Evaluating Named Entity Recognition Models on Indian English News Headlines

1st Arya Tayshete
Department of Electrical Engineering
Veermata Jijabai Technological Institute
Mumbai, India
avtayshete_b21@et.vjti.ac.in

2nd Aryaan Pandhare
Department of Computer Engineering
Veermata Jijabai Technological Institute
Mumbai, India
aapandhare_b21@it.vjti.ac.in

***Abstract*—Named Entity Recognition (NER) is a crucial part of Natural Language Processing and is used in numerous applications. Although its performance significantly varies across diverse linguistic contexts. Indian contexts are characterized by regional languages including their dialects. This presents unique challenges for NER models. The research assesses the performance of small Language Models-BERT, Flair, spaCy and Stanza. We used a custom annotated dataset tailored with Indian-contextual English news articles that contain English/Hindi words used largely by the public in Indian English. The evaluation highlights strengths and weaknesses of each model and their linguistic biases. The results provide insights that are actionable to optimize these NER models assisting educational institutes, developers and researchers in India and all over the world in making decisions and handle Indian-specific textual data better.**

I. Introduction

Natural Language Processing (NLP) is a branch of computer science and artificial intelligence that helps computers understand and work with human languages, like English or Hindi. In simple terms, NLP allows computers to read text, interpret speech, figure out what we mean, and reply in ways that feel natural to us. We can think of tools like Google Translate, Siri, or the text prediction on your phone—all of these use NLP. Basically, NLP is about teaching computers to "talk" and "listen" more like people, making it easier for us to communicate with technology.

NLP covers a wide range of tasks that help machines understand, interpret, and generate human language. Key tasks include **Language Identification (LID)**, which detects the language of a given text and is important in multilingual settings. **Multilingual Language Inference (MLI)** helps assess relationships like entailment and contradiction across various languages, supporting reasoning across languages. **Named Entity Recognition (NER)** finds and classifies entities such as names, organizations, and places, which is crucial for extracting structured information from unstructured text. **Part-of-Speech (POS) Tagging** assigns grammatical labels, like noun or verb, to each word, assisting in analyzing sentence structure. Machine Translation (MT) enables automatic translation between languages, often depending on effective tagging and entity recognition to keep meaning intact. Together, these basic tasks lay the groundwork for more advanced language understanding systems, especially in linguistically diverse areas like India.

NER is a key element in NLP, essential for applications that include information retrieval, sentiment analysis, question-answering systems, and conversational AI. Despite progress in NER methods thanks to deep learning and transformer-based language models, challenges still exist. These challenges are particularly evident when dealing with linguistically diverse and culturally complex texts, such as those found in Indian contexts. India has a unique linguistic environment, marked by multilingualism, code-switching, and a mix of formal and informal language usage, often referred to as Hinglish.

Small Language Models like **BERT, Flair, spaCy**, and **Stanza** have become popular because they are lightweight, efficient, and easy to deploy in both academic and practical settings. However, these models are mainly trained on datasets that provide limited exposure to Indian-specific named entities and contexts. This can lead to notable performance gaps. As a result, their effectiveness for tasks such as educational support, local news analysis, or extracting contextual information within Indian universities and colleges is uncertain.

This paper looks at how small LLMs perform specifically in Indian educational and practical applications. We created a custom annotated dataset that includes English news articles in an Indian context. These articles are labeled for entities such as `PERSON`, `LOCATION`, `ORGANIZATION`, `MISCELLANEOUS`, and `EVENT`. Using this dataset, we carried out comparative experiments with BERT-based models, Flair, spaCy, and Stanza NER models. Our evaluation not only identifies how well each model performs overall but also shows specific strengths, weaknesses, and linguistic biases. The insights from this research aim to help educational institutions, developers, and researchers choose and improve models more suited to the unique needs of Indian textual data.

