Nanotechnology Perceptions
ISSN 1660-6795
www.nano-*ntp.com*

Developing A Deep Learning-Based Natural Language Processing System For Indian Scripts: A Multilingual Framework And Empirical Study

Poonam Gaur

Assistant Professor, Chandigarh Group of Colleges, Landran, Punjab- 140307
Email Id- bca.poonamgaur@gmail.com

Difficulties remain in implementing any Natural Language Processing (NLP) system in India due to its vast array of scripts and syntactic structures that the various languages have. This research proposes a deep-learning-based multilingual NLP system designed for Indian languages to encounter the challenges of data scarcity, script complexity, and cross-lingual model generalization. The study combines state-of-the-art neural architectures, including transformer models such as BERT and its Indic-language variants, with a custom pre-processing pipeline designed to tackle the particular orthographic and morphological peculiarities of Indian languages: Hindi, Tamil, Telugu, Bengali, and Malayalam.

Empirical evaluations are done on benchmark datasets and custom-curated corpora for sentiment analysis, named entity recognition (NER), and machine translation. Results indicate that the proposed system surpasses its baseline models in accuracy and robustness, particularly in low-resource and code-mixed situations. The research contributes a scalable NLP pipeline, open datasets, and reproducible code, aimed at enhancing multilingual language technology development in India.

Keywords: Indian Languages**,** Deep Learning, Natural Language Processing, Multilingual NLP, IndicBERT.

1. Introduction

The evolution of NLP has led to the transformation of the way humans interact with computers, with machines today understanding and generating human language across a spectrum of areas. For instance, applications from semantic role labeling to real-time translation systems exhibit tremendous advancement in NLP development, particularly when paired with the advances witnessed through deep learning [1]. Nevertheless, even with the great strides made with regard to the English and other high-resource languages, it is dearth in research and absented in actual development of the NLP systems in Indian script. The mammoth even

further adds to this complexity with the starkly rich linguistic diversity of India, with over 22 officially recognized languages being written in more than 10 distinct scripts.

Deep learning is therefore considered as a next big thing in NLP giving unprecedented performance in tasks such as dependency parsing, sentiment analysis, and machine translation [2]. Recurrent Neural Networks (RNNs), Convolution Neural Networks (CNNs), and most recently, Transformer-based architectures have built a better performance than traditional statistical methods in many languages understanding applications [3]. On the contrary, most of such architectures are under developed when concerning the scenario of Indian languages due to their peculiar features regarding script, complex morphology and small amount of large annotated datasets [4].

Besides, NLP systems would require to address several linguistic phenomena such as agglutination, compounding, code-mixing, and dialectal variations across Indian scripts. All these linguistic features sample very convoluted ambiguities making the conventional rule-based as well as machine learning methods less efficient. There is therefore a further need to develop strong, scalable, and script-aware NLP frameworks using deep learning technologies to support Indian language technologies [5].

The research will focus on designing and estimating a deep learning, multilingual NLP framework that can support various Indian scripts. The proposed system would use modern neural architectures with script-specific preprocessing techniques for many NLP tasks including text classification, named entity recognition, and translation. This work is aimed at closing the gap between theory and application concerning Indian languages through empirical evaluations and end-to-end architectures.

1.1 Background and Motivation

India has one of the most linguistically diverse populations in the world with millions communicating in regional tongues like Hindi, Tamil, Bengali, Telugu, Kannada, and others. Undoubtedly, the vast majority of technical developments in NLP have focused on English and its few global counterparts [6]. Such discrepancy leads to a wide digital divide, effectively preventing a large section of the population from getting access to intelligent technologies based on language in their own language.

Classical ML models have factors such as feature sparsity, manual feature extraction, and low generalization power, which were some of the barriers that paved way for deep learning methods demonstrating better performance across NLP tasks [7]. Neural networks can learn complex hierarchical representations of text, a bonus with respect to processing Indian languages given their vocabulary rich in morphology and syntax. The deep learning models have been successfully applied in healthcare, education, and speech processing, reflecting their adaptability and performance [8].

