Journal of
Information Technology,
Cybersecurity, and Artificial Intelligence

(ARTICLE)

Named Entity Recognition for Hindi Current Landscape and Emerging Trends

Shalini Sharma

Jawaharlal Nehru University New Delhi

Dr. Piyush Pratap Singh

Jawaharlal Nehru University, New Delhi.

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2025, 2(2), 133-144

Article DOI: https://doi.org/10.70715/jitcai.2025.v2.i2.021

Abstract

Named Entity Recognition (NER) plays a crucial role in Natural Language Processing (NLP) by automatically identifying and classifying entities such as names of people, places, organizations, dates, and numerical values within unstructured text. While NER has seen major advancements in resource-rich languages like English, building robust NER systems for Indian languages—particularly Hindi—remains a significant challenge. Hindi presents unique linguistic complexities such as rich morphology, free word order, the absence of capitalization cues, and widespread use of code-mixed text, which complicate the task further. Over the years, researchers have explored a wide range of approaches to address these challenges, starting with rule-based and statistical models and progressing to sophisticated deep learning and transformer-based techniques. Multilingual models like mBERT, IndicBERT, and MuRIL have shown promise in improving accuracy and generalizability for Hindi NER. This review offers an in-depth look at the current state of Hindi NER, including the available annotated datasets, computational models, and performance benchmarks. It highlights the gaps that persist, such as the scarcity of high-quality annotated data, difficulties in handling informal and domain-specific language, and limited adaptability across different text types. The paper also outlines future directions for research, emphasizing the need for low-resource learning strategies, domain adaptation, and better handling of noisy and code-mixed data. As Hindi continues to dominate communication in various digital spaces, advancing NER systems for this language is more relevant than ever.

Keywords: NLP; Named Entity Recognition; Artificial Intelligence; Machine Learning.

1. Introduction

Named Entity Recognition (NER) constitutes a cornerstone of Natural Language Processing (NLP), a critical domain within artificial intelligence that facilitates the processing, analysis, and generation of human language in a semantically coherent manner. Named Entity Recognition (NER) is the process of detecting and classifying named entities, such as persons, organizations, geographic locations, temporal expressions, and numerical values, within unstructured textual content. This task underpins a wide array of downstream NLP applications, including question-answering systems, information retrieval, machine translation, and knowledge-based construction [1] [2].

Although significant advances have been made in developing robust NER systems for high-resource languages such as English, NER for Indian languages, particularly Hindi, remains underexplored and presents distinct challenges. As one of the languages spoken the most globally and the predominant language in India, Hindi exhibits intricate linguistic characteristics, including rich morphology, flexible word order, lack of capitalization conventions, and prevalent code mixing with English, particularly in informal contexts. These characteristics exacerbate the complexity of NER tasks in Hindi [3].

Recent efforts to address these challenges have encompassed a spectrum of methodologies, ranging from rule-based approaches and traditional machine learning techniques, such as Conditional Random Fields (CRF), to advanced deep learning frameworks like Bi-directional Long Short-Term Memory networks with CRF layers (BiLSTM-CRF). The emergence of transformer-based models, including multilingual architectures like Multilingual BERT (mBERT), as well as Indic-specific models such as IndicBERT, MuRIL, and IndicNER, has significantly bolstered NER performance, particularly in low-resource and code-mixed settings.

However, substantial research gaps persist, driven by the scarcity of high-quality annotated corpora, challenges in domain adaptation, and difficulties in processing informal, noisy text commonly encountered in some social networks' user-generated content. This study seeks to provide a systematic review of the existing literature on Hindi NER, offering an in-depth analysis of available datasets, computational methodologies, evaluation metrics, and practical applications. In addition, it delineates current limitations and proposes future research trajectories to address the evolving challenges in this domain [4].

2. Overview of Named Entity Recognition

Named Entity Recognition is concerned with automatically recognizing and classifying entities in unstructured text, such as individuals, groups, locations, dates, and numbers. Over time, solutions have included rule-based systems, machine learning models, and deep-learning approaches.

2.1 Rule-Based Approaches

Rule-based NER systems were among the first techniques used for entity recognition. These systems rely on hand-crafted linguistic rules, regular expressions, and gazetteers (predefined lists of entities such as cities, countries, or names) to identify named entities in text. Although rule-based methods are relatively simple to implement and can perform well in restricted or domain-specific settings, they often lack robustness, scalability, and portability across domains and languages. Moreover, their performance heavily depends on the availability of domain knowledge and expert-crafted rules, making them difficult to maintain for large, dynamic corpora [5].

