JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Artificial Intelligence in Indian Languages: A Comprehensive Overview

Ramya S
Department of Computer Science, St Philomena College, Puttur Karnataka

Abstract

India offers a distinct opportunity and challenge for applications of artificial intelligence (AI) due to its linguistic diversity. This study offers a thorough assessment of the current level of artificial intelligence (AI) in Indian languages, addressing the main issues, potential fixes, and future possibilities in this quickly developing field. There are several complications, ranging from low-resource languages to linguistic variety and script differences. We explore the subtleties of AI methods and tools used to support language generation, comprehension, and cultural relevance. We also talk about how AI will affect education, information accessibility, and Indian society.

Keywords: Artificial Intelligence, Indian Languages, Natural Language Processing, Machine Learning, Linguistic Diversity.

Introduction

Significant progress is being made in supporting and improving Indian languages with artificial intelligence (AI). AI is being used in India, a linguistically diverse nation with many different languages and dialects, to increase accessibility, encourage digital inclusion, and overcome language barriers. The following are some applications of AI in Indian languages:1. Major Languages: The main languages spoken in India are Hindi, English, Bengali, Telugu, Marathi, and Tamil. Due to their broad usage, these languages are typically used in the development of AI applications and content.

Linguistic Diversity in India:

Given the variety of Indian languages, artificial intelligence (AI) in Indian languages is a significant and developing field. India is a nation with hundreds of languages spoken in its different areas, creating a rich linguistic tapestry. India's linguistic diversity presents benefits as well as obstacles for the advancement and use of AI technologies.

When talking about AI in Indian languages and the variety of Indian languages, keep the following aspects in mind:

1. Script Diversity: Devanagari, Latin, Bengali, Telugu, Tamil, and other scripts are among those used in India. It's difficult to create AI models that can manage different scripts.

2. Linguistic Variants: There are numerous regional dialects and linguistic variances in Indian languages. For AI models to be effective nationwide, these variations must be taken into consideration.

3. Low-Resource Languages: A large number of India's lesser-known languages are low-resource languages, which means that there is a dearth of digital information and data for AI model training. The creation of AI applications in these languages is hampered as a result.

4. NLP Challenges: Because of India's linguistic diversity, natural language processing (NLP), a crucial component of artificial intelligence, is a challenging undertaking. Certain NLP models and resources are needed to comprehend and process the various languages and scripts.

5. Language Preservation: By developing digital tools and resources for endangered languages, AI can help promote and preserve them. These languages are being documented and digitalized.

6. Language Technologies: To allow AI to serve a more diversified linguistic audience, the Indian government and a number of organizations are working on language technologies. This covers the advancement of text-to-voice systems for many languages, machine translation, and speech recognition.

7. Content Localization: Applications and content must be translated into several languages in order to reach a larger audience. Through automation of translation and adaption, AI can help with the localization process.

Data Collection and Preprocessing

Preprocessing and data collection are essential phases in the creation of AI models for Indian languages. The linguistic diversity of India makes having high-quality data in a variety of languages and scripts imperative. The following are the main things to think about when gathering and preparing data:

1. Data Collection:

• **Diverse Sources**: Gather information from a variety of sources, such as books, websites, news stories, social media, and more. This aids in capturing various content kinds and writing styles.

• **Crowdsourcing:** Involve the public, particularly for languages with limited resources. Gathering text and audio data from native speakers can be aided via crowdsourcing.

• **Government Resources**: Publications from the government, texts in the public domain, and official papers can be excellent sources of information for a variety of Indian languages.

• **Domain-Specific Information**: Gather information specific to a given domain based on the application. For example, compile medical books in the target languages if you're developing an AI for medicine.

• **Translation:** Convert data between Indian languages and major languages like English. For machine translation jobs, parallel corpora—source text and translations—are quite helpful.

• **Speech Data:** Gather audio data in a variety of Indian languages for speech-related AI applications, such as text-to-speech synthesis and speech recognition.

2. Data Preprocessing

• **Text Normalization**: Address problems with ligatures, diacritical marks, and various script variations in order to normalize the text data. For consistent data processing, this is essential.