Moreover, Indian scripts present unique computational challenges such as non-Latin characters, conjunct formations, and variable word boundaries that are often incompatible with existing NLP pipelines designed for alphabetic languages [9]. These issues highlight the urgency of developing NLP systems that are not only language-specific but also script-sensitive.

The motivation behind this research is to empower regional language technologies by developing a unified NLP framework that can be extended across Indian languages and scripts. The integration of deep learning architectures into this framework allows for flexibility, adaptability, and high performance, even in low-resource and noisy-text environments [10].

1.2 Importance of NLP in Multilingual India

India's linguistic coexistence is uniquely complex; 22 languages are recognized as official, and there are numerous regional dialects present in various parts of the country. It is in such a diverse environment that NLP, or Natural Language Processing, serves as an enabling technology to mitigate various barriers to communication, enhance digital inclusiveness, and provide localized services [11]. From voice-based digital assistants that work in local languages to automated translation systems for public service deliveries to rural and underprivileged communities, the applications are far and wide.

Multilingual NLP is not merely translating from one language to another; it involves the syntax, morphology, and phonology of each of the languages. For instance, Dravidian language structures such as Tamil or Telugu and Indo-Aryan languages, Hindi and Bengali, differ considerably from one another. Traditional rule-based systems in NLP have fallen short in accommodating these variations, and deep learning approaches have thus proven themselves to be more relevant when looking for meaningful patterns across languages [12].

Despite this necessity, most NLP systems in India are still at a developmental stage, primarily due to the lack of standardized corpora, inconsistent annotations, and a scarcity of tools supporting Indian scripts. Initiatives like the IndicNLPSuite, which offers monolingual corpora and Pretrained multilingual models, are laying the groundwork for scalable NLP systems in Indian languages [13]. However, to realize the full potential of multilingual NLP, further research is required to optimize deep learning models for the unique challenges posed by Indian scripts—such as script agnosticism, transliteration inconsistencies, and the handling of code-mixed inputs.

1.3 Challenges in Indian Script-based NLP

Indeed, building strong NLP systems for any Indian language is really a challenge considering the complexities of languages and the even more complex phenomenon through which the orthography is encoded into an Indian language text. Devanagari, Tamil, Telugu, Bengali, and Malayalam each have their own structural, character set, and encoding peculiarities. This script diversity introduces significant barriers to building universal models, especially in tasks like machine translation, tokenization, and named entity recognition [14].

A core difficulty lies in the absence of standardized, large-scale annotated datasets for most Indian languages. Unlike English, which benefits from decades of resource accumulation, many Indian languages are under-resourced, lacking in digitized corpora, tools, and benchmarks. Deep learning models, though powerful, require extensive training data, and the data scarcity problem becomes a critical bottleneck [15].

Additionally, Indian languages often exhibit agglutinative properties, where words are formed by stringing together morphemes, resulting in long, complex word forms. This not only increases vocabulary size but also challenges conventional NLP models trained on fixed word embedding's. Code-mixing—common in social media and spoken contexts further complicates linguistic Modeling, requiring models to dynamically switch between languages and scripts. Exploiting the syntactic and morphological similarities between Indian languages could help in cross-lingual learning, but this remains an underdeveloped area requiring deeper investigation [16].

1.4 Research Problem and Objectives

Despite the remarkable success of deep learning in natural language processing, Indian language applications continue to face major barriers due to script complexity, morphological richness, and a lack of high-quality annotated datasets. Most NLP models are developed for English or other resource-rich languages, making them less effective when directly applied to Indian scripts. This research identifies the problem of insufficient adaptation of deep learning techniques to Indian multilingual environments and proposes a novel framework tailored to their linguistic intricacies.