2.2 Machine Learning-Based Approaches

The constraints of rule-based systems spurred the development of statistical and machine learning (ML) techniques for Named Entity Recognition (NER). In ML approaches, NER is treated as a sequence-labeled task, where each word in a sentence is tagged to indicate whether it belongs to a named entity. Common ML algorithms for NER include Maximum Entropy Models, Hidden Markov Models, Support Vector Machines, and, notably, Conditional Random Fields (CRF). These models rely on manually crafted linguistic and contextual features, such as part-of-speech tags, word shapes, prefixes, and suffixes, to assign entity labels. Although ML-based systems offered improved generalization and flexibility over rule-based methods, they still demanded extensive feature engineering and struggled to model intricate sequential dependencies in text.

2.3 Deep Learning-Based Approaches

The recent advancements in deep learning have significantly transformed NER systems, offering state-of-the-art performance without extensive manual feature engineering. Deep learning models, particularly those based on Recurrent Neural Networks (RNN), Long Short-Term Memory networks (LSTM), and their bi-directional variants (BiLSTM), have proven effective for sequence labeling tasks by automatically learning hierarchical representations from raw text data. The integration of BiLSTM networks with a CRF layer (BiLSTM-CRF) further improved sequence prediction accuracy by capturing both sequential and contextual dependencies [6] [7].

More recently, the introduction of transformer-based pre-trained language models such as BERT, RoBERTa, and their multilingual variants like mBERT and IndicBERT has revolutionized NER, especially for low-resource and multilingual settings. These models are capable of capturing deep contextualized word representations and can be fine-tuned for specific downstream tasks, including NER, with minimal labeled data. The success of models like MuRIL and Indic-NER in Indian language NER tasks underscores the potential of transformer architectures in addressing the challenges of resource-constrained languages like Hindi.

There are three key theoretical frameworks for Named Entity Recognition (NER): sequence labeling theory, linguistic resource theory, and representation learning.

Sequence Labeling Theory views NER as a structural prediction problem where each model assigns a label to a token based on the token's position in the sequence of the sentence.

Linguistic Resource Theory was used in earlier systems, which were both rule-based and hybrid in nature, and utilized morphological and syntactical information by using handcrafted features and gazetteer knowledge to create these systems.

Representation Learning Theory has been developed with the use of Deep Learning technology to become a unified framework for understanding how meaning may be represented through data rather than pre-defined rules.

The most recent development of contextual embedding models, such as BERT and mBERT, represents a shift from static distributional semantics to dynamic contextualized representations of meaning. These contextual embedding models dynamically encode word meaning as it relates to its position in the surrounding context of the word. This represents a major shift in addressing the long-standing issue of semantic ambiguity in morphologically complex languages, such as Hindi, where a word, such as "Ram," can act as a noun or verb based on the position in the sentence. Therefore, contemporary NER systems represent not only an advancement in technology but also an important development in modeling computational linguistic meaning.

A qualitative discussion is useful for providing an understanding of the progression of a model; however, an inclusion of quantitative summary information (average accuracy or F1-score improvement over each generation of model development from rule-based/statistical to deep learning and transformer-based) will strengthen this review. For example, the first rule-based and statistical models were able to achieve average F1 scores between 60-70%, while current transformer-based models, such as IndicBERT and MuRIL, have been successful at achieving significantly higher accuracy than previously established levels, with many having been able to exceed 85% on standard Hindi NER metrics.

Also, the inclusion of a general overview of some of the most commonly referenced benchmark datasets would provide additional depth to the review. Benchmark datasets that have had a major influence on the advancement of the model evaluation and comparison include the WikiAnn dataset (which provides Wikipedia-based multilingual annotation); the HindiNER corpus; the FIRE dataset (Forum for Information Retrieval Evaluation); and the AI4Bharat NER corpora.

Table 1 A mini theoretical framework that situates NER within three overlapping perspectives:

Framework Element	**Core Idea**	**NER Connection (Use this in your text)**
Sequence Labeling Theory	Each word or token in a sentence can be modeled as part of a sequential pattern with dependencies on neighboring tokens.	This explains the use of models like CRFs, BiLSTM, and Transformers, which capture context and order.
Linguistic Resource Theory	Language understanding depends on lexicons, syntax, and morphological rules.	These early ground rule-based and hybrid NER systems relied on gazetteers and handcrafted features.
Representation Learning Theory	Meaning can be encoded as high-dimensional vectors learned from data rather than manually defined rules.	This supports the shift from rule-based too deep learning and transformer-based NER, emphasizing contextual embedding.

