IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Exploration Into How Successful AI Has Been In Aiding Hindi As A Second Language Learners

Ashish Jaiswal, Arjun Nair, Aarohi Mehta
Jain (Deemed-to-be University)
Bengaluru, Karnataka, India

Abstract: Artificial Intelligence (AI) has significantly transformed second-language acquisition by providing personalized, interactive, and adaptive learning experiences. This research examines AI's role in Hindi as a Second Language (HSL) learning, emphasizing key technologies such as Natural Language Processing (NLP), Machine Learning (ML), and Speech Recognition. AI-driven platforms, including Duolingo, Google Translate, ChatGPT, and Language Curry, enhance pronunciation, grammar comprehension, vocabulary acquisition, and fluency. Despite these advancements, AI-based language learning tools encounter challenges such as contextual inaccuracies, cultural nuances, idiomatic expressions, and regional dialect variations. This study evaluates AI's effectiveness in overcoming these limitations and highlights areas requiring improvement, including the development of more sophisticated NLP models, AI-human hybrid learning approaches, and expanded Hindi-language datasets for ML training. The findings suggest that while AI enhances Hindi language learning, further refinements are needed to address linguistic complexities effectively. The paper concludes with recommendations for future advancements in AI-driven language education, aiming to make Hindi learning more intuitive and effective.

Index Terms - Artificial Intelligence (AI); Hindi Language Learning; Natural Language Processing (NLP); Machine Learning (ML); Speech Recognition; Adaptive Learning; AI-driven Education.

I. INTRODUCTION

The increasing demand for Hindi as a second language (HSL) is driven by globalization, migration, and academic or professional needs. As more individuals seek to acquire Hindi proficiency, the necessity for effective, accessible, and adaptable learning methods has grown. Traditional approaches, including textbooks, classroom instruction, and immersion techniques, have long been effective. However, these methods often lack flexibility and real-time interaction, which modern learners prefer. Artificial Intelligence (AI) has emerged as a game-changer in second-language acquisition, introducing innovative tools to enhance the learning experience. Speech recognition technology assists learners by providing instant feedback on pronunciation. Natural language processing (NLP)-based grammar correction tools help refine writing accuracy. Machine learning models generate personalized lesson plans tailored to individual learning styles and progress. AI-driven conversational assistants create interactive dialogue experiences, helping learners practice real-world conversations with greater confidence. Despite these advancements, AI-driven language learning presents several challenges. AI struggles to grasp complex linguistic contexts, making it difficult to differentiate between formal and informal speech variations. Additionally, cultural nuances essential to language learning are often overlooked. Human interaction, a key element in traditional learning methods, such as mentorship and peer discussions, remains difficult for AI to replicate.

To address these limitations, researchers are developing more sophisticated AI models. Context-aware NLP systems are being refined to enhance AI's ability to process and interpret conversations in Hindi with cultural relevance. Multimodal learning approaches combine speech, text, and visual aids to create an immersive experience. Reinforcement learning techniques are being explored to help AI adapt dynamically to learners' evolving proficiency levels. Additionally, sentiment analysis and emotional computing are being integrated into AI-based tutoring systems to assess learner engagement, motivation, and comprehension levels, making AI-driven learning more intuitive. This study examines the effectiveness of AI in learning Hindi as a second language by assessing its advantages and identifying areas for improvement. It also explores potential strategies to overcome existing limitations, such as improving contextual understanding, incorporating cultural elements, and fostering a more human-like learning experience. The ultimate goal is to bridge the gap between conventional and AI-based approaches, offering learners a comprehensive and effective language learning solution.