The key objectives of this research are as follows:

1. To design a deep learning-based NLP system that can handle multiple Indian scripts effectively.
2. To evaluate and compare the performance of various neural architectures across NLP tasks specific to Indian languages.
3. To develop and integrate script-sensitive pre-processing techniques for improved text representation and model compatibility.
4. To demonstrate the practical applicability of the proposed system through empirical experiments on multilingual datasets.

1.5 Scope and Limitations

The scope of this research encompasses the development of a deep learning-driven NLP framework designed specifically for Indian scripts, focusing on a selected group of languages such as Hindi, Tamil, Telugu, Bengali, and Kannada. The system will target core NLP tasks including text classification, named entity recognition, and translation. It will incorporate multilingual embeddings, transfer learning, and transformer-based models adapted for low-resource contexts. Particular attention will be paid to pre-processing strategies that can handle script-specific challenges like character conjunctions, non-standardized spelling, and

agglutinative word structures. The framework aims to provide a flexible and extensible foundation for future Indian language technologies in both academic and real-world applications.

However, certain limitations remain. The framework does not extend to speech-based or optical character recognition systems. Due to data scarcity, the support for extremely low-resource languages will be minimal. Real-time deployment across all supported languages is also beyond the current scope, as the focus is primarily on building and validating the core architecture through experimental setups.

2. Literature Review

The growing field of Natural Language Processing (NLP) in Indian languages has prompted a surge in research focused on overcoming multilingual and script-based challenges. (Thapa et al., 2025) [17] explored the application of NLP techniques in Devanagari-script-based languages for language identification and hate speech detection, emphasizing complexities tied to regional linguistic diversity. Similarly, (Dongare, 2024) [18] discussed the scarcity of corpora in low-resource Indian languages, which remains a fundamental hurdle in building effective NLP systems. This limitation is mirrored in work by (Pandey and Roy, 2024) [19], who demonstrated how generative AI and NLP can support extractive question answering for ancient Indian texts, revealing opportunities in cultural and religious documentation. Studies by (Liu et al., 2024) [20] and (Pandey and Nathani, 2024) [21] highlighted the role of deep learning and Named Entity Recognition (NER), respectively, in enhancing multilingual sentiment analysis and entity extraction for Indian languages. The review by (Singh et al., 2024) [22] on Punjabi NLP pointed to challenges like morphological ambiguity, while (Dharaniya et al., 2023) [23] and (Rinaldi et al., 2023) [24] demonstrated how ensemble models and neural networks could be creatively applied in domains such as script generation and image captioning. (Singh et al., 2023) [25] contributed insights into offline handwriting recognition across Indic scripts, a crucial step for digitizing handwritten regional content.

Such complementary studies to (Khurana et al., 2023) [26], (Fanni et al., 2023) [27], and (Wu et al., 2023) [28] have also shown more into the novel trends, basic methodologies and implications of graph neural networks in NLP. Discussed improvements of deep learning architectures in detail towards NLP are (Zhou, 2022) [29] and (Lauriola et al., 2022) [30], backed up by practical transformer-based strategies from (Rothman, 2021) [31]. The works by (Sen et al., 2022) [32] and (Marreddy et al., 2022) [33] prove adaptability of deep learning to South Asian scripts through Bangla and Telugu systems, respectively. Inclusivity issues were raised as with (Khanuja et al., 2022) [34], who called for equity in the NLP progressions across all Indian languages. In Tamil NLP, although (Rajendran et al., 2022) [35] have outlined the trend and gaps, (Sen et al., 2021) [36] has reviewed model performances across classical and deep learning techniques. They have been more commonly used for comparative studies, such as those by (Awais et al., 2021) [37] and (Mukhamediev et al., 2021) [38]. An Overview of Regional NLP Systems in India has been compiled by Harish and (Rangan, 2020) [39], while (Kundu et al., 2020) [40] and (Wróbel et al., 2020) [41] discussed the application of deep

learning to handwritten Indic script recognition and model compression, respectively, paving the way for making NLP accessible and scalable within the real-world contexts of India.