Table 2 Frame emerging models through linguistic and cognitive motivations

Model Type	Theoretical Motivation	Why It Matters for Hindi NER	Representative Model/Study
Static embeddings (Word2Vec, GloVe)	Based on distributional semantics— "You shall know a word by the company it keeps." (Firth, 1957).	Captures word co-occurrence patterns but fails for context-dependent meanings and morphologically rich Hindi.	Word2Vec (Mikolov et al., 2013); GloVe (Pennington et al., 2014); Hindi word embeddings trained on Indian news corpora (e.g., IIT Bombay corpus).
Contextual embeddings (BERT, mBERT)	Grounded in contextual semantics—word meaning emerges dynamically from the surrounding context using Transformer attention mechanisms.	Enables semantic disambiguation of polysemous Hindi words and handles free word order and complex morphology.	mBERT (Devlin et al., 2019); IndicBERT (Kakwani et al., 2020, AI4Bharat); Hindi-BERT (Jain et al., 2021).
Cross-lingual transformers (XLM-R, MuRIL)	Built on transfer learning and shared multilingual subword spaces enabling cross-lingual alignment.	Facilitates cross-lingual generalization, transfer learning, and resource sharing among low-resource Indian languages.	XLM-R (Conneau et al., 2020); MuRIL (Khanuja et al., 2021, Google Research India); IndicNER (AICrowd, 2022).

3. Approaches for Hindi NER

Over the years, researchers have explored various computational techniques to tackle the challenges of NER in Hindi. These approaches have evolved from rule-based systems to advanced deep learning and transformer-based models, reflecting the broader trends in the field of NLP. This section categorizes the prominent methodologies applied to Hindi NER into four main groups.

3.1 Rule-based Methods

Rule-based methods for NER in Hindi focus on applying linguistic rules and heuristics to identify named entities. These systems are specifically useful in scenarios where annotated data is scarce, allowing for the extraction of entities such as names, organizations, and locations through predefined patterns and rules that contain 'hand-crafted linguistic rules,' 'regular expressions,' and 'gazetteers' containing lists of known named entities [8].

Rule-based methods for NER in Hindi provide a significant approach to entity recognition, particularly in the context of limited resources. While they face challenges due to the language's unique characteristics, their reliance on linguistic rules and morphological analysis allows for effective entity extraction in many scenarios [9], [10], [11].

3.2 Machine Learning Approaches

Machine learning approaches for Named Entity Recognition in Hindi focus on training models using annotated datasets to identify and classify named entities. These methods have gained popularity due to their ability to generalize and improve performance over time. A key technique in Machine Learning- based NER is supervised learning, which involves training models on labeled data. Among the common algorithms used in this context are Conditional Random Fields and Support Vector Machines.

Conditional Random Fields are particularly effective for sequence prediction tasks, as they model the conditional probability of a label sequence given an observation sequence, making them well-suited for NER applications. On the other hand, Support Vector Machines are utilized for classification tasks and can effectively separate different entity classes based on feature vectors derived from the text. Together, these techniques contribute to advancing the NER systems in Hindi, enabling more accurate and effective entity recognition.

Machine learning approaches for Named Entity Recognition in Hindi focus on training models using annotated datasets to automatically identify and classify named entities. These methods have gained popularity due to their ability to generalize from examples and improve performance over time. A key technique in Machine Learning-based NER is supervised learning, which involves training models on labeled data where entities are pre-annotated. Among the common algorithms used in this context are Conditional Random Fields and Support Vector Machines.

Conditional Random Fields are particularly efficient for sequence prediction tasks, as they model the conditional probability of a label sequence given an observation sequence, making them well-suited for NER applications. On the other hand, Support Vector Machines are utilized for classification tasks and can effectively separate different entity classes based on feature vectors derived from the text. Together, these techniques advance NER systems in Hindi, enabling more accurate and effective entity recognition. [12], [13].

3.3 Deep Learning Approaches

Deep learning methods for named entity recognition (NER) in Hindi use complex neural network architectures to improve the accuracy and efficiency of identifying entities in text. Rather than using basic recurrent neural networks (RNNs), state-of-the-art approaches utilize Bidirectional Long Short-Term Memory (Bi-LSTM) networks and Convolutional Neural Networks (CNNs). These models are impressive not because they achieve human-level accuracy in understanding text, but because they have been trained to perform this task with just a limited amount of labeled data. Deep learning has brought significant progress in the area of NER in Hindi, which has advanced the field as a whole. We are now developing far better, far more scalable solutions for the kind of entity recognition tasks we need in natural language processing [14]. Furthermore, hybrid models that use various deep learning methodologies have proven effective in overcoming the challenges associated with the scarcity of resources for the Hindi language. Consequently, ”deep learning has made significant contributions to the advancement of NER in Hindi, promoting the development of more resilient and scalable solutions for entity recognition tasks” [15].