II. Background

A. Natural Language Processing (NLP) in AI

Natural Language Processing (NLP) is a part of artificial intelligence (AI) that helps machines understand, interpret, and generate human language. In the last ten years, NLP has changed quickly because of advancements in deep learning and pre-trained language models. NLP is now essential for many

AI applications, including machine translation, sentiment analysis, chatbots, and automated summarization..

B. Named Entity Recognition (NER) in NLP

Among the main tasks in NLP, Named Entity Recognition (NER) is particularly important. NER involves identifying and classifying segments of text into specific categories like people, organizations, locations, dates, and other entities. It plays a key role in tasks such as information extraction, question answering, knowledge graph construction, and content tagging.

Traditional rule-based NER systems relied heavily on manually crafted features and linguistic rules. These systems have gradually been replaced by machine learning, and more recently, deep learning methods. Models like Conditional Random Fields (CRFs) gave way to neural architectures such as LSTMs and GRUs, which were then surpassed by transformer-based models like BERT. These models have achieved top performance on a wide range of benchmarks, especially when fine-tuned for specific domains or languages.

Despite these advancements, using NER with Indian English text presents significant challenges. Indian news articles, social media posts, and educational documents often have code-switched text (mixing English with Hindi), local named entities, non-standard spellings, and culturally unique references that are not well covered in Western training data. As a result, pre-trained models often have poor generalization and confuse entities when applied to Indian data.

This study aims to explore how small Language Models, which require less computing power than larger models, perform in these contexts. By focusing on models like BERT, Flair, spaCy, and Stanza, we evaluate their effectiveness in real-world Indian NLP scenarios, especially in resource-limited academic and research settings.

III. Problem Statement

Despite rapid advancements in Named Entity Recognition (NER) through deep learning and transformer-based models, most existing systems have been developed and benchmarked on datasets rooted in Western linguistic and cultural contexts. This results in suboptimal performance when these models are applied to Indian English, which features code-switching, unique named entities, cultural references, and non-standard syntactic structures. Furthermore, small language models like BERT, Flair, spaCy, and Stanza, while efficient, often lack the contextual training necessary to accurately capture and classify named entities specific to Indian domains, such as regional locations, festivals, institutional acronyms, or colloquial expressions. The absence of comprehensive evaluation on Indian-contextual data creates a blind spot in deploying reliable NER systems for local educational, administrative, and journalistic applications.

IV. Research Objectives

The primary aim of this study is to evaluate the performance of small LLM-based NER models in understanding and extracting named entities from Indian English texts. Specifically, the objectives are:

- To construct a manually annotated NER dataset using Indian English news articles, covering entity classes such as `PERSON`, `LOC`, `ORG`, `EVENT`, and `MISC`.
- To evaluate and compare the performance of BERT-based, Flair, spaCy, and Stanza NER models on this dataset.
- To identify domain-specific misclassifications, linguistic biases, and strengths or weaknesses of each model.
- To offer recommendations for improving NER performance in culturally diverse, resource-constrained Indian NLP environments.

V. Significance of Study

This research offers a focused look at NER systems in a less explored language area. By creating and using a dataset designed for Indian English contexts, the study provides important insights into how small LLMs perform with non-standard, culturally rich data. The findings matter for both academic NLP research and practical uses in Indian education, media monitoring, digital governance, and AI tools. This work addresses a significant gap by revealing the strengths and weaknesses of commonly used models. It ultimately helps in developing more inclusive and context-aware NER systems.

VI. Literature Review

COMI-LINGUA—Large-Scale Hinglish NER Dataset : The COMI-LINGUA dataset is a high-quality, expert-annotated dataset for code-mixed Hindi–English text focusing on various NLP tasks including NER. It spans over 27k raw instances while 25k filtered ones in NER. The authors observed that leveraging multilingual data and transformer-based approaches, such as multilingual BERT, greatly improves performance on code-mixed NER. This dataset highlights the importance of domain-specific corpora for Indian languages including names, events, and colloquial expressions.