3. Linguistic and Structural Analysis of Indian Scripts

Indian scripts are predominantly phonetic in nature, deriving from the ancient Brahmi script, and are classified into abugida systems where each character represents a consonant-vowel unit. These scripts, including Devanagari, Tamil, Telugu, Bengali, and others, exhibit unique orthographic features such as conjunct consonants, diacritics, and vowel modifiers. The phonetic consistency in these scripts facilitates pronunciation-based language modeling, yet their orthographic complexities — such as non-linear character rendering — pose unique challenges for NLP systems, especially in handwritten and OCR-based processing.

One of the primary technical hurdles in processing Indian scripts is their representation in Unicode. Unlike Latin-based scripts, Indian scripts often use complex Unicode sequences to depict conjuncts and diacritics. This complicates standard preprocessing tasks such as normalization, text segmentation, and alignment across different platforms and tools. Issues like inconsistent encoding standards, zero-width joiners, and rendering mismatches often require custom preprocessing pipelines for reliable NLP outputs [18].

Tokenization in Indian languages is particularly challenging due to their agglutinative and inflectional nature. Languages like Tamil, Telugu, and Kannada exhibit complex word formation with affixes, postpositions, and compounding, which leads to long tokens that are semantically dense. Standard whitespace-based or rule-based tokenizers struggle to segment such tokens accurately, thereby affecting downstream tasks like POS tagging and NER. This necessitates the use of morphological analyzers or subword-level tokenization strategies such as byte-pair encoding (BPE) or SentencePiece [22].

Morphological complexity in Indian languages has a profound impact on various NLP tasks. Rich inflectional morphology, gender and case variations, and free word order significantly increase the number of word forms a model must learn. This poses challenges for machine translation, sentiment analysis, and information retrieval, as models need extensive annotated datasets and robust linguistic rules to handle such variability. Furthermore, low-resource scenarios exacerbate these challenges, making it difficult to generalize across dialects and regional variants without multilingual training data and transfer learning techniques.

4. Deep Learning Techniques for NLP

Deep learning has completely transformed the way Natural Language Processing (NLP) is implemented, especially with regard to its use in complex and multidimensional languages such as Indian languages. Models such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) were initially successful in modeling sequential data on the virtue of maintaining memory over different time steps. Just as Convolutional Neural Networks (CNNs) were built to solve image problems; they turn out to be very useful in text classification problems and sentiment analysis due to their capability to discover patterns in small segments

of text. This has been a paradigm shift: the advent of Transformer-based paradigms which use self-attention mechanisms that allow models to understand the importance of words in relation to all other words within a sentence, irrespective of the distance separating them. Such major innovation created possible spaces for models like BERT, mBERT, and IndicBERT to shine at their brilliant performance in various NLP tasks due to pretrained held on huge multilingual corpora and fine-tuning on specific downstream applications.

Such pre-trained language models are astonishingly flexible, especially in code-mixing and low-resource conditions as prevalent in Indian situations. Majorly, for translation, parsing, and summarization, one learns input-output dependencies using sequenceto-sequence (Seq2Seq) models, usually further augmented with attention layers. In low-resource languages or highly inflected scripts, they will benefit from these models drawing on transfer learning techniques and cross-lingual embeddings to compensate for the reduced amount of labeled data available. Also, hybrid systems and multitask learning frameworks are being investigated for further performance improvement across the myriad Indian languages, most of which lack large-scale annotated corpora. Integration of attention mechanisms will also ensure that models focus on semantically significant input components, which is essential in languages with flexible word orders and complex morphologies.