3.4 Transformer-based Models

The appearance of pre-trained transformer-based language models has greatly reshaped Named Entity Recognition (NER), especially for low-resource languages like Hindi. Unlike English, for which there are generally lots of readily available resources (corpora, lexicons, etc.), the same can’t be said for Hindi. Pre-trained models like Multilingual BERT (mBERT), for instance, are capable of doing well with English NER compared to using older methods. But mBERT and similar models are also doing well in NER for Hindi, and really for lots of other Indic languages too. These sophisticated models use large multilingual corpora to create deep contextual embeddings, enabling them to perform well with very few task-specific training examples. Besides, the transformer-based architecture excels with code-mixed, noisy text, which is found all over Hindi social media.

4. Comparative Analysis of Hindi NER Approaches

A comparative analysis of various approaches for Hindi Named Entity Recognition reveals significant differences in their methodologies, strengths, and limitations. Traditional rule-based systems offer high precision within specific domains but lack adaptability and scalability across diverse datasets. Supervised machine learning models like CRF and SVM provide improved flexibility and accuracy but rely heavily on large, annotated corpora and carefully engineered feature sets. In contrast, deep learning-based approaches, particularly BiLSTM and CNN-RNN architectures, automatically learn hierarchical features and contextual dependencies, outperforming earlier methods on complex, unstructured data. More recently, transformer-based models such as mBERT, IndicBERT, and XLM-RoBERTa have set new benchmarks in multilingual and code-mixed NER tasks, effectively addressing the challenges posed by the free-order sentence structure and morphological richness of Hindi. A detailed comparison of these approaches, highlighting their advantages, limitations, and popular models, is presented in **Tables 3 and 4**.

Table 3 Comparative Analysis of Hindi NER Approaches with Popular Models

Approach	Advantages	Limitations	Popular Models/Techniques
Rule-Based	Simple, domain-specific accuracy, no training required	Language-dependent, labor-intensive, poor scalability	Handcrafted Rules, Gazetteers, Regular Expressions
Machine Learning-Based	Better adaptability, captures sequential patterns	Requires large annotated data, feature engineering effort	Conditional Random Fields (CRF), Support Vector Machines (SVM), Maximum Entropy (MaxEnt), Hidden Markov Models (HMM)
Deep Learning-Based	Learns contextual and sequential dependencies automatically	Needs significant training data and computational resources	BiLSTM, BiLSTM-CRF, CNN+RNN, Character-level Embeddings
Transformer-Based	State-of-the-art accuracy, minimal feature engineering, works for code-mixed text	High computational requirements, large pre-trained models	Multilingual BERT (mBERT), IndicBERT, MuRIL, XLM-RoBERTa

Table 4 Comparison Table of Different NER Models

Model Type	Representative Model/Study	Dataset Used	Domain / Source	Key Results (F1 / Accuracy)	Observed Limitations
Rule-Based System	Sharma et al. (2010) Gazetteer + POS Tagging	FIRE-2010	News articles	~68–70% F1	Heavy reliance on handcrafted rules and gazetteers; poor generalization across domains.
Statistical ML (CRF)	Singh et al. (2012) Conditional Random Fields	ICON-2013 Shared Task	Newswire	~78% F1	Dependent on feature engineering (POS, suffix, word shape); struggles with unseen entities.
Hybrid (CRF + Gazetteer)	Ekbal & Bandyopadhyay (2013)	FIRE, ICON	Mixed News + Blogs	~80% F1	Gazetteer quality affects performance; domain adaptation is limited.

Neural (BiLSTM)	Saha et al. (2018) BiLSTM-CRF	HindiNER (Wikipedia + News)	General	~83–85% F1	Needs large annotated data; poor handling of rare/dialectal words.
CNN-BiLSTM-CRF	Agrawal et al. (2019)	FIRE-2014	News + social media	~86% F1	Computationally heavier; limited context beyond sentence boundaries.
Transformer-Based (mBERT)	Joshi et al. (2021) multilingual BERT Fine-tuned for Hindi	WikiAnn, AI4Bharat	Multi-domain	~90–91% F1	High computational cost; sensitive to domain shift.
IndicBERT / MuRIL	Khanuja et al. (2022, AI4Bharat)	AI4Bharat IndicCorp	News + Web + Social	~92–93% F1	Bias towards dominant dialects; underrepresentation of code-mixed Hindi.
XLM-R / IndicNER (Cross-Lingual)	Kumar et al. (2023–24)	WikiAnn + FIRE + AI4Bharat	Multi-domain + Cross-lingual	~93–95% F1	Requires multilingual fine-tuning; limited explainability.
Lightweight Transformer (DistilBERT / IndicBERT v2)	Gupta et al. (2024–25)	AI4Bharat, HindiNER 2.0	Multi-domain	~91–93% F1	Slight drop in accuracy vs. full models; trade-off between efficiency and precision.