Fine-Tuning Pretrained NER Models for Indian Languages: This recent study presents a manually curated NER dataset for four Indian languages, containing 40,000 sentences annotated for entities across domains. By fine-tuning both multilingual transformer models and monolingual versions (e.g., Hindi BERT, Marathi BERT), they achieved aggregate F1-scores around 0.80. Their findings emphasize that multilingual training helps performance, but blindly mixing datasets can sometimes hurt accuracy—suggesting dataset selection still matters. Although not specific to Hinglish, this work informs our approach with small LLM fine-tuning for Indian data.

Large-scale Indic NER Dataset : Naamapadam is currently the largest publicly available NER dataset for eleven Indian languages, encompassing over 400,000 sentences and 9.4 million annotated entities in PERSON, LOCATION, and ORGANIZATION categories. The dataset provides extensive coverage across diverse languages and regional contexts, enabling benchmarking across traditionally underrepresented Indian languages. While Naamapadam does not include EVENT

or MISC labels, it sets a strong standard for scale and diversity in Indian NER.

Hindi–English Code-Mixed Social Media NER: Vinay Singh et al. presented one of the early corpora and evaluation studies on NER for Hindi–English code-mixed tweets. They experimented with traditional machine learning models such as Decision Trees, LSTMs, and Conditional Random Fields (CRFs). The best-performing models (CRF and LSTM) achieved F1-scores around 0.95 on their annotated corpus, highlighting the feasibility of structured approaches in code-mixed social media texts. Despite the high performance, the study focused only on social media data and used standard labels rather than Indian context labels like "EVENT" or "MISC."

VII. Methodology

A. Dataset Creation

To evaluate NER models on Indian-contextual data, we curated a dataset comprising around 500 English news headlines and article snippets sourced from regional Indian news websites. These samples reflected common patterns in Indian journalism, including named entities such as local places, politicians, educational institutes, government programs, festivals, and infrastructure projects. Annotation was done using the Label Studio tool and reviewed to ensure quality. The annotation schema followed five entity classes:

- `PERSON`: Names of people (e.g., "Narendra Modi")
- `LOC`: Geographical locations (e.g., "Ayodhya", "Tamil Nadu")
- `ORG`: Organizations (e.g., "BJP", "Gujarat University")
- `EVENT`: Events and public programs (e.g., "Janmabhoomi", "New Year")
- `MISC`: Entities not fitting above, like product names, currency mentions, religious texts (e.g., "Quran", "iPhone X")

The final dataset included 472 labeled entities from 238 news headlines.

B. Data Annotation

To prepare the dataset for NER evaluation, we performed manual annotation of 238 English-language news headlines. The annotation process was conducted using **Label Studio**, an open-source data labeling platform. For convenience and collaborative access, Label Studio was hosted locally on a machine within a private network and accessed remotely from multiple PCs through a secure setup.

The annotation interface allowed tagging of text spans with five predefined entity types: `PERSON`, `LOC`, `ORG`, `EVENT`, and `MISC`. Each entity span was selected manually to ensure accurate boundary detection and semantic correctness. Ambiguous cases were reviewed through cross-checking and iterative discussion. The annotated data was exported in JSON format compatible with the standard Label Studio NER schema and later normalized for downstream evaluation tasks.

C. Selected Models

We evaluate four widely used NER models in this study: **BERT**, **Flair**, **spaCy**, and **Stanza**. Each model has unique architectural strengths and trade-offs relevant to multilingual and code-mixed entity recognition tasks.

1) BERT (Fine-Tuned for NER): BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model pretrained on large-scale corpora. For NER, it is fine-tuned using token-level classification objectives. Its bidirectional context awareness allows it to capture subtle dependencies in sentences. However, it requires alignment of WordPiece tokenizations and is computationally intensive. Despite its powerful representations, BERT may underperform on niche domains without sufficient fine-tuning data.

2) Flair: Flair leverages character-level contextual string embeddings trained using LSTM architectures. These embeddings can be stacked with traditional word embeddings (like GloVe or FastText), improving performance on morphologically rich and informal texts. Flair handles out-of-vocabulary and misspelled words better due to its character-level nature, but it is relatively slower than transformer-based models.