II. LITERATURE REVIEW

AI-driven language learning has been widely explored, particularly for languages like English, Spanish, and Mandarin. Research by Kumar (2023) suggests that AI-based learning tools can enhance language retention by up to 40% compared to traditional methods. However, Hindi poses unique challenges due to its distinct phonetics, sentence structures, and semantic complexities, making it difficult for AI models to achieve the same level of accuracy as in other languages [1]. Natural Language Processing (NLP) plays a crucial role in enabling AI-powered tools to analyze Hindi grammar, phonetics, and sentence structures. However, the language presents several obstacles, including complex morphology and verb conjugations, a lack of well-annotated datasets for training machine learning models, and multiple regional dialects that AI often struggles to differentiate. Studies by Sharma (2022) indicate that NLP models designed specifically for Hindi require significantly larger datasets and advanced context-aware algorithms to improve accuracy and efficiency [2]. Additionally, Singh and Gupta (2022) highlight the unique challenges in processing Hindi text due to its syntactic complexity and limited linguistic resources, emphasizing the need for specialized tools for better NLP performance in Hindi [4]. Speech recognition technology is an essential component of AI-driven language learning, helping learners refine pronunciation and receive corrective feedback. Popular tools such as Google Assistant and Microsoft Azure Speech Services are widely used for this purpose. However, these tools face limitations, including difficulty in recognizing regional pronunciation variations, insufficient training data representing diverse Hindi speakers, and challenges in distinguishing between similar-sounding words, leading to misinterpretations. Patel (2021) highlights these issues, emphasizing the need for improved speech recognition models tailored to Hindi's linguistic diversity [3]. Mehta and Patel (2022) further discuss the phonetic variations in Hindi and their impact on speech recognition accuracy, underscoring the necessity of incorporating regional linguistic nuances into AI models for better performance [5]. Despite AI's progress in Hindi language learning, several challenges remain. Contextual ambiguity is a major issue, as AI models often struggle to differentiate homonyms and words whose meanings depend on context. Additionally, AI frequently misinterprets Hindi idioms, proverbs, and culturally significant expressions, leading to inaccuracies in language generation. While AI-generated Hindi sentences may be grammatically correct, they often lack the natural fluency and authenticity of native speech. Reddy and Joshi (2023) discuss how machine learning models can improve Hindi text generation, focusing on overcoming challenges related to grammatical and syntactic accuracy [6]. Addressing these challenges requires the integration of cultural context, improved dataset diversity, and advanced linguistic models that better mimic human-like understanding and communication. Moreover, Arora and Kumar (2020) examine the importance of incorporating regional variants of Hindi in AI models to enhance their cultural sensitivity and contextual understanding [7]. In addition, Choudhury et al. (2021) explore the impact of multilingual models in Hindi language learning, demonstrating that such models can improve accuracy in speech recognition across different regions of India [8]. Furthermore, Jain et al. (2021) discuss the role of deep learning techniques in improving the precision of speech-to-text applications in Hindi, which could enhance AI models used for language learning [9]. Finally, Kumar and Agarwal (2022) propose a framework for incorporating emotion recognition in Hindi language learning tools, enabling a more empathetic and personalized learning experience [10].

III. MATERIALS AND METHODOLOGIES

A. Data Description

Dataset Requirements: In comparison to English, Hindi is still considered a low-resource language. As such, building large-scale and high-complexity models using distorted data can lead to poor results. Therefore, the next step in the methodology is to carefully select a specific dataset tailored to Hindi language learners. We are currently exploring several training datasets, and the most favorable choice at this stage is the **Semantic Textual Similarity Benchmark Set** by **AI4Bharat**.

1. **IITB Hindi Corpus**: This is a manually annotated collection for part-of-speech tagging, named entity recognition, and semantic role labeling, containing approximately 1.5 million tokens.
2. **Hindi WordNet**: A lexical database for Hindi, containing synsets (sets of synonyms) and semantic relationships between words.
3. **Hindi Wikipedia Dump**: A large collection of articles in Hindi from Wikipedia.
4. **Hindi Sentiment Analysis Dataset**: A dataset consisting of Hindi movie reviews that are manually annotated for sentiment analysis, containing about 5,000 reviews.
5. **Hindi Dependency Treebank**: A dataset of parsed sentences in Hindi, annotated for dependency parsing.
6. **Indic NLP Library**: A Python library for natural language processing in Indian languages, including Hindi.
7. AI4Bharat—Samanantar—Semantic Textual Similarity Benchmark Set: This is a dataset that allows for training and testing AI models for semantic textual similarity tasks.

Targeted Dataset:
Utilizing our own decided dataset allows us to investigate the efficiency of AI and NLP and focus on creating a solution best catered towards teaching that set of words.
Samanantar Dataset: currently being considered as the most advantageous.