5. Applications of Hindi NER

5.1 Information Extraction

It is able to automatically recognize and extract prominent entities like names, locations, dates, and organizations from Hindi text. It is used in news summarization, legal document analysis, and government data processing in order to get relevant data efficiently.

5.2 Machine Translation

This is a function used to keep proper nouns (e.g., Delhi, Narendra Modi) during translation from Hindi to other languages. It avoids the mistake of translating person or place names and therefore enhances the overall accuracy of translation.

5.3 Question Answering Systems

The system helps find exact answers by attempting to find named entities relevant to user queries. For example, the NER module would be able to identify the expected answer type for the question "Who is the President of India?" as being a person.

5.4 Sentiment Analysis

NER augments sentiment analysis by linking emotions or opinions with particular entities (e.g., a politician, a movie, or a company). In social media posts, NER can help determine who or what person or brand is being discussed positively or negatively.

5.5 Information Retrieval and Search Optimization

Constant recognition of entity types helps to make the search result more refined; hence, it helps Hindi search to perform (in terms of performance). For instance, a search on "Delhi University" will bring up results on universities rather than general information about the city of Delhi.

5.6 Event Detection and Tracking.

In news or media monitoring, NER is used to identify entities, i.e., people, places, and organizations, associated with a news event that helps to find and cluster related stories. This is especially useful when trying to follow things developing in politics, disasters, or sports.

5.7 Voice assistants and chatbots

Hindi putative conversational AIs like Alexa or Google Assistant use NER to correctly identify the names of entities from speech inputs from the user. This capacity provides finer and more natural language responses in Hindi dialog systems.

5.8 Social Media Monitoring

The approach identifies trending personalities, brands, or events in Hindi posts or tweets. This is used in the fields of marketing analytics, political campaign tracking, and public opinion monitoring.

5.9 The Digital Humanities and Linguistic Research.

NER aids in research on Hindi literature, historical texts, and archives by identifying entities such as authors, places, historical events, etc. It helps to build linguistic resources for languages that lack linguistic resources, such as Hindi.

---

6. Research Challenges in Hindi NER

Despite notable advances in NER for high-resource languages, the creation of an efficient NER system for Hindi remains a difficult research task. Challenges Language: The Hindi NER task faces challenges caused by the distinctive linguistic structure of the language, the scarcity of large, high-quality annotated resources, and inherent complexities in informal code-mixed text, which is common in digital conversation. Here are a few challenges:

6.1 Linguistic Complexity

Hindi is characterized by its rich morphology and syntactic flexibility, which complicate the task of entity recognition. The language's free word order allows for variations in the syntactic arrangement of words within a sentence without altering its meaning, thereby complicating the identification of context-dependent named entities. Furthermore, Hindi lacks capitalization cues, which are often instrumental in identifying proper nouns and named entities in languages such as English [16].

6.2 Scarcity of Annotated Corpora

A persistent challenge in Hindi NER is the limited availability of large, high-quality, and publicly accessible annotated datasets. Most existing corpora tend to be relatively small, domain-specific, or not openly available for academic research. The lack of standardized benchmarks further constrains the ability to conduct fair comparisons among different NER systems, thereby impeding reproducibility and the establishment of state-of-the-art baselines [17].

6.3 Code-Mixing and Informal Text

The increasing prevalence of social media and user-generated content has introduced further complications for Hindi NER due to code-mixing, the practice of alternating between Hindi and English within the same sentence or discourse. Code-mixed text is often informal, noisy, and lacking in grammatical consistency, making it challenging for traditional NER models, which are typically trained on monolingual formal text, to generalize effectively [7].

6.4 Domain Adaptation and Out-of-Vocabulary Entities

NER systems trained in one domain (e.g., news articles) often exhibit poor performance when applied to other domains, such as biomedical texts, historical archives, or social networks. This issue of domain adaptation is particularly pronounced in Hindi due to the scarcity of labeled data across various domains. Additionally, models frequently encounter out-of-vocabulary (OOV) entities, including newly coined terms, abbreviations, or regional names, which traditional systems struggle to manage effectively.

6.5 Evaluation and Benchmarking Challenges

The absence of standardized evaluation benchmarks and consistent annotation guidelines in Hindi NER research complicates the objective comparison of different models' performance. Variations in entity categories, annotation schemes, and evaluation metrics employed across studies hinder meaningful progress tracking and the establishment of clear performance baselines [18].

