The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

40. Advancements and Challenges in Natural Language Processing for Low-Resource Languages: A Comprehensive Review

Anu Priya
Research Scholar, Jharkhand Rai University, Jharkhand, India
anupriya4658@gmail.com
Dr.Md. Irfan Alam
Associate Professor, Jharkhand Rai University, Jharkhand, India
irfan2.alam2@gmail.com

Abstract

This paper explores the improvements and demanding situations in Natural Language Processing (NLP) for Indian languages, specializing in textual content classification, part-of-speech tagging, computational linguistics applications, and the improvement of language processing equipment for lesser-studied languages. The paper synthesizes findings from more than one studies, highlighting key problems inclusive of linguistic diversity, aid limitations, and the effectiveness of numerous system gaining knowledge of algorithms. It additionally discusses the effect of new technology at the examine and alertness of NLP in Indian languages.

This review discusses advancements in processing Indian regional languages, which are resource-limited and complex due to rich morphology and diverse structures. It covers key tasks—machine translation, Named Entity Recognition, sentiment analysis, and POS tagging—using rule-based, statistical, and neural methods. Essential techniques like tokenization, transliteration, stemming, and lemmatization are outlined, emphasizing their role in natural language processing. Despite progress, challenges remain, including language diversity, translation ambiguity, and limited resources, like WordNets, for regional languages. The paper also mentions available datasets that aid research. Looking forward, it highlights the potential of neural and transfer learning to improve machine understanding and support seamless interactions in Maithili language. This ongoing research is essential as India's linguistic diversity grows in digital spaces, calling for better tools to facilitate multilingual communication and accessibility.

The role of Indian knowledge systems, rooted in traditional and indigenous languages, emerges as an essential factor in understanding cultural context and semantics in NLP applications. These systems preserve ancient philosophies, sciences, and cultural expressions, influencing tasks such as machine translation, Named Entity Recognition, sentiment analysis, and POS tagging.

Keywords
Natural Language Processing, Indian Languages, Text Classification, Part-of-Speech Tagging, Computational Linguistics, Machine Learning, Linguistic Diversity, Indian Knowledge System.

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

1. Introduction

1.1 Overview of Natural Language Processing

Natural Language Processing (NLP) is a pivotal department of synthetic intelligence that permits the interplay among computer systems and human languages via the improvement of algorithms and structures able to processing, analyzing, and producing herbal language textual content and speech. It incorporates a big range of duties which includes textual content classification, part-of-speech tagging, named entity recognition, gadget translation, and sentiment analysis, that are important for packages like facts retrieval, query answering, textual content summarization, and language generation(Kaur & Saini, 2015). Recent improvements in gadget mastering, in particular deep mastering strategies like neural networks, transformer models, and switch mastering, have notably improved the accuracy and performance of NLP structures. Despite those improvements, NLP faces demanding situations, mainly with the numerous and resource-scarce languages, which includes the ones in India, which own wealthy linguistic range and complicated scripts. Addressing those demanding situations calls for tackling problems associated with textual content encoding, linguistic range, and the shortage of categorized data, necessitating a aggregate of rule-primarily based totally methods, statistical models, and neural-primarily based totally strategies. This overview pursuits to discover those components comprehensively.

1.2 Importance of NLP for Indian Languages

Natural Language Processing (NLP) holds huge importance for Indian languages, given the linguistic range and richness gift withinside the country. India is domestic to 22 formally identified languages and loads of dialects, every with its particular syntax, semantics, and phonetic nuances. The significance of NLP on this context can't be overstated, because it helps the improvement of equipment and technology which can bridge conversation gaps, sell virtual inclusivity, and keep linguistic heritage.

NLP allows the advent of programs like computerized translation offerings, speech popularity systems, and text-to-speech converters tailor-made to Indian languages. This equipment can empower non-English-talking populations with the aid of using presenting get admission to virtual resources, academic content, and authorities' offerings of their local languages. For instance, an NLP-powered chatbot can help farmers with the aid of using presenting agricultural recommendation of their neighborhood dialect, thereby improving their productiveness and livelihoods.

1.3 Importance of Indian Knowledge System in Low-Resource languages

Deeply ingrained cultural and linguistic roots found in Indian languages are highlighted by the Indian Knowledge System, which helps improve contextual comprehension in NLP applications. Tools that incorporate IKS into NLP are better able to catch subtleties that are ingrained in the

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

cultural environment, producing more accurate interpretations and results in tasks like sentiment analysis and machine translation.