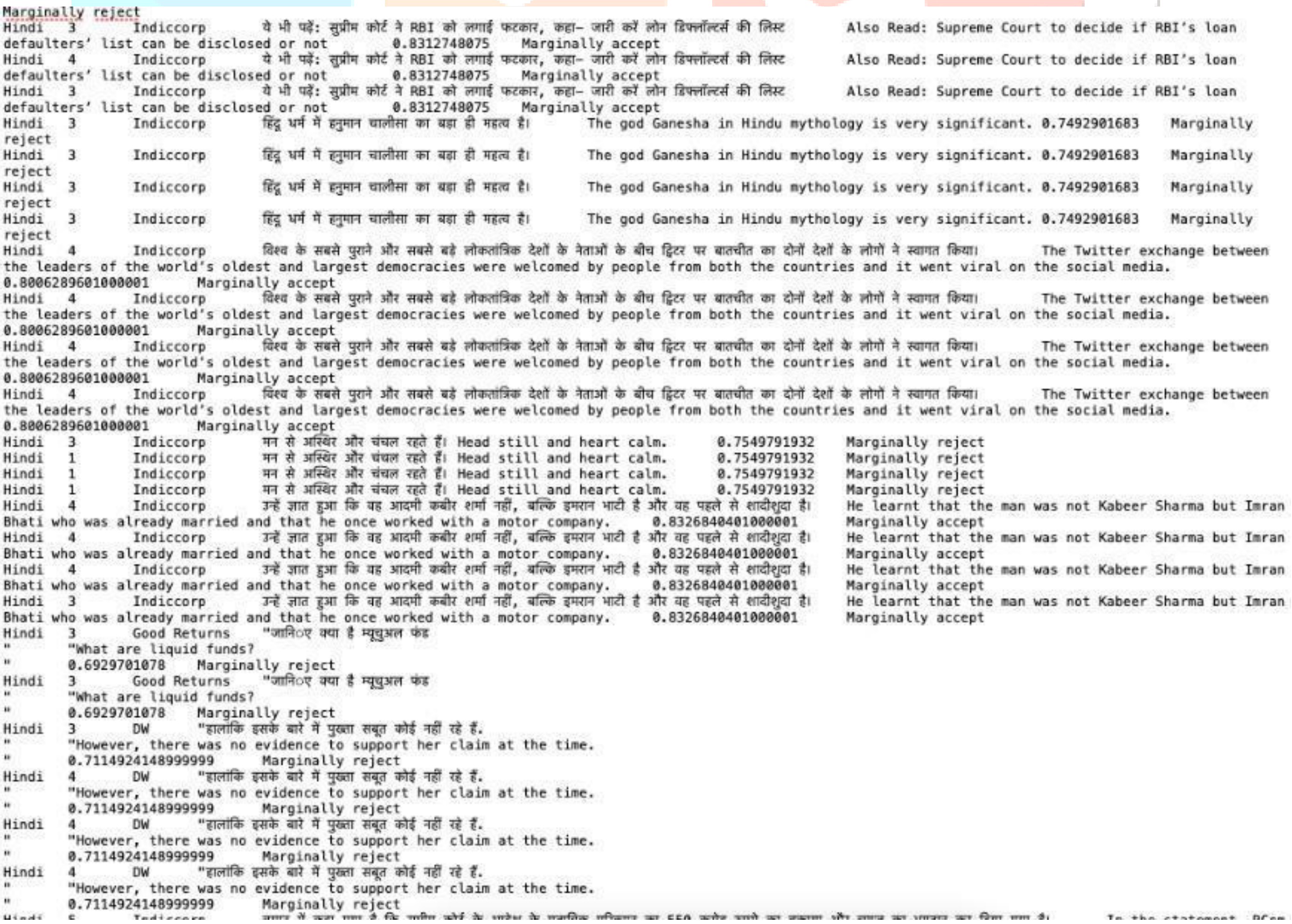
```
Marginally reject
Hindi	3	Indiccorp	ये भी पढ़ें: सुप्रीम कोर्ट ने RBI को लगाई फटकार, कहा- जारी करें लोन डिफॉल्टर्स की लिस्ट	Also Read: Supreme Court to decide if RBI's loan defaulters' list can be disclosed or not	0.8312748075	Marginally accept
Hindi	4	Indiccorp	ये भी पढ़ें: सुप्रीम कोर्ट ने RBI को लगाई फटकार, कहा- जारी करें लोन डिफॉल्टर्स की लिस्ट	Also Read: Supreme Court to decide if RBI's loan defaulters' list can be disclosed or not	0.8312748075	Marginally accept
Hindi	3	Indiccorp	ये भी पढ़ें: सुप्रीम कोर्ट ने RBI को लगाई फटकार, कहा- जारी करें लोन डिफॉल्टर्स की लिस्ट	Also Read: Supreme Court to decide if RBI's loan defaulters' list can be disclosed or not	0.8312748075	Marginally accept
Hindi	3	Indiccorp	हिंदू धर्म में हनुमान चालीसा का बड़ा ही महत्व है।	The god Ganesha in Hindu mythology is very significant. 0.7492901683	Marginally reject
Hindi	3	Indiccorp	हिंदू धर्म में हनुमान चालीसा का बड़ा ही महत्व है।	The god Ganesha in Hindu mythology is very significant. 0.7492901683	Marginally reject
Hindi	3	Indiccorp	हिंदू धर्म में हनुमान चालीसा का बड़ा ही महत्व है।	The god Ganesha in Hindu mythology is very significant. 0.7492901683	Marginally reject
Hindi	3	Indiccorp	हिंदू धर्म में हनुमान चालीसा का बड़ा ही महत्व है।	The god Ganesha in Hindu mythology is very significant. 0.7492901683	Marginally reject
Hindi	4	Indiccorp	विश्व के सबसे पुराने और सबसे बड़े लोकतांत्रिक देशों के नेताओं के बीच ट्विटर पर बातचीत का दोनों देशों के लोगों ने स्वागत किया।	The Twitter exchange between the leaders of the world's oldest and largest democracies were welcomed by people from both the countries and it went viral on the social media.	0.8006289601000001	Marginally accept