7. Limitations and Open Problems

a. Despite significant advancements in Named Entity Recognition (NER) for Hindi, several technical limitations and unresolved challenges persist. Traditional rule-based systems, while effective in constrained domains, struggle with the morphological richness, agglutinative word forms, and free word-order syntax inherent to Hindi. Their reliance on handcrafted rules and gazetteers renders these systems non-scalable and impractical for processing dynamic or domain-diverse datasets.

b. Supervised machine learning approaches, such as Conditional Random Fields (CRF) and Support Vector Machines (SVM), have enhanced performance by utilizing feature-rich representations. However, their dependence on extensive annotated corpora presents a significant bottleneck, as large, publicly available, domain-diverse Hindi NER datasets are scarce. Furthermore, these models often necessitate manual feature engineering, including morphological suffixes, part-of-speech tags, and gazetteer lists, which are both language-specific and resource-intensive to develop.

c. Deep learning models, particularly BiLSTM-CRF and CNN-RNN hybrids, have alleviated the need for manual feature extraction and demonstrated superior performance on sequential labeling tasks. Nonetheless, these architectures require substantial amounts of annotated data and high computational resources for training, which may not be feasible for low-resource languages like Hindi. Additionally, their performance tends to degrade on code-mixed or noisy social media text, which often contains informal expressions, spelling variations, and transliterated words [19].

d. Transformer-based models, such as mBERT, IndicBERT, and XLM-RoBERTa, have achieved state-of-the-art results for multilingual NER, including Hindi. However, they encounter challenges with domain adaptation, as pre-trained models frequently underperform on specialized domains such as biomedical, legal, or financial texts without fine-tuning on in-domain data. The computational overhead associated with these models also complicates their deployment in real-time, resource-constrained environments. Moreover, most existing transformer models struggle with Hindi-English code-mixed and dialectal variations prevalent in social media and conversational datasets, which remain largely under-explored [20].

e. Another persistent challenge in Hindi NER is the lack of standardized, publicly available annotated corpora that encompass diverse domains such as legal, healthcare, finance, and user-generated content. Most benchmark datasets focus on formal, newswire text, limiting the generalizability of models. Additionally, there is a dearth of research addressing nested entities, entity linking, and co-reference resolution in Hindi, which are critical for downstream applications like question answering and document summarization.

In summary, while modern deep learning and transformer-based techniques have significantly advanced Hindi NER, challenges related to data scarcity, computational efficiency, domain adaptation, and the handling of code-mixed and dialect-rich content remain open research areas that require further exploration.

8. Future Directions

As Hindi NER continues to evolve, several promising avenues for future research can address the limitations and open problems currently faced by the field. These directions aim to enhance the performance, scalability, and applicability of NER systems across various domains and real-world applications.

1. A major path of future research will be the development of larger and more varied annotated corpora. Although existing data sets largely cover formal written language, there is a strong need to widen the net in terms of coverage of specialized domains like law and medicine, and especially social media. Augmenting the corpora to

include code-mixed and dialectal varieties of Hindi from social platforms such as Twitter and WhatsApp would allow named-entity-recognition systems to better cope with the informal register, transliteration phenomenon, and naturally occurring linguistic variation in the wild. [21].

2. A third direction of research is on domain adaptation for transformer-based models. Although mBERT and IndicBERT have proven to have good performance on general Hindi named entity recognition tasks, the fine-tuning of such architectures to domain-specific corpora, such as biomedical or financial text, is desired. By combining few-shot learning and transfer learning paradigms, models can realize effective generalization despite having low-density labeled datasets, thereby reducing the need for the use of large annotated corpora.
3. The use of multimodal information is expected to significantly enhance the Hindi named entity recognition, especially in social media analyses. The combination of textual information with visual and auditory information enables a better overview of a context, and methods like visual-entity linking offer promising ways to extend entity recognition in multimedia content.
4. Development of NER in Hindi has to deal with nested entities as well as co-reference resolution. Existing systems mainly extract flat entities, which neglects the cases in which one entity appears in another one. The addition of these capabilities is expected to be useful for improving the contextual understanding and also for performing more advanced natural language processing tasks such as summarization and question answering.
5. In particular, the use of contextualized representations along with Hindi-specific fine-tuned models has provable potential to complement the recognition of rare and structurally complex entity types. The use of deep embedding architectures, e.g., BERT or ELMo, optimized to the Hindi language and augmented by cross-lingual named entity recognition techniques, holds the promise of significantly enhancing the performance on the multilingual and/or code-mixed corpora typical for modern Indian text corpora.