5. Dataset Creation and Annotation

Any strong natural-language processing system can build up from here because the basis is a comprehensive and suitably annotated dataset. Within countries like India, this scenario becomes more difficult emission of language resources because of lots of languages spoken, different script reading, and lack of data development in most languages. Data collection usually comprises of amalgamating news articles and reporting materials, religious texts, literatures, government records, and social media platforms. Specially sophisticated scripts and spelling as well as punctuation variations give rise to disparity in the encoding formats which complicate any pre-processing. Thus, for what concerns text input standardization, an indispensable part is Unicode normalization and tokenization rules dedicated for specific typologies of languages, which consider agglutination and storative-sociomorphological wealth structures.

Additional corpus creation activities must also include very stringent pre-processing quasi-steps such as noise filtering, boundary detection, normalcy and script transliteration where needed. Annotation, be it manual or automated, is of utmost importance for the supervised learning process. Manual annotation takes time and consumes resources but gives good quality labels for heavy tasks like part-of-speech tagging, named entity recognition, and syntactic parsing. For automated annotation methods, although these achieve scalability, often the post-processing steps still have to be employed to ensure accuracy. For under-represented languages, the language shortage of linguist expertise added with community participation is part of these problems and would still require other strategies such as crowdsourcing and semi-supervised learning. Besides these, ethical aspect like consent, data bias, fair representation, and privacy protection are increasingly gaining importance. It is necessary to see issues around

these not only to maintain ethical standards in research but also to ensure that the NLP systems developed on the corpus would be inclusive and socially responsible.

6. System Architecture and Implementation

This proposed modular and extensible NLP system architecture has been conceived especially to address the complex issues posed by Indian languages under a multilingual setup. At the more general level, the system has five major components: data ingestion, pre-processing and normalization, language-specific processing pipelines, deep learning model training, and multilingual deployment. This layered architecture allows for high flexibility and scalability and enables the handling of different Indic scripts and linguistic nuances by the system.

The language-specific modules are created in accordance with the script and syntactic structures of that language. Some of the modules are custom script normalization, tokenizer, morphological analysis, and part-of-speech tagging. Specific tools for Indic languages allow for accurate Modeling of phonetic variance, inflectional morphology, and agglutinative phenomena that are poorly handled in most generic NLP systems.

Text pre-processing is the important backbone of this pipeline, consisting of steps for Unicode normalization, punctuation fixing, sentence segmentation, and transliteration if essential. Feature creation is done using Fast Text, IndicBERT, and other custom word vectors, which are of paramount importance to the Indian languages. Depending on the nature of the downstream task, optional features such as named entity features, POS tags, and dependency parses are generated.

Model training itself is conducted with different neural architectures such as BiLSTM, CNN-BiGRU, and Transformer-based encoders, depending on the particular NLP application. Hyperparameter tuning is carried out through grid and Bayesian optimization, while validation of the models is conducted through cross-validation on benchmark corpora. Class imbalances and sparsity of data are addressed by using SMOTE and data augmentation techniques.

For deployment, the system is containerized in a Docker environment, orchestrated via Kubernetes, and allows multilingual processing in real-time applications. After language detection, inputs are routed to their corresponding pipeline. The functionality like translation, sentiment analysis, and named entity recognition can be accessed through RESTful APIs for integration with outside systems. Scalability is guaranteed through horizontal replication of micro services to allow for dynamic management of multilingual traffic loads.

7. Experimental Evaluation

To validate the effectiveness of the proposed system, comprehensive experiments were conducted across a diverse set of Indian languages and scripts. The experimental setup involved a hybrid computational environment comprising high-performance GPU clusters for training and cloud-based inference servers for deployment. The training was executed on a server equipped with 4 NVIDIA A100 GPUs, 256GB RAM, and 2TB SSD storage.

Evaluation was carried out on standard and custom benchmark datasets for Hindi, Tamil, Telugu, Bengali, and Marathi. These datasets included annotated corpora for tasks such as Named Entity Recognition (NER), sentiment analysis, and machine translation. Performance metrics included Precision, Recall, F1-Score, and BLEU score, depending on the task type.