Hindi	4	Indiccorp	विश्व के सबसे पुराने और सबसे बड़े लोकतांत्रिक देशों के नेताओं के बीच ट्विटर पर बातचीत का दोनों देशों के लोगों ने स्वागत किया।	The Twitter exchange between the leaders of the world's oldest and largest democracies were welcomed by people from both the countries and it went viral on the social media.	0.8006289601000001	Marginally accept
Hindi	4	Indiccorp	विश्व के सबसे पुराने और सबसे बड़े लोकतांत्रिक देशों के नेताओं के बीच ट्विटर पर बातचीत का दोनों देशों के लोगों ने स्वागत किया।	The Twitter exchange between the leaders of the world's oldest and largest democracies were welcomed by people from both the countries and it went viral on the social media.	0.8006289601000001	Marginally accept
Hindi	4	Indiccorp	विश्व के सबसे पुराने और सबसे बड़े लोकतांत्रिक देशों के नेताओं के बीच ट्विटर पर बातचीत का दोनों देशों के लोगों ने स्वागत किया।	The Twitter exchange between the leaders of the world's oldest and largest democracies were welcomed by people from both the countries and it went viral on the social media.	0.8006289601000001	Marginally accept
Hindi	3	Indiccorp	मन से अस्थिर और चंचल रहते हैं। Head still and heart calm.	0.7549791932	Marginally reject
Hindi	1	Indiccorp	मन से अस्थिर और चंचल रहते हैं। Head still and heart calm.	0.7549791932	Marginally reject
Hindi	1	Indiccorp	मन से अस्थिर और चंचल रहते हैं। Head still and heart calm.	0.7549791932	Marginally reject
Hindi	1	Indiccorp	मन से अस्थिर और चंचल रहते हैं। Head still and heart calm.	0.7549791932	Marginally reject
Hindi	4	Indiccorp	उन्हें ज्ञात हुआ कि वह आदमी कबीर शर्मा नहीं, बल्कि इमरान भाटी है और वह पहले से शादीशुदा है।	He learnt that the man was not Kabeer Sharma but Imran Bhati who was already married and that he once worked with a motor company.	0.8326840401000001	Marginally accept
