER tasks in languages that don't have a lot of resources.
- *Dependence on Relatedness:* When languages have comparable grammar or words, they work better.

iv. Neural Machine Translation (NMT)

- *Statistical Machine Translation (SMT):* Early NMT systems utilised parallel corpora, which are hard to find for many Indian languages; therefore, they didn't operate as well as they could have.

v. Ways to Learn from Transfer

- *Backtranslation*: Prasada and Rao (2024) demonstrated that backtranslation generates synthetic data by translating from the target language to the source language. This makes datasets larger for languages like Kannada that lack many resources.
- *NMT Models (Multilanguage):* These models generally learn from a common available vocabulary across all languages, which helps them move knowledge

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

from languages with a lot of resources, like Hindi, Punjabi or Malayalam, to languages with fewer resources, like Manipuri.

vi. *Important:*
- *Resource Efficiency while using TL:* Transfer learning diminishes the need for vast parallel corpora; yet, it necessitates high-quality data for resource-abundant languages.
- *Effectiveness of TL:* The transfer learning and backtranslation significantly enhance translations with limited resources, evidenced by elevated BLEU scores.

vii. *Classification of Relations:*
- *Statistical Methods:* as per history, every individual formulated rule that are generally generated manually or employ statistical models that require substantial amounts of labelled data. This data is presently unavailable for Indian languages.

viii. *Learn from Transfer:*
- *Pre*-Trained *Models:* Utilising datasets such as IndoRE and pre-trained models, eg, RoBERTa and XLM-R, which are fine-tuned explicitly for classification tasks, demonstrated their superior performance.
- *Cross-Lingual* Transfer*:* Utilizing annotated data from resource-rich languages considerably simplifies the annotation process for languages with limited resources.

ix. *Needed for it:*
- *Scalability*: Transfer learning works well for languages that have similar grammar rules
- *Data Dependency:* Pre-trained models need fewer labelled instances, although they still operate better with datasets that are of high quality.

B. Sentimental Analysis [10]

i. *Conventional Methods:*
- *Lexicon-Based Models:* These models used sentiment lexicons to provide polarity scores; however, it's an issue for them that they didn't work well for material that mixed languages or depended on the context.

ii. *Transfer Learning Methodology:*
- *Code-Mixed Data:* Research employing syntactic-semantic curricular learning frameworks for English-Hindi code-mixed sentiment analysis indicates significant enhancements.
- *Large Language Models:* Models such as GPT and mBERT, optimized for sentiment analysis, perform effectively with code-mixing and domain-specific content.

iii. *Keep takeaways:*
- *Code-Mixing Challenge:* Transfer learning facilitates the analysis of sentiments regarding code-mixed text, which poses challenges due to the utilization of disparate languages and scripts.
- *Context Sensitivity:* Fine-tuned transformer-based models exhibit superior sensitivity to nuanced sensations compared to conventional lexicon-based models.

Table I: Comparison Table for NET v/s NMT

Aspect	NER	NMT	Relation Classification	Sentiment Analysis
Data Dependency	Moderate to High	High (parallel corpora)	Moderate	Low to Moderate
Model Preference	Multilingual Models (e.g., mBERT)	Multilingual NMT, Backtranslation	Pre-Trained Models (e.g., XLM-R)	Fine-Tuned Transformers (e.g., GPT)
Challenges	Data scarcity, language diversity	Parallel data scarcity, domain adaptation	Annotation effort, domain generalization	Code-mixing, sentiment ambiguity
Performance Gains	Significant with cross-lingual transfer	High with backtranslation and multilingual NMT	Moderate to High with IndoRE	High with curriculum learning frameworks
Unique Needs	Syntactic and semantic alignment	Domain-specific vocabulary	Shared feature representation	Handling mixed-language and contextual cues

V. CHALLENGES AND OPPORTUNITIES

A. Common Challenges

i. The lack of high-quality annotated data for several Indian languages remains a significant obstacle.

ii. Addressing mixed scripts and languages, prevalent on social media, necessitates specialized models and preprocessing techniques.

iii. The varied syntactic structures and morphological alterations of linguistic diversity complicate the generalization of models.

B. Opportunities

i. Methods such as backtranslation and data augmentation can compensate for insufficient training data.

ii. Task-specific learning frameworks enhance model performance in contexts characterized by extensive contextual and code mixing.

iii. Models such as mBERT and XLM-R provide a practical

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

foundation for scalable cross-lingual transfer.