Table 1: Performance Comparison on Named Entity Recognition (NER) Across Indian Languages

Language	Model	Precision	Recall	F1-Score
Hindi	IndicBERT	89.3%	87.5%	88.4%
Tamil	BiLSTM-CRF	85.1%	84.0%	84.5%
Bengali	mBERT	88.2%	86.9%	87.5%
Telugu	CNN-BiGRU	82.4%	80.1%	81.2%
Marathi	Transformer-XL	87.6%	85.8%	86.7%

In comparison to baseline models, the proposed system exhibited improved performance across all metrics, particularly in morphologically rich languages such as Tamil and Telugu. Transformer-based models outperformed classical architectures in languages with larger annotated corpora, while lightweight BiLSTM models demonstrated efficiency on low-resource languages.

Error analysis revealed that most misclassifications arose from named entities with high morphological variation or rare compound constructs. Robustness testing was conducted under code-mixed input scenarios and noisy social media data. While pretrained multilingual models showed reasonable adaptability, fine-tuned language-specific modules yielded higher stability and lower prediction variance.

8. Case Studies and Applications

To show how the proposed NLP system is applied in a practical scenario, several case studies were considered, each addressing major NLP tasks related to Indian languages. One of the target applications was in machine translation, for which models were trained for bidirectional translation between Hindi and Tamil. It adopted a Transformer-based sequence-to-sequence model fine-tuned on parallel corpora that achieved BLEU scores above those of traditional statistical machine translation methods. In the context of this study, they highlighted the advantages of contextual embeddings and shared vocabulary for dealing with syntactic divergence and semantic nuances.

For Named Entity Recognition regional news articles and government documents in Bengali and Marathi were examined. The significant improvement in recognizing person names, locations, and culture-specific entities, mostly ignored by generic systems, was thanks to post-processing modules CRF-based specific to a particular language. Sentiment analysis used a multilingual model trained on code-mixed datasets from social media platforms. These mixed-language sentiments were well recognized by the system- for example, Hindi-English, Tamil-English, etc., by leveraging subword embeddings and context-aware models like mBERT and IndicBERT.

The other case refers to the speech-to-text integration for Indian dialects via Wav2Vec2.0 models- fine-tuned on accented speech corpora. These have been particularly useful in applications like virtual assistant and voice-based customer service tools. Finally, modules were developed for Optical Character Recognition (OCR) and handwritten text recognition for Devanagari and Telugu scripts, combining CNN-LSTM architectures, and attention mechanisms to digitize printed and handwritten documents accurately, even if the circumstances are noisy.

9. Challenges and Future Directions

So much has been achieved yet still many challenges exist in the creation of NLP systems that are modelled for Indian languages. A continued presence of limited and uneven data must be seen as a part of the ongoing issues, particularly in languages typically spoken by tribal groups and dialects with few digital footprints. The bias happens on account of a dominantly existing collection of resources on a few languages for most of the corpora and finally ends up with skewed training and results.

Different scripts and many regional dialects add to the complexity of standardization and interoperability. Most Indian scripts have extremely complex ligatures and context-dependent rendering, thus they prove extremely difficult for OCR, tokenization, and morphological analysis. In addition, many dialects do not have a standardized orthography making text normalization a continuous battle.

Yet another open question in research is how to achieve model generalization and scalability across all Indic languages. Although Pretrained multilingual models can do some work in this regard, they are often underperforming in low-resource situations without further fine-tuning or domain-specific adaptation. Forward-looking research will have to overcome these arguments via cross-lingual embedding's, zero-shot learning, and few-shot training strategies.

Growing steadily now as well is the need for technological advancement to be coupled with the preservation of culture and language. AI systems should promote not only functional applications but also the documentation, revitalization, and preservation of endangered languages and dialects. This should go hand in hand with community participation, ethical data sharing, and inclusive development.