• **Tokenization:** Using tokenizers tailored to a certain language, divide text into tokens (words or sub words). Tokenization methods may vary depending on the language.

• **Character Encoding:** Take into account the different scripts used in India when encoding text data. When encoding multilingual text, UTF-8 is frequently utilized.

• **Cleaning and Filtering:** Get rid of extraneous material, HTML tags, and non-textual content. Remove entries that are duplicates.

• **Part-of-Speech Tagging:** Use part-of-speech tagging to classify words with their grammatical categories if your AI application requires understanding language structure.

• **Lemmatization and stemming:** To streamline text analysis and enhance models, reduce inflected or derived words to their base or root form.

• **Lemmatization and stemming:** To streamline text analysis and enhance model performance, reduce inflected or derived words to their base or root form.

• **Language Identification:** When working with multilingual datasets, it is important to use language identification techniques to label text data with the appropriate language.

Language Identification

Natural language processing (NLP) has advanced significantly thanks to artificial intelligence (AI), and this includes the identification and recognition of Indian languages. Given that there are hundreds of languages and dialects spoken in India, language identification is crucial in this setting. The following describes how AI is applied to the identification of Indian language:

1. Character-Level Analysis: In order to discern between several Indian languages, AI algorithms are able to examine the text at the character level. For languages with similar scripts, like Hindi, Marathi, and Sanskrit, this method works well. Devanagari is one such script.

2. Word-Level Analysis: Language identification can also be done at the word level, or word-level analysis. Artificial intelligence models are capable of identifying the unique word structures and vocabularies found in various Indian languages.

3. Phonetic Analysis: Despite having similar characters, certain Indian languages have different phonemes. These phonetic differences allow AI to be trained to distinguish between different languages.

4. Language Models: For certain language identification tasks, pretrained language models like BERT, GPT-3, and their Indian language variations can be adjusted. These models are capable of analysing text and predicting the language in which it was written.

5. Dialect Recognition: Different dialects of the same language are common. It is possible to train artificial intelligence (AI) to distinguish between dialects and the main language. For regional communication and localized content, this is essential.

6. Statistical Methods: To determine the language of a document, AI models can make use of statistical features such as n-grams, which are word or character sequences, and language-specific patterns.

7. Deep Learning and Machine Learning: Using text inputs, deep learning algorithms such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) can be trained to recognize Indian languages. Models of supervised machine learning can also be applied here.

Part-of-Speech Tagging and Named Entity Recognition

Part-of-Speech Tagging (POS) and Named Entity Recognition (NER) are two examples of artificial intelligence (AI) applications in Indian languages that have gained importance because they make it possible to create natural language processing systems that are sensitive to the country's linguistic diversity. An outline of these two language processing assignments with relation to Indian languages is provided below:

Part-of-Speech Tagging (POS)

• **Definition:** Labelling every word in a document with the appropriate grammatical category (part of speech), such as noun, verb, adjective, etc., is known as POS tagging.

• **Importance in Indian Languages:** For numerous NLP tasks, such as machine translation, sentiment analysis, and information retrieval in Indian languages, POS tagging is necessary for comprehending the syntactic structure of sentences.

• **Difficulties:** The morphological and syntactic patterns of Indian languages vary greatly, necessitating the customization of POS tagging models for every language, dialect, and script.

• **Approaches:** Deep learning, rule-based systems, and hybrid approaches are being used by numerous researchers and organizations to create POS tagging models and datasets for Indian languages. Examples of projects that have helped in this area are the Universal Dependencies project and the Indian Language Toolkit (ILMT).

Named Entity Recognition (NER):

• **Definition:** Name recognition and enumeration (NER) is the process of locating and classifying specific entities in text, including names of individuals, locations, businesses, dates, and more.

• **Significance in Indian Languages:** NER is essential for activities like question-answering, information extraction, and other language comprehension. Because of the wide variety of names and entities, problems with transliteration, and different word boundaries, it can be difficult to recognize named entities in Indian languages.