Semantic Depth and Philosophy: Classical literature and deep philosophical ideas that represent distinctive approaches to language and meaning structure are part of Indian traditional knowledge. By incorporating this element into NLP, models that comprehend the semantic complexity and metaphors inherent in Indian languages—found in ancient books such as the Vedas and Upanishads—can be developed.
Endangered Language Preservation: IKS can be very helpful in protecting endangered languages.

2. Preprocessing Techniques

Preprocessing for transliteration in Indian regional language NLP involves several essential techniques to prepare text for accurate processing and analysis. These techniques include text normalization, which ensures consistency by converting text to a uniform format, addressing variations in scripts, spellings, and diacritics. Tokenization is another critical step, where text is split into manageable units like words or phrases while respecting language-specific rules and word boundaries. For transliteration, script conversion is crucial, involving the mapping of characters from one script (e.g., Devanagari) to another (e.g., Roman script) using phonetic or rule-based approaches. Phonetic mapping algorithms help maintain the sound structure of words across different scripts, ensuring that pronunciation remains consistent. Additionally, noise removal helps eliminate unwanted symbols, punctuation, or non-language-specific characters, while language identification is often performed to handle code-switching cases common in Indian languages. Finally, stemming and lemmatization may be applied to reduce words to their base forms, aiding in uniformity and improving model performance. Together, these preprocessing techniques facilitate seamless transliteration, enabling better NLP outcomes in multilingual and diverse language processing environments.

3. Objectives of the Review

- Examine NLP`s Role in Digital Inclusivity: Highlight how NLP technology can bridge communication gaps and offer non-English-talking populations with get right of entry to virtual assets and services.
- Assess Technological Developments: Analyze the contemporary country of NLP gear and programs, which includes automatic translation services, speech reputation systems, and text-to-speech converters, tailor-made to Indian languages.
- Promote Linguistic Preservation: Discuss the significance of NLP in retaining endangered languages via digitization and the advent of linguistic databases, thereby preserving cultural heritage.

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

- Identify Research and Innovation Opportunities: Identify new studies avenues and capability improvements in NLP that cater to the numerous linguistic panorama of India.
- Evaluate Socio-Economic Impact: Assess the socio-financial advantages of NLP programs in improving productiveness and livelihoods, specifically in sectors like agriculture and education.

4. Challenges in Classifying Documents in Indian Languages

Natural Language Processing (NLP) complexity arises from the inherent intricacies of human language, which consist of syntax, semantics, pragmatics, and the widespread variability in linguistic expression(Devi & Purkayastha, 2018). The demanding situations multiply while addressing multilingual and low-useful resource languages, inclusive of the ones observed in India. Complexities consist of:

4.1 Ambiguity and Context: Words and sentences could have more than one meaning relying on context, requiring state-of-the-art algorithms to interpret accurately.

4.2 Morphological Variations: Indian languages show off wealthy morphological structures, along with large use of inflections and compound words, complicating tokenization and parsing.

4.3 Diverse Syntax and Grammar Rules: Each language has specific syntactic and grammatical rules, worrying custom fashions and large linguistic datasets.

4.4 Resource Scarcity: Many Indian languages lack large, annotated corpora essential for education powerful NLP fashions, proscribing the overall performance and applicability of NLP tools.

4.5 Multilingual Interactions: Code-switching and transliteration are not unusualplace in multilingual societies, including layers of complexity to language processing tasks.

The Indian Knowledge System (IKS) plays a crucial role in addressing the challenges faced by low-resourced languages in NLP. By emphasizing the preservation and documentation of indigenous languages, IKS supports efforts to maintain linguistic diversity and cultural heritage, aiding in the development of digital resources for lesser-documented languages. Additionally, IKS acknowledges the importance of oral traditions, providing a framework that can enhance NLP systems' ability to process speech and incorporate context from non-written sources. Classical linguistic works, such as *Panini's Ashtadhyayi*, offer valuable insights into the complex morphology and syntax of Indian languages, informing rule-based models and assisting in overcoming tokenization and parsing challenges. The inherent support for multilingualism within IKS can help NLP models better handle code-switching and transliteration, common in real-world Indian language use. Furthermore, the emphasis on semantic richness and contextual understanding in IKS can enhance NLP systems' ability to process language with greater depth

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

and accuracy. Finally, IKS can guide the creation of new linguistic resources through the digitization of traditional stories, folklore, and ancient texts, helping to alleviate the data scarcity issue that hinders the development of NLP tools for low-resourced languages.

5. Performance of Supervised Learning Algorithms

Supervised algorithms have proven sizable efficacy in numerous NLP obligations, especially in textual content class. Among these, Naive Bayes (NB) and Support Vector Machine (SVM) are usually used because of their robustness and simplicity(Harish, Bhuvaneswari, & Bhavani, 2020).