VI. CONCLUSION AND FUTURE SCOPE

Transfer learning in NLP for Indian languages with limited resources has significant room for further expansion. There are various approaches to address current problems and improve tasks such as NER, NMT, sentiment analysis, and relation categorisation, as technology advances rapidly and people become more aware of the need for linguistic inclusion. Here are some key areas for the future:

A. Making Datasets of High Quality

Low-resource languages still have a big problem because there aren't any annotated datasets. Future initiatives should focus on:

i. *Making Data with the Help of the Community:* Getting local communities, linguists, and native speakers to work together to produce annotated datasets for languages that aren't well-represented.
ii. *Cross-Lingual Data Sharing:* Making systems that help you share datasets between languages that have similar rules for how to use them.

B. Making Models That are Specific to a Certain Field. Many Fields Need Models, such as Healthcare, Agriculture, and the Law. Subsequent Research May Examine

i. *Transfer Learning for Domain Adaptation:* Fine-tuning models that have already been taught to operate better on tasks that are peculiar to Indian languages.
ii. *Multimodal Models:* Employing textual, auditory, and visual data concurrently to address language-specific issues in sectors such as healthcare and education.

C. Talking about Code-Mixed Language Processing [11]

Code-mixing, prevalent in informal communications such as social media interactions, remains challenging to rectify. Subsequent research may concentrate on:

i. *Integrated Frameworks:* Developing models that seamlessly operate across several scripts and languages.
ii. *Linguistic Context Alignment:* Utilizing embeddings from many *languages* to enhance individuals' understanding of the implications of mixed-language sentences.

D. Increasing the Number of Multilingual Pre-Trained Models

We can make and use already existing multilingual models like mBERT and XLM-R for the betterment of Indian languages in the following manner:

i. *Adding More Indian/Local Languages:* Upending the languages that aren't well represented to us, such as Konkani, Bhojpuri, and Manipuri, to the pre-training datasets so that we can find better results.
ii. *Tokenisation for Diverse Languages:* The tokenisation approaches generally work with the diverse *morphological* orders as well as the syntax of Indian languages.

E. Seeking new ways as part of a learning journey

New approaches to learning hold promise for improving Natural Language Processing in languages with limited resources.

i. *Few-Shot and Zero-Shot Learning:* Some training models use very little or even no tagged data. [3]
ii. *In-Context Learning (ICL):* Use of enormous language models for completing tasks with few examples, which *means* you won't require as many annotated datasets.

F. Inclusion of Linguistic and Their Cultural Subtitles

NLP systems are built to require knowledge of the linguistic and cultural specifications of Indian languages.

i. *Morphological Understanding:* For making models in a better way so that they can handle the complex morphology and the inflexions of languages like Kannada, Konkani, and Sanskrit.
ii. *Cultural Relevance:* Creating algorithms for sentiment analysis [12] and *relation* classification that consider cultural expressions and idiomatic language.

G. Working with an Open Research

It will be beneficial as far as considering the future and its growth in the sector, only if we work together:

i. *Cross-Institutional Partnerships*: Cooperations, Encouragements in schools, NGO, and other government organisations for working together to achieve the pace and share resources with desired information.
ii. *Open-Source Contributions:* The Main motive is to encourage people to share their datasets as part of open datasets, models, and research contributions to make it easier for everyone to get included and come up with some boosted ideas.

H. Real Life Implementations

Future studies must focus on the practical uses of transfer learning models in real-world scenarios:

i. *Digital Inclusion:* To achieve it, we need voice assistants, some chatbots, and a brief translation system that can offer more Indian languages for groups that don't have enough of them [12].
ii. *Schooling:* Creation of language learning aids or TLM that can be used as content translation systems to fill in the gaps in schooling.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article

does not necessitate ethical approval or consent to participate with supporting documentation.

- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. M. Sabane, A. Ranade, O. Litake, P. Patil, R. Joshi and D. Kadam, "Enhancing Low-Resource NER using Assisting Language and Transfer Learning," *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 2023, pp. 1666-1671, DOI: https://doi.org/10.1109/ICAAIC56838.2023.10141204
2. Transfer Learning for Low-Resource Multilingual Relation Classification Arijit Nag, Biswarup Samanta, Animesh Mukherjee, Niloy Ganguly, Soumen Chakrabarti 08 Aug 2022. 22, Iss: 2, pp 1-24 DOI: http://doi.org/10.1145/3554734.
3. Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon Fajri Koto, Tilman Beck, Zeerak Talat, Iryna Gurevych, Timothy Baldwin. DOI: https://doi.org/10.48550/arXiv.2402.02113.
4. Nitin Sharma, Bhumica Verma; Recent Advances in Transfer Learning for Natural Language Processing (NLP), A Handbook of Computational Linguistics: Artificial Intelligence in Natural Language Processing Federated learning for Internet of Vehicles: IoV Image Processing, Vision and Intelligent Systems (2024) 2: 228. DOI: https://doi.org/10.2174/9789815238488124020014
5. Transfer Learning for Low-Resource Neural Machine Translation Barret Zoph, Deniz Yuret, Jonathan May, Kevin Knight 01 Apr 2016 (Association for Computational Linguistics) - pp 1568-1575, DOI: https://doi.org/10.48550/arXiv.1604.02201.
6. Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings Aiusha Vellintihun Hujon, Thoudam Doren Singh, Khwairakpam Amitab 01 Jan 2023 - Procedia Computer Science - Vol. 218, pp 1-8, DOI: https://doi.org/10.1016/j.procs.2022.12.396.
7. Exploring Benefits of Transfer Learning in Neural Machine Translation Tom Kocmi 05 Dec 2019, DOI: https://doi.org/10.48550/arXiv.2001.01622.
8. M. Kumar, S. Ali Khan, A. Bhatia, V. Sharma and P. Jain, "A Conceptual Introduction of Machine Learning Algorithms," *2023 1st International Conference on Intelligent Computing and Research Trends (ICRT)*, Roorkee, India, 2023, pp. 1-7, DOI: https://doi.org/10.1109/ICRT57042.2023.10146676
9. Kumar, A., Bhatia, A., Kashyap, A., & Kumar, M. (2023). LSTM network: a deep learning approach and applications. In *Advanced Applications of NLP and Deep Learning in Social Media Data* (pp. 130-150). IGI Global, DOI: https://doi.org/10.4018/978-1-6684-6909-5.
10. Verma, Praveen, et al. "Sentiment analysis "using SVM, KNN and SVM with PCA." *Artificial Intelligence in Cyber Security: Theories and Applications*. Cham: Springer International Publishing, 2023. 35-53, DOI: https://doi.org/10.1007/978-3-031-28581-3.
11. Bhatia, A. (2024). The Role of Cutting-Edge Technologies in Revolutionary Industry 5.0. In *Artificial Intelligence and Communication Techniques in Industry 5.0* (pp. 128-153). CRC Press, DOI: https://doi.org/10.1201/9781003494027.
12. Bhatia, A., Bhatia, P., & Sood, D. (2024). Leveraging AI to transform online higher education: Focusing on personalized learning, assessment, and student engagement. *International Journal of Management and Humanities (IJMH) Volume-11 Issue-1*, DOI: https://doi.org/10.35940/ijmh.A1753.11010924.

AUTHOR'S PROFILE

Dr. Abhay Bhatia is an accomplished academician and researcher, serving as an Associate Professor in the Department of Computer Science and Engineering at Roorkee Institute of Technology, Uttarakhand. With over 13 years of teaching and research experience, he holds a B. Tech and an M. Tech in Computer Science, and a PhD in Wireless Sensor Networks. An active IEEE member, he has published over 34 papers, authored 11 book chapters, and filed seven patents. His authored books include Fundamentals of IoT and Practical Approach to Machine Learning with TensorFlow. His research interests include Artificial Intelligence, Machine Learning, and Wireless Sensor Networks.

Dr. Anil Kumar is presently associated with Roorkee Institute of Technology (RIT), Roorkee, Haridwar, Uttarakhand, as an Associate Professor in the Department of Computer Science and Engineering. He has more than 19 years of academic experience and has worked with various reputed engineering institutions. He has completed his B.Tech in IT from AKTU (formerly UPTU), M. Tech in Computer Science and Engineering, and PhD in Natural Language Processing. He is currently an active member of IEEE and has also reviewed several journal articles. He has a distinguished record of research papers, with more than 18 international, Scopus, IEEE, and SCI papers. He is also in his bucket as a researcher with research areas in NLP, Machine Learning, and Image Processing.