• **Difficulties:** One major difficulty is the unavailability of extensive labelled datasets for NER in Indian languages, particularly for low-resource languages. NER is further complicated by the fact that complex noun phrases and compound words are common in Indian languages.

•**Approaches:** Conditional random fields (CRF), deep learning, and transfer learning are some of the methods that researchers have been employing to create NER models for Indian languages. The goal of initiatives such as the Indo NLP project has been to develop NER models and datasets for Indian languages.

Machine Translation

Indian language machine translation has advanced significantly thanks to artificial intelligence (AI). Machine translation is the process of translating text or speech automatically between languages using artificial intelligence (AI) models and algorithms. In India, where multiple languages and dialects are spoken, machine translation can be extremely helpful in removing barriers between languages and promoting communication.

Here are some key points related to AI and machine translation in Indian languages:

1. Richness of Languages: With hundreds of regional dialects and over 22 officially recognized languages, India is renowned for its linguistic richness. In order to improve communication, machine translation is being used in India to translate between various languages and dialects.

2. Difficulties: Indian languages pose particular difficulties for machine translation because of their intricate grammar, differing scripts, and the dearth of large digital text corpora for numerous languages.

3. Government Initiatives: The development of AI and machine translation technologies for Indian languages has garnered the active interest of the Indian government. To encourage research and development in this field, programs like "Technology Development for Indian Languages" (TDIL) have been introduced.

4. Research and Development: In India, machine translation models for Indian languages are being actively researched and developed by both public and private institutions. This entails building translation software, training AI models, and producing datasets tailored to individual languages.

5. Commercial Solutions: A number of Indian startups and businesses provide machine translation services for Indian languages that are available for purchase. Numerous industries, including as e-commerce, customer service, and content localization, use these tools.

6. Open-Source Projects: To enhance the quality of translations for Indian languages, there are open-source machine translation projects. Researchers and developers are encouraged to contribute under the community-driven development paradigm.

Speech Recognition and Generation:

Although there has been a lot of progress in translating Indian languages, machine translation is still in its early stages of development. There is constant effort to increase translation coverage, accuracy, and usability. Machine translation in Indian languages is probably going to becoming much easier to use and more efficient as AI and NLP technologies develop. This will help with cross-lingual communication and content localization.

Speech Recognition:

1. Language Diversity: Given the diversity of languages and dialects spoken in India, speech detection is a difficult task. AI systems must be able to recognize and comprehend a variety of language variances, including accents and pronunciations.

2. Voice Assistants: AI-driven voice assistants that support Indian languages include Apple's Siri, Amazon Alexa, Google Assistant, and Apple Assistant. These devices allow users to communicate in multiple languages, including Bengali, Tamil, Hindi, and others.

3. Voice Search: A lot of mobile apps and search engines now provide voice search options in Indian languages. Users can ask questions in the language of their choice, and the AI system will respond with relevant results.

4. Transcription Services: AI-powered transcription services are being used to turn audio files into text for a range of purposes, including interview transcription and the creation of video subtitles.

5. Accessibility: Real-time transcription for people with hearing problems is made possible by speech recognition technology. It makes communication easier by translating spoken words into text.

Speech Generation:

1. Text-to-Speech (TTS): Text input is transformed into spoken words via TTS technology. TTS systems for Indian languages have advanced, producing more expressive and natural speech creation.

2. Localized material: Audiobooks, automated voice responses in customer support, and language-learning apps are just a few examples of the types of localized material that may be produced thanks to TTS technology.

3. Voice Cloning: Brands and content producers seeking for distinctive voice branding in Indian languages may find it useful because certain AI systems enable users to generate personalized voices for TTS.

4. Conversational AI: Chatbots and virtual assistants may react in Indian languages with natural-sounding voices thanks to the integration of TTS into these systems.

5. Content Accessibility: written-to-speech software (TTS) makes digital material more readable by providing audio versions of written content for people with visual impairments.

6. Voice Model Development: Ongoing research and development endeavours aim to enhance the expressiveness and Caliber of voices produced by TTS systems, rendering them more akin to human voices.

7. Integration with Translation: To offer a comprehensive solution for translating and producing speech in several languages, TTS can be integrated with machine translation.