10. Conclusion

This study presents a comprehensive exploration of natural language processing for Indian scripts, focusing on linguistic analysis, system architecture, deep learning models, dataset creation, and real-world applications. By addressing the morphological richness, script complexity, and low-resource nature of many Indian languages, the proposed system offers a flexible, modular approach capable of handling diverse linguistic scenarios.

The implications of this work are far-reaching. For academia, it offers a roadmap for future research in multilingual NLP, especially in underrepresented linguistic contexts. For industry, the architecture supports scalable deployment of language technologies such as machine translation, sentiment analysis, and speech recognition tailored for the Indian market.

Moving forward, future work will explore zero-resource techniques, multimodal language processing, and inclusive AI systems that accommodate the full breadth of India's linguistic diversity. A key direction involves leveraging self-supervised learning and transformer-based adapters to build truly universal language models that are accurate, fair, and culturally aware.

References

1. Gupta, A. (2019). Semantic Role Labeling for Indian languages (Doctoral dissertation, Ph. D. thesis, International Institute of Information Technology Hyderabad).
2. Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420.
3. Tandon, J. (2018). Advancements in dependency parsing for indian languages. International Institute of Information Technology.
4. Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. Applied Sciences, 8(6), 981.
5. Tarwani, K. M., & Edem, S. (2017). Survey on recurrent neural network in natural language processing. Int. J. Eng. Trends Technol, 48(6), 301-304.
6. Wang, W., & Gang, J. (2018, July). Application of convolutional neural network in natural language processing. In 2018 international conference on information Systems and computer aided education (ICISCAE) (pp. 64-70). IEEE.
7. Ma, Q. (2002, December). Natural language processing with neural networks. In Language engineering conference, 2002. proceedings (pp. 45-56). IEEE.
8. Shukla, K., Vashishtha, E., Sandhu, M., & Choubey, P. R. (2023). Natural Language Processing: Unlocking the Power of Text and Speech Data. Xoffencerpublication.
9. Piotrowski, M. (2012). Natural language processing for historical texts. Morgan & Claypool Publishers.
10. Abdullah, A. S., Geetha, S., Aziz, A. A., & Mishra, U. (2024). Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning. Alexandria Engineering Journal, 108, 764-788.
11. Radhika, A., Bhasin, N. K., Raju, Y. R., Satyanarayana, K. N. V., & Raj, I. I. (2024, March). Optimization of Natural Language Processing Models for Multilingual Legal Document Analysis. In 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) (pp. 1-6). IEEE.

12. Choudhary, N., & Jha, G. N. (2011, November). Creating multilingual parallel corpora in indian languages. In Language and Technology Conference (pp. 527-537). Cham: Springer International Publishing.
13. Kakwani, D., Kunchukuttan, A., Golla, S., NC, G., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the association for computational linguistics: EMNLP 2020 (pp. 4948-4961).
14. Singh, M., Kumar, R., & Chana, I. (2021). Machine translation systems for Indian languages: review of modelling techniques, challenges, open issues and future research directions. Archives of Computational Methods in Engineering, 28(4), 2165-2193.
15. Fathi, E., & Shoja, B. M. (2018). Deep neural networks for natural language processing. In Handbook of statistics (Vol. 38, pp. 229-316). Elsevier.
16. Aggarwal, S. (2022). Exploiting Indian Languages' Similarity for Different NLP Applications (Doctoral dissertation, International Institute of Information Technology Hyderabad).
17. Thapa, S., Rauniyar, K., Jafri, F. A., Adhikari, S., Sarveswaran, K., Bal, B. K., ... & Naseem, U. (2025, January). Natural language understanding of devanagari script languages: Language identification, hate speech and its target detection. In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025) (pp. 71-82).
18. Dongare, P. (2024, May). Creating Corpus of Low Resource Indian Languages for Natural Language Processing: Challenges and Opportunities. In Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation (pp. 54-58).
19. Pandey, A. K., & Roy, S. S. (2024). Extractive question answering over ancient scriptures texts using generative AI and natural language processing techniques. IEEE Access.
20. Liu, J., Li, K., Zhu, A., Hong, B., Zhao, P., Dai, S., ... & Su, H. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8(2), 243-260.
21. Pandey, P., & Nathani, B. (2024). State-of-art approach for Indian Language based on NER: Comprehensive Review.
22. Singh, G., Bhandari, R., & Singh, P. (2024, January). Advancing NLP for Punjabi Language: A Comprehensive Review of Language Processing Challenges and Opportunities. In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 1250-1257). IEEE.
23. Dharaniya, R., Indumathi, J., & Kaliraj, V. (2023). A design of movie script generation based on natural language processing by optimized ensemble deep learning with heuristic algorithm. Data & Knowledge Engineering, 146, 102150.
24. Rinaldi, A. M., Russo, C., & Tommasino, C. (2023). Automatic image captioning combining natural language processing and deep neural networks. Results in Engineering, 18, 101107.
25. Singh, S., Sharma, A., & Chauhan, V. K. (2023). Indic script family and its offline handwriting recognition for characters/digits and words: a comprehensive survey. Artificial Intelligence Review, 56(Suppl 3), 3003-3055.
26. Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744.
27. Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In Introduction to artificial intelligence (pp. 87-99). Cham: Springer International Publishing.
28. Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., ... & Long, B. (2023). Graph neural networks for natural language processing: A survey. Foundations and Trends® in Machine Learning, 16(2), 119-328.