- **Naive Bayes (NB) for Text Classification**
  Naive Bayes is a probabilistic classifier primarily based totally on Bayes` theorem, with the idea of characteristic independence. Despite its simplicity, NB plays remarkably nicely in textual content class obligations which includes junk mail detection, sentiment analysis, and subject matter categorization. Its electricity lies in its capacity to deal with big vocabularies and high-dimensional records efficiently. However, the independence assumption can occasionally restrict its accuracy, especially in instances wherein phrase dependencies are sizable.

- **Support Vector Machine (SVM) for Text Classification**
  Support Vector Machine is a powerful, linear classifier acknowledged for its effectiveness in high-dimensional spaces. SVM works through locating the top-rated hyperplane that separates training withinside the characteristic space. In textual content class, SVM has been efficaciously implemented to obligations which includes e mail filtering, file categorization, and sentiment analysis. It plays nicely with each linearly separable and non-separable records through the usage of kernel hints to convert the records right into a higher-dimensional space. SVM's fundamental gain is its robustness and capacity to deal with noisy records, despite the fact that it is able to be computationally intensive, especially with big datasets.

6. Part-of-Speech Tagging for Indian Languages

Part-of-speech (POS) tagging is a essential challenge in NLP that entails assigning grammatical categories, consisting of nouns, verbs, and adjectives, to every phrase in a sentence (Priyadarshi & Saha, 2020). Developing POS taggers for Indian languages is essential because of the wealthy linguistic range and complicated morphology of those languages.

Language	Word	Pronunciation
English	anybody	e n i : b o d i :
Beganli	ভালবাসা	Bhālabāsā
Kannad	ಭಾರತ	Bhārata
Telugu	ప్రేమ	Prēma

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

Table 01:- Examples of pronunciation dictionary of few languages.

7. Case Study: Development of a Maithili POS Tagger

Maithili, a language spoken predominantly withinside the Indian states of Bihar and Jharkhand, affords precise demanding situations for POS tagging because of its syntactic and morphological characteristics. A case take a look at on growing a Maithili POS tagger illustrates the procedure and hurdles worried. The improvement worried amassing a complete corpus, annotating it with POS tags, and schooling a system studying version to carry out tagging. The task confronted demanding situations along with the shortage of annotated facts and the want for specialised linguistic know-how to as it should be tag Maithili words. Despite those demanding situations, the ensuing POS tagger substantially advanced the accuracy of language processing duties for Maithili.

7.1 NLP tools available in real world

- SpaCy:
  Overview: spaCy is an open-supply software program library for superior herbal language processing in Python. It`s designed for large-scale records extraction and herbal language understanding(Kumar & Kumar, 2022).
  Features: Tokenization, part-of-speech tagging, dependency parsing, named entity recognition, and textual content classification.
  Applications: Used in manufacturing structures for responsibilities which include textual content analytics, records extraction, and statistics mining.
- NLTK (Natural Language Toolkit):
  Overview: NLTK is a main platform for constructing Python applications to paintings with human language statistics(Kumar & Kumar, 2022).
  Features: Includes a huge variety of textual content processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
  Applications: Popular in academic settings and studies for coaching and studying approximately NLP.

8. NLP application in real world

The Indian Knowledge System (IKS) holds significant practical applications in real-world settings, particularly in addressing the challenges faced by low-resourced language processing. IKS emphasizes the preservation and promotion of cultural and linguistic diversity, which aligns with the goal of developing NLP systems that cater to underrepresented languages. In a country like India, which is home to numerous languages and dialects, most of which are low-resourced, integrating IKS into NLP can enhance the creation of language processing tools that respect and reflect cultural nuances.

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

One of the primary applications of IKS in this context is the documentation and digitization of indigenous languages. Many Indian languages have rich oral traditions that may not be well-documented in written form. IKS, with its focus on preserving oral and traditional knowledge, can guide efforts to develop speech recognition and synthesis models that incorporate context from oral literature, folklore, and traditional stories. This not only aids in the creation of linguistic corpora but also ensures that NLP models are more culturally aware and contextually accurate.

IKS also provides insights into linguistic frameworks through classical texts, such as *Panini's Ashtadhyayi*, which offer detailed analyses of syntax and grammar. These insights can inform NLP models for complex morphological and syntactic structures, helping overcome challenges in tokenization, part-of-speech tagging, and parsing that are common in low-resourced languages. By incorporating the grammatical rules and linguistic structures outlined in such texts, NLP applications can be adapted to handle the intricate characteristics of these languages more effectively.