It is anticipated that AI-driven voice generation and recognition in Indian languages will revolutionize digital communication, accessibility, content development, and user experience in general. We should expect much more precise and natural speech generation and recognition for Indian languages as technology develops.

Sentiment Analysis and Emotion Detection

For a variety of languages, including Indian languages, artificial intelligence (AI) has made substantial progress in the area of natural language processing, including sentiment analysis and emotion recognition. These technologies can be used for a multitude of tasks, such as analysing client feedback and monitoring social media. An overview of the application of AI to sentiment analysis and emotion identification in Indian languages is provided below:

1. Data Collection and Language Processing: AI systems require access to sizable datasets in Indian languages in order to do sentiment analysis and emotion identification in those languages. A crucial first step is the gathering, preprocessing, and annotation of such data. This data is getting easier to get as more Indian languages are available online.

2. NLP Models and Algorithms: Artificial intelligence (AI) models, such as transformer-based models like BERT and recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been adjusted and refined for Indian languages. The purpose of these models is to comprehend the context and semantics of text in Indian languages.

3. Sentiment analysis: Also referred to as opinion mining, sentiment analysis is figuring out if a passage of text conveys a favourable, negative, or neutral emotion. This can be used for news stories, consumer evaluations, social media updates, and more in Indian languages. Businesses can use sentiment analysis technologies to measure customer satisfaction and public opinion.

4. Emotion Detection: This feature goes one step further by classifying text according to distinct emotions such as joy, sorrow, rage, etc. Understanding user responses and emotional responses in text data can be aided by this. For every Indian language, emotion detection models must be trained using labelled data.

5. Difficulties: The linguistic diversity of Indian languages creates particular difficulties. India recognizes hundreds of dialects in addition to 22 official languages. Developing reliable models for each of these dialects and languages can need a lot of resources. Additionally, there might not be as much data available to train AI models in some Indian languages because they are low-resource languages.

6. Tools and Libraries: For NLP jobs in Indian languages, a variety of tools and libraries are available. For instance, libraries like Hugging Face Transformers and spaCy as well as tools like Indic NLP have helped to design AI applications for Indian languages.

7. Use Cases: Applications such as social media monitoring for political campaigns or brand reputation management make use of AI-powered sentiment analysis and emotion recognition in Indian languages.

• Analysis of consumer feedback for companies.

• Examining news stories for societal sentiment.

• Emotional and sentiment analysis in patient evaluations and discussion boards pertaining to healthcare.

8. Prospective Advancements: The field of AI research and development for Indian languages is still expanding. Models and tools for Indian languages will improve in accuracy and versatility as more data becomes accessible.

Chatbots and Virtual Assistants

Recent years have seen a notable increase in the use of artificial intelligence (AI) in Indian languages, including chatbots and virtual assistants. India is a multilingual nation with many regional languages and dialects, therefore developing AI systems that can converse in Indian languages is crucial to expanding one's customer base and offering more inclusive services. In the framework of chatbots and virtual assistants, the following are some significant advancements and uses of AI in Indian languages:

1. Language Support: A variety of Indian languages, including Hindi, Bengali, Tamil, Telugu, Marathi, and more, are being supported by AI-powered chatbots and virtual assistants. Users can now communicate with these devices in the language of their choice thanks to this.

2. Customer assistance: To offer customer care in Indian languages, some Indian enterprises have integrated chatbots powered by artificial intelligence. The ability of these chatbots to respond to frequently asked questions, handle problems, and help users in their preferred language improves the user experience.

3. Voice Assistants: Multiple Indian languages are now available for voice-activated virtual assistants like Google Assistant and Amazon's Alexa, making them more accessible to users who might not speak English well.

4. Language Translation: To facilitate cross-linguistic communication, chatbots and virtual assistants are being equipped with AI-based language translation capabilities. For instance, a chatbot might reply in Hindi after translating a user's question from Hindi to English.

5. Content Generation: News stories, reports, and marketing materials can all be produced in Indian languages with the help of AI. For media firms and content creators trying to reach a larger audience, this is very helpful.