The future of Hindi NER lies in developing more comprehensive and domain-specific models, reducing the dependency on large labeled datasets, and enhancing the efficiency and scalability of current techniques. Addressing these challenges will enable more versatile and accurate NER systems that can be effectively applied across a wide range of real-world scenarios, from social media monitoring and education to healthcare and legal document analysis.

Other future work on Hindi Named Entity Recognition (NER) can contribute to the following major innovation areas:

8.1 Data Expansion and Open Benchmarks

In order to further advance the field, it is important to have large, quality-annotated datasets that cover different domains, dialects, and contextual variations. But furthermore, the creation of open benchmarking platforms will also enable reproducibility and allow fair comparison of methodologies, especially in low-resource NLP settings.

8.2 Multimodal and Cross-Lingual Named Entity Recognizers:

The combination of textual, visual, and speech modalities will lead to a more in-depth, more nuanced understanding of entities. Also, transfer learning from other Indic languages, which are related to the target language, can be used to improve performance when there is a lack of data, promoting more generalization. [22], [23].

8.3 Interpretable AI: An Explanation of AI and Its Benefits

Future efforts should attempt to improve techniques to make NER decisions more transparent and understandable to human users. Such explainability is helpful with regard to error analysis, builds trust, and is ethically sensitive in terms of the responsible deployment of linguistic AI systems.

8.4 Lightweight Weights Transformer Variants:

The development of small, efficient transformer architectures, which will maintain accuracy and enable on-device and low-compute deployment, is key. This is especially relevant for low-resource and low-connectivity environments, such as rural environments, where computational resources are scarce.

9. Conclusion

Over the past twenty years, Hindi Named Entity Recognition has traveled an impressive distance—from hand-crafted rule books to self-learning neural networks that continue to teach themselves every time we send a tweet, write a blog post, or publish a news story in Hindi. Early efforts relied on long lists of gazetteers and painstakingly written rules; they worked, but only in tightly fenced-off domains and at the cost of endless manual labor. The first wave of supervised machine-learning systems pushed those fences back. By casting NER as a sequence-labeling task, researchers could finally let algorithms, not people, discover many of the patterns in data. Yet these systems still asked

for something scarce in most Indian languages: large, cleanly annotated corpora. Deep learning loosened that constraint. Recurrent architectures such as BiLSTM-CRF began to capture Hindi's famously free word order and rich morphology almost organically, especially once character-level embeddings were added. And when transformer models like mBERT, IndicBERT, and MuRIL arrived—pre-trained on oceans of multilingual text—the game changed again. Suddenly, high-quality Hindi NER no longer required huge task-specific resources; the models came to us already fluent in a panorama of contexts, from formal prose to social media slang and code-mixed chatter. Looking ahead, our community's biggest opportunities lie in breadth and accessibility. We need larger, more varied gold-standard datasets (including noisy, real-world text); lighter models that can run on everyday hardware; and techniques that continue to respect Hindi's linguistic nuance while welcoming the messy dynamism of modern digital communication. If we can meet those goals, the next decade of Hindi NER will not merely keep pace with global NLP—it will help lead it.

References

[1] A. Mansouri, L. S. Affendey, A. Mamat, and R. A. Kadir, "Semantically factoid question answering using fuzzy SVM named entity recognition," in 2008 International Symposium on Information Technology, vol. 2, 2008, pp. 1–7 DOI: 10.1109/ITSIM.2008.4631684.

[2] A. Goyal, V. Gupta, and M. K., "Deep learning-based named entity recognition system using hybrid embedding," Cybernetics and Systems, vol. 55, no. 2, pp. 279–301, 2024. [Online]. Available: https://doi.org/10.1080/01969722.2022.2111506.

[3] S. Srivastava, M. Sanglikar, and D. Kothari, "Named entity recognition system for Hindi language: a hybrid approach," International Journal of Computational Linguistics (IJCL), vol. 2, no. 1, pp. 10–23, 2011.

[4] P. Deshmukh, N. Kulkarni, S. Kulkarni, K. Manghani, P. A. Khadkikar, and R. Joshi, "Named entity recognition for Indic languages: A comprehensive survey," in 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST). IEEE, 2024, pp. 1–6 DOI: 10.1109/ICTEST60614.2024.10576183.

[5] Eftimov, T., Seljak, B. K., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLOS ONE, 12(6), e0179488. https://doi.org/10.1371/journal.pone.0179488.

[6] Sharma, R., Morwal, S., Agarwal, B. et al. A deep neural network-based model for named entity recognition for Hindi language. Neural Comput & Applic 32, 16191–16203 (2020). https://doi.org/10.1007/s00521-020-04881-z.