3) spaCy: spaCy uses CNN-based architectures with transition-based parsing for efficient NER. It offers fast inference and ease of integration, making it suitable for production-level NLP tasks. However, it relies on static embeddings and thus lacks the deep contextual understanding found in transformer models. Its performance may degrade on informal or domain-specific corpora.

4) Stanza: Stanza is Stanford NLP's modern neural pipeline offering multilingual support and modular design. It uses deep BiLSTM networks with CRF decoding for NER. Although not based on transformers, it offers strong baseline performance on a variety of languages and can handle morphologically complex structures well. However, its inference time is higher and contextual representation is limited compared to BERT.

These models were selected due to their lightweight nature, ease of deployment, and popularity in educational and applied NLP tasks.

VIII. Normalization and Evaluation

To account for label variations across models (e.g., "GPE" vs "LOC"), we implemented a normalization function mapping raw model outputs to our 5-tag schema. Predictions were converted into spans of the format (start_char, end_char, label). We computed span-level Precision, Recall, and F1-score for each label category and for each model individually. Predictions were matched with ground-truth spans based on exact character alignment and normalized labels. An error analysis was also performed to understand the frequent mismatches, entity fragmentations, and model biases.

.

IX. Results and Analysis

A. Per-Label Performance

Table I presents the precision, recall, and F1-scores for each of the four evaluated models for PERSON, LOC and ORG. The evaluation was done at the span-level with strict character alignment and label normalization.

TABLE I
ENTITY-WISE PRECISION, RECALL, AND F1-SCORE FOR PERSON, LOC, AND ORG ACROSS NER MODELS

Model	Entity	Precision	Recall	F1-score
BERT	PERSON	0.16	0.32	0.22
	LOC	0.34	0.44	0.39
	ORG	0.33	0.54	0.41
Flair	PERSON	0.77	0.80	0.78
	LOC	0.57	0.62	0.59
	ORG	0.38	0.56	0.45
spaCy	PERSON	0.40	0.38	0.39
	LOC	0.45	0.29	0.35
	ORG	0.26	0.44	0.33
Stanza	PERSON	0.56	0.72	0.63
	LOC	0.59	0.55	0.57
	ORG	0.49	0.56	0.52

B. Analysis

Based on the evaluation results in Table I, we observe the following insights regarding the performance of the four NER models—BERT, Flair, spaCy, and Stanza—on the key entity types: PERSON, LOC, and ORG.

- **Flair and Stanza outperformed others for the PERSON entity.** Flair achieved the highest F1-score of 0.78, followed by Stanza at 0.63. In contrast, BERT performed significantly worse with an F1-score of just 0.22, suggesting poor recall and weak entity span detection for persons.
- **Stanza showed robust overall performance.** Across all three entities, Stanza maintained strong precision and recall values, especially for PERSON (0.56/0.72) and ORG (0.49/0.56), indicating balanced performance.
- **spaCy lagged in recall.** Although spaCy had moderate precision, it struggled to identify many relevant entities, especially LOC and ORG, resulting in lower F1-scores of 0.35 and 0.33 respectively.
- **BERT's underperformance is notable.** Despite being a transformer-based model, the fine-tuned BERT model failed to generalize effectively on the custom annotated dataset. Its F1-scores for all three entities were significantly lower than expected, particularly for PERSON.
- **Flair maintained high precision and recall balance.** It consistently performed well across all entities with F1-scores exceeding 0.45, demonstrating that it may be better suited for low-resource or code-mixed data scenarios in Indian contexts.