6. Educational Applications: To assist students in learning Indian languages, enhance their language proficiency, and offer explanations in the user's native tongue, educational apps and platforms employ AI chatbots and virtual assistants.

7. Healthcare: To make healthcare services more accessible to a varied population, virtual assistants are being used in the healthcare industry to deliver medical information and help in Indian languages.

All things considered, the varied and multilingual population of India might greatly benefit from the accessibility, user experience, and convenience that the integration of AI in Indian languages for chatbots and virtual assistants offers.

Dialect and Regional Variations

India's rich linguistic landscape—there are hundreds of languages and thousands of dialects—presents a unique mix of potential and problems for artificial intelligence (AI) in Indian languages. An outline of the difficulties and differences in AI for Indian languages is provided below:

1. Linguistic Diversity: With over 1,600 spoken languages and dialects and 22 officially recognized languages, India is linguistically diverse. Because of this diversity, it is difficult for AI systems to comprehend and process different languages and dialects.

2. Script Variations: India has several scripts, such as Tamil, Telugu, Bengali, Gujarati, and Devanagari (used for Hindi, Marathi, and other languages). These scripts must be reliably recognized and processed by AI systems.

3. Differences in Dialect: Even within the same language, Indian dialects can differ greatly from one another. For example, there are several regional dialects of Hindi, such as Rajasthani, Bhojpuri, and Haryanvi. For AI systems to produce and comprehend text in various dialects, they must be trained.

4. Low-Resource Languages: There is a dearth of digital resources and information for many Indian languages. Because of this, creating reliable AI models for these languages is difficult. It's possible that poorly annotated data from low-resource languages is used to train AI algorithms.

5. Code-Switching: Indians frequently use code-switching, or combining different languages in everyday conversation. AI systems must be capable of managing this code-switching.

6. Speech Recognition: Due to the distinctive phonetic features of Indian languages, developing precise speech recognition systems is a major difficulty. This process is made more difficult by the variety of accents and dialects found in spoken Indian languages.

7. Machine Translation: Translating between Indian languages and English or other languages is tough due to linguistic variances. AI systems must be specifically designed to appropriately manage these subtleties.

Researchers and developers in AI are striving to create language models tailored to Indian languages in order to address these issues. Furthermore, attempts are being made to enhance speech recognition and machine translation skills for Indian languages, as well as to create databases and resources for low-resource languages and dialects. Collaborations between the public and private sectors and government initiatives are also advancing AI technologies in Indian languages.

Low-Resource Languages

The field of artificial intelligence (AI) in low-resource Indian languages is expanding and offers potential as well as obstacles. Low-resource languages are those that have a restricted availability of digital resources, including

speech, text, and natural language processing (NLP) technologies. India is a multilingual nation with many different languages spoken there; but, because to the dearth of digital material and research funding, many of these languages are regarded as low-resource languages. Here are some important things to think about while using AI in Indian languages with limited resources:

Challenges

• **Data Availability**: Insufficient text and speech data for AI model training is frequently present in low-resource languages. The creation of language models and NLP tools is hampered by this lack of data.

• **Lack of Language Expertise:** The instruments and linguistic experts required for language annotation are lacking in several low-resource Indian languages.

• **Script Diversity:** Low-resource languages with non-Latin scripts are more difficult to interpret and comprehend due to India's numerous writing systems and scripts.

Opportunities:

• **NLP Research:** In an effort to provide NLP tools in low-resource languages, AI researchers are concentrating more on these languages. This includes gathering information, developing language models, and developing low-resource language applications.

• **Language Preservation**: AI can help with the documentation and preservation of endangered languages. Communities can preserve their linguistic and cultural legacy with the aid of digital preservation.

• **Local Language Access:** AI can improve local language access to information and services, increasing the inclusivity and accessibility of digital material for a larger demographic.

Initiatives

• **Government Support:** To encourage the development and use of Indian languages in IT and other applications, the Indian government has launched programs like the Technology Development for Indian Languages (TDIL).