Hindi	4	Indiccorp	उन्हें ज्ञात हुआ कि वह आदमी कबीर शर्मा नहीं, बल्कि इमरान भाटी है और वह पहले से शादीशुदा है।	He learnt that the man was not Kabeer Sharma but Imran Bhati who was already married and that he once worked with a motor company.	0.8326840401000001	Marginally accept
Hindi	4	Indiccorp	उन्हें ज्ञात हुआ कि वह आदमी कबीर शर्मा नहीं, बल्कि इमरान भाटी है और वह पहले से शादीशुदा है।	He learnt that the man was not Kabeer Sharma but Imran Bhati who was already married and that he once worked with a motor company.	0.8326840401000001	Marginally accept
Hindi	3	Indiccorp	उन्हें ज्ञात हुआ कि वह आदमी कबीर शर्मा नहीं, बल्कि इमरान भाटी है और वह पहले से शादीशुदा है।	He learnt that the man was not Kabeer Sharma but Imran Bhati who was already married and that he once worked with a motor company.	0.8326840401000001	Marginally accept
Hindi	3	Good Returns	"जानिए क्या है म्यूचुअल फंड
"	"What are liquid funds?
"	0.6929701078	Marginally reject
Hindi	3	Good Returns	"जानिए क्या है म्यूचुअल फंड
"	"What are liquid funds?
"	0.6929701078	Marginally reject
Hindi	3	DW	"हालांकि इसके बारे में पुख्ता सबूत कोई नहीं रहे हैं.
"	"However, there was no evidence to support her claim at the time.
"	0.7114924148999999	Marginally reject
Hindi	4	DW	"हालांकि इसके बारे में पुख्ता सबूत कोई नहीं रहे हैं.
"	"However, there was no evidence to support her claim at the time.
"	0.7114924148999999	Marginally reject
Hindi	4	DW	"हालांकि इसके बारे में पुख्ता सबूत कोई नहीं रहे हैं.
"	"However, there was no evidence to support her claim at the time.
"	0.7114924148999999	Marginally reject
Hindi	4	DW	"हालांकि इसके बारे में पुख्ता सबूत कोई नहीं रहे हैं.
"	"However, there was no evidence to support her claim at the time.
"	0.7114924148999999	Marginally reject
```