C. NER Model Performance by Entity Label

The bar plot in Fig. 1 illustrates the F1-scores of four NER models—BERT, Flair, spaCy, and Stanza—across five major entity types. Flair consistently outperforms the others for the PERSON entity with an impressive F1-score of 0.78, followed by Stanza at 0.63. In contrast, BERT significantly underperforms for PERSON, with an F1-score of only 0.22. For LOC entities, both Flair (0.59) and Stanza (0.57) show strong performance, while BERT and spaCy lag slightly behind. The ORG entity sees moderate success across all models, with Stanza (0.52) leading, and spaCy (0.33) trailing. All models struggle to identify EVENT and MISC entities, with F1-scores below 0.2 in all cases, indicating these categories are particularly challenging in the given dataset. These results highlight that while Flair and Stanza are generally more robust, performance varies significantly across entity types.

This plot served as a critical diagnostic tool to identify model strengths, weaknesses, and potential biases with respect to specific Indian-context entity categories.

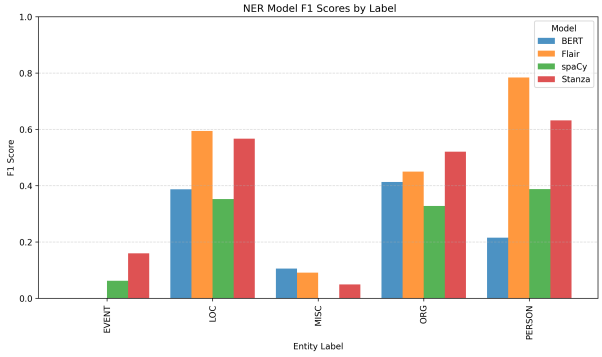

Fig. 1. F1 Score by Label for each Model

D. Token-Level Agreement Analysis Between NER Models

To better understand inter-model consistency, we compute the pairwise token-level agreement between all four NER models. The heatmap in Figure 2 illustrates the percentage of tokens for which each model pair assigned the same label, including the "O" (non-entity) class.

- **High Agreement Between BERT and Flair:** These two models show the strongest agreement (91.99%), likely due to their deep contextual embeddings and similar transformer-based architectures. Their ability to capture subword and character-level context helps align their token predictions more closely.
- **Moderate Agreement Between spaCy and Stanza:** The spaCy-Stanza pair shows relatively high agreement (84.72%), suggesting comparable tokenization and decision boundaries, possibly due to similar linguistic pipelines or shared pretraining corpora.
- **Low Agreement Between spaCy and BERT/Flair:** Token agreement between spaCy and the transformer-based models is notably lower (74.15% to 74.66%), which could be attributed to differences in tokenization strategies. For instance, BERT uses WordPiece tokenization, whereas spaCy uses rule-based methods, leading to mismatches in subword units and label spans.

- **Disparity Due to Fine-Tuning and Training Corpora:** The BERT model in our experiment is fine-tuned on the custom annotated dataset, which might introduce distributional shifts not seen by the off-the-shelf versions of spaCy or Stanza. This contributes to divergence in labeling, especially for less common entity types.

Overall, token-level agreement reflects not only architecture differences but also downstream factors like pretraining corpora, tokenization, and entity boundary assumptions. High agreement between BERT and Flair indicates shared capabilities in handling context-rich entities, while low agreement with spaCy highlights the challenge of aligning different NER pipelines.

Several key patterns emerge from the token-level agreement matrix:

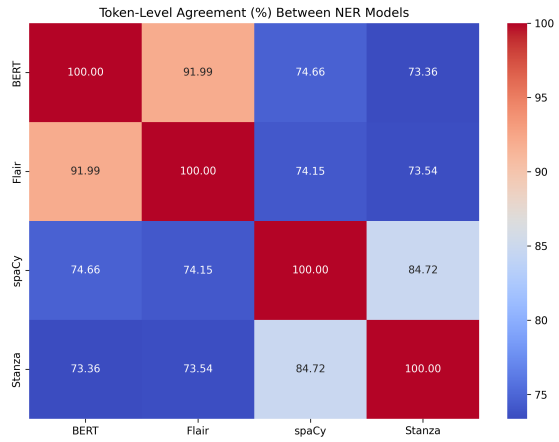

Fig. 2. Token-level % agreement between NER Models

X. Discussion

A. Insights from Performance

The evaluation of NER models on the custom annotated dataset revealed several interesting trends. As observed in Table I, Flair and Stanza consistently outperformed spaCy and BERT across most entity types. Specifically, Flair achieved the highest F1-score of **0.78** for `PERSON`, closely followed by Stanza (**0.63**). In contrast, BERT struggled significantly with `PERSON` entities, scoring just **0.22**, possibly due to insufficient fine-tuning or its sensitivity to low-resource or informal domain data.