• **Academic and Industry Collaboration:** The creation of AI resources in low-resource languages depends on cooperation between academic institutions, language specialists, and tech firms.

• **Open-Source Tools:** Research and development can be aided by the creation of open-source NLP tools and resources for languages with limited resources.

Community Involvement

Successful AI programs in low-resource languages require active engagement with local people, language enthusiasts, and linguists. They can work together to collect and annotate data and offer insightful advice.

Artificial Intelligence in low-resource Indian languages is a developing topic that needs coordinated efforts from multiple stakeholders to overcome obstacles and capitalize on the potential advantages of increasing technological inclusivity for all Indian linguistic minorities.

Implications for Indian Society

Indian society stands to gain much from the application of artificial intelligence (AI), especially in the context of Indian languages. The following are some effects of AI in Indian languages on many facets of Indian society:

1. Information Accessibility: AI can help create digital content in Indian languages, increasing the number of people who can access information. For a multilingual nation like India with a multitude of regional languages and dialects, this is essential.

2. Language Preservation: AI can support the revival and preservation of Indian languages that are in danger of extinction. Tools for translation and language modelling can help with language promotion and documentation.

3. Education: AI-driven tools and applications for language acquisition can increase accessibility and engagement in the classroom, especially for people who speak and comprehend Indian languages.

4. Government Services: By offering chatbots for questions, automated language translation, and more user-friendly interfaces for citizens who feel more at ease in their native tongues, artificial intelligence (AI) can enhance the effectiveness and accessibility of government services.

5. Business and Commerce: When e-commerce, digital marketing, and customer service are customized for local languages, they can work better. This enhances user experience and increases market potential.

6. Healthcare: AI can help healthcare providers and patients who speak different Indian languages communicate more effectively. Translation services and medical chatbots can be quite helpful in enhancing patient results.

But there are also drawbacks and hazards, such as concerns about security, privacy of data, loss of employment, and biases in AI algorithms. It's critical that Indian society create and put into place laws and policies that address these issues, guaranteeing that the positive effects of AI in Indian languages are maximized while reducing any potential negative effects.

Conclusion

With significant ramifications for India's linguistic and cultural diversity, artificial intelligence in Indian languages is a vibrant and developing topic. The creation of reliable and contextually aware AI models for Indian languages would be crucial as AI technology develops. This paper presents a comprehensive overview of AI in Indian languages as it is today, emphasizing both the obstacles and the intriguing future prospects.

References:

1. A comprehensive survey for automatic speech recognition of Indian languages, A Singh, V Kadyan, M Kumar, N Bassan - Artificial Intelligence Review, 2020 – Springer.
2. A comprehensive survey on Indian regional language processing, BS Harish, RK Rangan - SN Applied Sciences, 2020 – Springer
3. A comprehensive survey on machine translation for English, Hindi and Sanskrit languages, Sitender, S Bawa, M Kumar, Sangeeta - Journal of Ambient Intelligence, 2021 – Springer
4. Study of automatic text summarization approaches in different languages ,Y Kumar, K Kaur, S Kaur - Artificial Intelligence Review, 2021 – Springer

5. A journey of Indian languages over sentiment analysis: a systematic review, S Rani, P Kumar - Artificial Intelligence Review, 2019 – Springer

6. Machine translation systems for Indian languages: review of modelling techniques, challenges, open issues and future research directions, M Singh, R Kumar, I Chana - Archives of Computational Methods in, 2021 – Springer

7. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR), J Memon, M Sami, RA Khan, M Uddin - IEEE Access, 2020 ieeexplore.ieee.org

8. A survey on artificial intelligence in Chinese sign language recognition, X Jiang, SC Satapathy, L Yang, SH Wang… - Arabian Journal for, 2020 – Springer
9. A Comprehensive Overview of World Mapping Analysis Research Trends on Impact of Artificial Intelligence in Tourism from 2000 to 2022: A Literature Review, MF Ab Rashid, MAA Aziz - researchgate.net

10. Artificial intelligence and the public sector—applications and challenges BW Wirtz, JC Weyerer, C Geyer - International Journal of Public, 2019 - Taylor & Francis