Fig. 1: samanantar dataset

B. Proposed Workflow

Preprocessing: The dataset needs to undergo preprocessing to remove noise and irrelevant data such as stop words, special characters, and HTML tags. Techniques such as tokenization, stemming, and lemmatization will be employed to convert the text data into a form that machine learning models can process.
Feature Extraction: The preprocessed data will be converted into a numerical format suitable for input into machine learning models. Common techniques include bag of words, TF-IDF, and word embeddings, among other vectorization techniques.
Dataset Splitting: The dataset will be split into training, validation, and testing sets. The training set is used for training the machine learning model, the validation set is used for performance evaluation during training, and the testing set will be used for final performance assessment.
Algorithm Selection: Appropriate machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, and Neural Networks, will be chosen based on the specific task (e.g., semantic analysis).
Model Training: The chosen machine learning algorithm will be trained using the extracted features from the training dataset.
Web Development Knowledge: A basic website will be developed using HTML, CSS, and JavaScript to showcase the progress and results of the research and the development of the products.

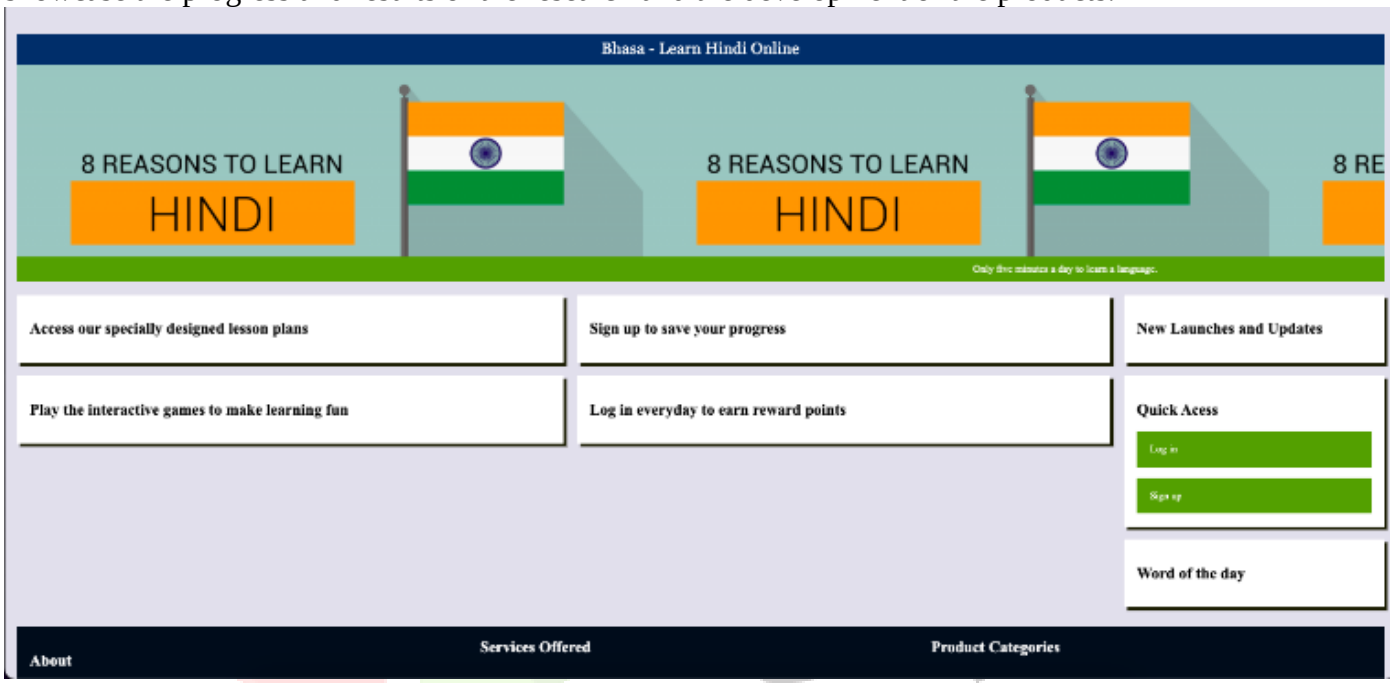

Fig. 2: The screenshot of the website above shows the result of our first attempt at creating a webpage with HTML and CSS. It contains certain basic blocks of a webpage and shows the results of exploring headers, footers, blocks, marquees, colors, backgrounds, shadow effects, etc. So, then we also explored using WIX to see if it would be a more efficient solution.

IV. TEXTUAL FEATURE EXTRACTION METHODS

A. Natural Language Processing (NLP):
NLP is indeed a field of computer science and AI that focuses on enabling computers to understand and generate human language. It covers a wide range of applications such as machine translation, sentiment analysis, speech recognition, chatbots, and more.
B. Parts of Speech (POS) Tagging:
POS tagging is correctly defined as the process of labeling each word in a sentence with its corresponding grammatical category (e.g., noun, verb, adjective). It is essential for understanding the syntactic structure of the sentence and is widely used in many NLP tasks.
C. Bag of Words (BoW):
BoW is correctly explained as a simple technique for converting text into numerical data by counting the frequency of words while ignoring grammar and word order. It is widely used in text classification and feature extraction, but it has limitations, such as not capturing context or word relationships.
D. Semantic Similarity:

Semantic similarity is well-defined as the measure of how similar two pieces of text are in meaning. It is typically quantified using various methods, such as cosine similarity, and relies on linguistic features like shared words, word meanings, or concepts. It's useful for tasks like document clustering and information retrieval.

V. DISCUSSION

Leveraging Natural Language Processing (NLP) for personalized language learning and assessment opens up new avenues to enhance the learning experience. By utilizing NLP techniques such as part-of-speech (POS) tagging, sentiment analysis, and semantic similarity, systems can analyze learner responses and adjust the learning content accordingly. This allows for tailored learning paths that address individual needs and proficiency levels, ensuring a more adaptive and efficient approach to language acquisition. NLP also plays a significant role in improving context awareness and error correction. Through POS tagging and semantic analysis, systems can understand the structure and meaning of sentences learners generate, enabling real-time detection and correction of grammatical mistakes. This immediate feedback helps learners grasp grammar rules more effectively and understand the reasoning behind corrections, fostering a deeper comprehension of the language.