Moreover, the inherent support for multilingualism within IKS helps address code-switching and transliteration, which are prevalent in multilingual Indian societies. This capability can be translated into practical NLP tools that better handle real-world language interactions. By leveraging these aspects, the Indian Knowledge System contributes to developing more inclusive, culturally sensitive, and effective NLP solutions for low-resourced languages.

9. Key Findings

- The main findings include the analysis of text classification works on Indian language content, the effectiveness of supervised learning algorithms for text classification tasks, and the need for further exploration in text classification for Indian languages.
- There is a strong need for language processing systems in Indian regional languages, as most customers prefer to interact in their native languages.
    - There is a lack of high-quality resources such as dictionaries and corpora for many Indian regional languages, which hinders the development of language processing systems.
    - Neural-based approaches, which have shown promising results in other NLP tasks, are yet to be extensively explored for Indian regional language processing.
    - The authors developed a CRF-based POS tagger for Maithili that achieved 82.67% accuracy using various features.
    - Incorporating neural word embeddings generated from a large raw corpus further improved the accuracy to 85.88%.

10. Conclusion

This paper gives an in depth survey of the cutting-edge repute and destiny instructions of NLP programs evolved for the Manipuri language, an underrepresented language in computational

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

linguistics. It analyzes numerous textual content category strategies and their utility to Indian languages, with a focal point at the overall performance of supervised studying algorithms including Naive Bayes, SVM, ANN, and N-gram models. The paper additionally highlights the demanding situations in processing Indian local languages, that are morphologically wealthy and feature obtained much less studies interest as compared to English. It emphasizes the want for extra sources and similarly exploration in numerous language processing tasks. Additionally, the paper describes the improvement of the primary Maithili part-of-speech tagger, which performed an accuracy of 85.88% the use of a CRF-primarily based totally version with neural phrase embeddings skilled on a manually annotated corpus. Finally, it gives an outline of the sphere of herbal language processing, detailing its history, key developments, programs, successes, and cutting-edge demanding situations, thereby imparting a complete knowledge of the sphere`s evolution and the unique wishes of Indian languages in NLP studies.

While significant progress has been made in text classification, machine translation, and part-of-speech tagging, many challenges remain, particularly due to the linguistic diversity, complex morphology, and resource scarcity associated with Indian languages. The Indian Knowledge System (IKS) offers valuable solutions to these challenges by emphasizing the preservation and documentation of indigenous languages, which supports efforts to maintain linguistic diversity and cultural heritage through the development of new digital resources. IKS's recognition of rich oral traditions can enhance NLP systems by incorporating context from non-written sources, thus improving speech recognition and synthesis models.

Moreover, classical works such as *Panini's Ashtadhyayi* provide insights into the intricate morphology and syntax of Indian languages, which can inform the development of rule-based and hybrid models to tackle issues like tokenization and syntactic parsing. The multilingual nature of IKS supports effective handling of code-switching and transliteration, reflecting real-world usage patterns in Indian contexts. The system's focus on semantic richness and deep contextual understanding can improve the precision and depth of NLP applications, including machine translation and named entity recognition. Lastly, IKS can guide the creation of comprehensive linguistic resources by digitizing traditional stories, folklore, and classical literature, which addresses the data scarcity problem faced by many low-resource languages. This holistic approach can significantly contribute to the advancement and inclusivity of NLP systems for India's diverse linguistic landscape.

References

1.Kaur, J., & Saini, J. R. (2015). A Study of Text Classification Natural Language Processing Algorithms for Indian Languages. VNSGU Journal of Science and Technology, 4(1), 162-167.

2.Devi, M. I., & Purkayastha, B. S. (2018). Advancements on NLP Applications for Manipuri Language. International Journal on Natural Language Computing (IJNLC), 7(5), 47-58.

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-7 Volume: IV (Special), October-December, 2025
Issue-28 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

3.Priyadarshi, A., & Saha, S. K. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. Computer Speech & Language, 62, 101054.
4.Harish, B., Bhuvaneswari, T., & Bhavani, R. (2020). An Extensive Review on Deep Learning Architectures. Journal of Artificial Intelligence, 16(4), 251-269.
5.Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.
6. Priyadarshi, A., Saha, S. K., & Singh, P. K. (2019). A Comprehensive Survey on Cross-Language Information Retrieval. Journal of Information Science Theory and Practice, 7(3), 12-28.
7. Kumar, A., & Kumar, D. (2022). Analysis of Natural Language Processing Techniques for Sentiment Analysis. Multimedia Tools and Applications, 81(19), 27233-27255.