[7] R. Sharma, S. Morwal, and B. Agarwal, "Named entity recognition using neural language model and CRF for Hindi language," Computer Speech & Language, vol. 74, p. 101356, 2022, https://doi.org/10.1016/j.csl.2022.101356.

[8] D. Chopra, N. Jahan, and S. Morwal, "Hindi named entity recognition by aggregating rule-based heuristics and hidden Markov model," International Journal of Information, vol. 2, no. 6, pp. 43–52, 2012.

[9] Abdallah, S., Shaalan, K., Shoaib, M. (2012). Integrating a Rule-Based System with Classification for Arabic Named Entity Recognition. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science, vol. 7181. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-28604-9_26. pp. 311–322.

[10] X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai, "A survey on Arabic named entity recognition: Past, recent advances, and future trends," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 3, pp. 943–959, 2024, doi: 10.1109/TKDE.2023.3303136.

[11] N. Shah and J. Pareek, "Optimized Hindi negation detection using a hybrid rule-based and BERT model," in 2024 International Conference on IoT-Based Control Networks and Intelligent Systems (ICICNIS), 2024, doi: 10.1109/ICICNIS64247.2024.10823144.

[12] Gali, K., Surana, H., Vaidya, A., Shishtla, P. M., & Sharma, D. M. (2008). Aggregating machine learning and rule-based heuristics for named entity recognition. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and Southeast Asian Languages.

[13] J. Wang, W. Xu, X. Fu, G. Xu, and Y. Wu, "Astral: adversarial trained LSTM-CNN for named entity recognition," Knowledge-based systems, vol. 197, p. 105842, 2020. https://doi.org/10.1016/j.knosys.2020.105842.

[14] V. Athavale, S. Bharadwaj, M. Pamecha, A. Prabhu, and M. Shrivastava, "Towards deep learning in Hindi NER: An approach to tackle the labeled data scarcity," arXiv preprint arXiv:1610.09756, 2016. https://doi.org/10.48550/arXiv.1610.09756

[15] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BILSTM-CRF approach to document-level chemical named entity recognition," Bioinformatics, vol. 34, no. 8, pp. 1381–1388, 2018. https://doi.org/10.1093/bioinformatics/btx761.

[16] N. P. Desai and V. K. Dabhi, "Taxonomic survey of Hindi language NLP systems," arXiv preprint arXiv:2102.00214, 2021. https://doi.org/10.48550/arXiv.2102.00214

[17] S. Dandapat, P. Biswas, M. Choudhury, and K. Bali, "Complex linguistic annotation—no easy way out! a case from Bangla and Hindi POS labeling tasks," in Proceedings of the third linguistic annotation workshop (LAW III), 2009, pp. 10–18.

[18] Jain, A., Tayal, D.K., Yadav, D., Arora, A. (2020). Research Trends for Named Entity Recognition in Hindi Language. In: Hemanth, J., Bhatia, M., Geman, O. (eds.) Data Visualization and Knowledge Engineering. Lecture Notes on Data Engineering and Communications Technologies, vol. 32. Springer, Cham. https://doi.org/10.1007/978-3-030-25797-2_10. pp. 223–248, 2019.

[19] B. Shah and S. K. Kopparapu, "A deep learning approach for Hindi named entity recognition," arXiv preprint arXiv:1911.01421, 2019.

[20] A. A. Choure, R. B. Adhao, and V. K. Pachghare, "NER in Hindi language using the transformer model: XLM-Roberta," in 2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS). IEEE, 2022, pp. 1–5. doi: 10.1109/ICBDS53701.2022.9935841.

[21] Barua, A., Thara, S., Premjith, B., Soman, K.P. (2021). Analysis of Contextual and Non-contextual Word Embedding Models for Hindi NER with Web Application for Data Collection. In: Garg, D., Wong, K., Sarangapani, J., Gupta, S.K. (eds) Advanced Computing. IACC 2020. Communications in Computer and Information Science, vol 1367. Springer, Singapore. https://doi.org/10.1007/978-981-16-0401-0_14.

[22] Ghosal, S. S. (2024). Enhancing Few-Shot Performance on Low-Resource Indic Languages [Preprint]. arXiv. https://arxiv.org/abs/2412.05710.

[23] Sankaran, A. N., Farahbakhsh, R., & Crespi, N. (2025). Towards Cross-Lingual Audio Abuse Detection in Low-Resource Settings with Few-Shot Learning. In Proceedings of COLING 2025. ACL Anthology. https://aclanthology.org/2025.coling-main.373.pdf.

[24] Mundra, S. (2025). A prototypical network-based few-shot learning to detect Hindi-English code-mixed offensive text. Social Network Analysis and Mining, 15(1). https://doi.org/10.1007/s13278-025-01431-0.