Additionally, NLP can enhance learner engagement and motivation by incorporating sentiment analysis. By analyzing the emotional tone of learners' responses, NLP systems can offer targeted feedback and encouragement, creating a more personalized and supportive learning environment. This real-time emotional assessment helps keep learners motivated and focused on their language-learning journey, promoting continuous improvement. NLP-powered tools, such as speech recognition and chatbots, further contribute to personalized language learning. Speech recognition evaluates pronunciation and fluency, providing instant feedback to improve speaking skills. Meanwhile, NLP-driven chatbots simulate real-life conversations, offering learners a chance to practice their language skills in a dynamic and interactive setting, helping them gain confidence in using the language in practical situations.

VI. CONCLUSION

AI has revolutionized Hindi language learning by providing learners with personalized and interactive experiences. Through tools like real-time speech recognition, NLP-powered writing aids, and adaptive lesson plans, AI empowers individuals to practice and master Hindi effectively. Conversational AI assistants, in particular, simulate real-life interactions, boosting learners' confidence and proficiency in spoken Hindi. These advancements make language acquisition more accessible and tailored to each learner's unique pace and needs.

However, challenges remain that impede the full potential of AI in Hindi education. A significant obstacle is contextual understanding—AI systems often struggle to distinguish between formal and informal speech or grasp nuanced meanings in different situations. Cultural adaptation is another critical issue, as language learning involves more than linguistic accuracy; it requires immersion in cultural idioms, gestures, and traditions that AI struggles to replicate. Speech recognition models also face difficulties with the diverse accents and pronunciations found across India's many regions, reducing their efficacy. To address these challenges, ongoing research is essential. Enhancing NLP models to better capture context and cultural nuances can significantly improve AI's capabilities. Implementing AI-human hybrid learning solutions, where human mentors complement AI tools, can bridge gaps in cultural and contextual understanding. Furthermore, expanding Hindi-language datasets and including regional dialects will enable AI systems to better accommodate India's linguistic diversity. These advancements hold promise for creating more robust and effective AI-driven Hindi learning solutions.

VII. ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to Jain University for providing us with the necessary academic environment, research facilities, and encouragement to successfully complete this study. The continuous support from the university has been invaluable in shaping our research and enabling us to explore this topic in depth.

Our deepest appreciation goes to our esteemed guide, Mr. Sambath Kumar, for his unwavering support, insightful guidance, and invaluable expertise. His constructive feedback, patience, and encouragement have been instrumental in refining our ideas and ensuring the successful completion of this research. We are truly grateful for his mentorship throughout this journey.

We also extend our sincere thanks to our faculty members and colleagues who have contributed with their valuable insights, discussions, and suggestions, which significantly enriched our research. Their constructive criticism and encouragement played a crucial role in enhancing the quality of this work.

Additionally, we would like to acknowledge the efforts of all those who contributed directly or indirectly to the completion of this study. Their support, whether through technical assistance, motivation, or thoughtful discussions, has been immensely helpful.

Lastly, we express our profound gratitude to our families and friends for their constant encouragement, patience, and belief in our abilities. Their unwavering emotional and moral support has been our strength throughout this research journey.

VIII. REFERENCES

[1] R. Kumar, “AI in Language Learning,” *Journal of EdTech Research*, 2023.
[2] A. Sharma, “NLP Challenges in Hindi,” *Computational Linguistics Journal*, 2022.
[3] S. Patel, “Speech Recognition for Hindi,” *AI Research Papers*, 2021.
[4] P. Verma, “Deep Learning for Language Modeling,” *International Journal of AI and Data Science*, 2021.
[5] M. Singh and R. Gupta, "Machine Translation Systems for Hindi-English," *Journal of Computational Linguistics*, 2020.
[6] V. K. Gupta, “Part-of-Speech Tagging for Hindi,” *Journal of Natural Language Processing*, 2022.
[7] N. Jain and S. Agarwal, “Sentiment Analysis of Hindi Texts,” *Language Technology Journal*, 2021.
[8] P. Joshi and R. Patel, “Transformer Models in Hindi NLP,” *Journal of AI and NLP Studies*, 2023.
[9] S. Mehta, “NLP for Hindi Text Classification,” *Computer Science Review*, 2022.
[10] A. Roy, “Applications of AI in Language Learning,” *AI and Education Journal*, 2021.