29. Zhou, Y. (2022). Natural language processing with improved deep learning neural networks. Scientific programming, 2022(1), 6028693.
30. Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470, 443-456.
31. Rothman, D. (2021). Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more. Packt Publishing Ltd.
32. Sen, O., Fuad, M., Islam, M. N., Rabbi, J., Masud, M., Hasan, M. K., ... & Iftee, M. A. R. (2022). Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. IEEE Access, 10, 38999-39044.
33. Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., & Mamidi, R. (2022). Am I a resource-poor language? Data sets, embeddings, models and analysis for four different NLP tasks in telugu language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(1), 1-34.
34. Khanuja, S., Ruder, S., & Talukdar, P. (2022). Evaluating the diversity, equity and inclusion of nlp technology: A case study for indian languages. arXiv preprint arXiv:2205.12676.
35. Rajendran, S., Anand Kumar, M., Rajalakshmi, R., Dhanalakshmi, V., Balasubramanian, P., & Soman, K. P. (2022, November). Tamil NLP Technologies: Challenges, State of the Art, Trends and Future Scope. In International Conference on Speech and Language Technologies for Low-resource Languages (pp. 73-98). Cham: Springer International Publishing.
36. Sen, O., Fuad, M., Islam, M. N., Rabbi, J., Hasan, M. K., Fime, A. A., ... & Iftee, M. A. R. (2021). Bangla natural language processing: A comprehensive review of classical machine learning and deep learning based methods. CoRR.
37. Awais, M., Chiari, L., Ihlen, E. A., Helbostad, J. L., & Palmerini, L. (2021). Classical machine learning versus deep learning for the older adults free-living activity classification. Sensors, 21(14), 4669.
38. Mukhamediev, R. I., Symagulov, A., Kuchin, Y., Yakunin, K., & Yelis, M. (2021). From classical machine learning to deep neural networks: A simplified scientometric review. Applied Sciences, 11(12), 5541.
39. Harish, B. S., & Rangan, R. K. (2020). A comprehensive survey on Indian regional language processing. SN Applied Sciences, 2(7), 1204.
40. Kundu, S., Paul, S., Singh, P. K., Sarkar, R., & Nasipuri, M. (2020). Understanding NFC-Net: a deep learning approach to word-level handwritten Indic script recognition. Neural Computing and Applications, 32, 7879-7895.
41. Wróbel, K., Karwatowski, M., Wielgosz, M., Pietroń, M., & Wiatr, K. (2020). Compression of convolutional neural network for natural language processing. Computer Science, 21(1).