In terms of `LOC` and `ORG` entities, all models exhibited relatively moderate performance. Flair again led in `LOC` with an F1 of **0.59**, whereas Stanza showed balanced performance across both `LOC` (**0.57**) and `ORG` (**0.52**). spaCy, while weaker in general, showed consistency but lower recall across the board. BERT's relatively poor results, especially on `MISC` and `EVENT`, indicate it may be misclassifying rare or domain-specific entities due to limited training examples.

The token-level agreement matrix in Figure 2 further supports these findings. BERT and Flair showed the highest agreement of **91.99%**, which suggests their token labeling patterns align significantly—likely due to their transformer-based architectures. Meanwhile, Stanza showed higher agreement with spaCy (**84.72%**) compared to BERT or Flair, possibly because of similar syntactic feature usage and segmentation strategies. Despite these agreements, F1 differences arise due to variation in entity boundary predictions and class assignments, especially for non-`O` tokens.

Overall, the analysis highlights that while some models share similar token labeling behavior, their ability to correctly classify and extract entities varies greatly, especially in low-resource or code-mixed domains. This underscores the importance of evaluating both aggregate agreement and label-wise performance in NER tasks.

B. Challenges in Indian NER

Named Entity Recognition in the Indian context poses several unique challenges due to the socio-linguistic and structural complexity of the language landscape. One of the most significant hurdles is the prevalence of code-mixing, especially between Indian languages and English, which confuses traditional tokenization and entity boundaries. This often results in inconsistent predictions for entities like names of people, places, or institutions that may be written partially in English or transliterated formats.

Additionally, Indian names and locations often exhibit high morphological and contextual ambiguity. For instance, the same word may refer to a person in one context and to a place or festival in another. This lack of disambiguation leads to confusion among models not explicitly fine-tuned for such variations. Moreover, named entities like government schemes, regional festivals, or caste-based organizations are typically absent from pretrained vocabularies of most models, thereby degrading performance.

The scarcity of large, high-quality annotated datasets covering Indian-specific domains (e.g., politics, local news, education, public policy) further hinders model generalization. Most open-source models like BERT or spaCy are trained on generic corpora from Western contexts, leading to poor recall on domain-specific Indian entities, particularly for underrepresented tags like `EVENT` and `MISC`.

Lastly, orthographic inconsistency—such as multiple spellings of the same name (e.g., "Modi ji" vs "Modiji") or inconsistent use of whitespace and punctuation—further complicates entity span alignment. These subtle but impactful issues must be addressed through better preprocessing, context-aware normalization, and culturally-aware model fine-tuning to improve the robustness of NER systems in Indian use cases.

C. Limitations of This Study

While this study offers valuable insights into how small NER models perform on Indian English data, it is important to recognize several limitations that affect the generalizability and scope of the findings.

First, the dataset used for evaluation contains only 238 manually annotated English news headlines. Although it includes a variety of content and sources, the dataset may not fully capture the linguistic and domain diversity found in broader Indian text collections, such as social media, parliamentary debates, or multilingual educational materials.

Second, the annotation process, although reviewed for consistency, was completed by two annotators. This introduces the possibility of subjective bias in label assignments and span selections. There was no analysis of multi-annotator agreement or expert validation, which could affect the reliability of the labeled ground truth.

Third, the models assessed—BERT (fine-tuned for NER), Flair, spaCy, and Stanza—were all used as they are, without additional training on Indian-specific datasets. Their performance, particularly on low-resource or culturally specific entities, such as local festivals or political schemes, may not show their full potential after domain-specific training or additional fine-tuning.

Additionally, span-level evaluation metrics depend on exact character alignment, which does not allow for partial matches or minor annotation errors. This strict requirement may underestimate model effectiveness in real-world situations where approximate or fuzzy entity detection can still be beneficial.

Lastly, deeper issues such as error propagation in downstream tasks, computational efficiency, and real-time inference feasibility were not assessed. These factors are important for practical deployment in real-world NLP systems, especially in low-resource or edge environments typical in Indian applications.

Future research could address these gaps by expanding the dataset, involving multiple annotators, exploring domain-specific fine-tuning, and performing qualitative error analysis in addition to quantitative metrics.

D. Implications for Future NER Models

The results of this study highlight several points to consider for developing future NER models, especially in Indian and multilingual settings.

First, model performance varied significantly across different entity types. This suggests that uniform architectures may not be enough. For instance, Flair and Stanza did well with PERSON entities but had difficulties with `ORGANIZATION` and `MISC` classes. This indicates a need for enhancements tailored to specific labels, such as hybrid rule-based or knowledge-aware modules that work alongside deep learning methods.

Second, token-level agreement analysis showed that even when macro scores differ, model predictions often agree on certain token spans. This reveals a hidden shared understanding. Future models could use ensemble learning or confidence-based aggregation of several models to improve recall and precision.

Additionally, the consistently poor performance of all models on `EVENT` and `MISC` entities indicates that training data needs to be more diverse. These labels are closely linked to cultural and domain-specific knowledge. Pretraining on Indian news, policy documents, or vernacular-English corpora could greatly enhance results.

Another important point is the annotation strategy. The difficulty in fine-grained span alignment and disagreements on boundary decisions suggest that sequence tagging could benefit from span-based architectures. These would allow for joint predictions of the start, end, and label without relying solely on token-wise BIO tagging.

Finally, the success of older models like Flair in some areas shows that model architecture is not always the main issue. We must also focus on data quality, label balance, and normalization strategies, which heavily impact performance.

Future NER research for Indian languages and code-mixed data should focus not just on larger models but also on better contextual grounding, smarter fine-tuning, and culturally aware training datasets.

XI. Conclusion

This study presents a systematic evaluation of four lightweight Named Entity Recognition (NER) models—BERT (fine-tuned for NER), Flair, spaCy, and Stanza—on a custom-annotated dataset of Indian English news headlines. The dataset, designed to reflect real-world usage and contextual nuances in Indian journalism, featured five key entity categories: `PERSON`, `LOC`, `ORG`, `EVENT`, and `MISC`.

Our analysis revealed notable differences in performance across models and entity types. Flair and Stanza performed well on PERSON entities, while spaCy struggled on most tags, and BERT underperformed on several span-level metrics despite its strong architecture. We also introduced a token-level agreement heatmap to visualize cross-model consensus, uncovering interesting overlaps even between models with diverging F1 scores.

The findings highlight critical challenges in applying off-the-shelf NER tools to Indian contexts and point toward key areas for future improvement. These include developing culturally grounded training data, designing span-aware architectures, and leveraging model ensembles to enhance reliability.

Overall, this research contributes both an empirical benchmark and practical insights for advancing NER systems tailored for Indian and code-mixed English-language content.

References

[1] Rajvee Sheth, Himanshu Beniwal, Mayank Singh. (2025). COMI-LINGUA: Expert Annotated Large-Scale Dataset for Multitask NLP in Hindi-English Code-Mixing. https://doi.org/10.48550/arXiv.2503.21670

[2] Sankalp Bahad, Pruthwik Mishra, Karunesh Arora, Rakesh Chandra Balabantaray, Dipti Misra Sharma, Parameswari Krishnamurthy, Fine-Tuning Pretrained NER Models for Indian Languages. https://arxiv.org/abs/2405.04829

[3] Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy V, Anoop Kunchukuttan, Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages. https://arxiv.org/abs/2212.10168

[4] Vinay Singh, Deepanshu Vijay, Syed Sarfaraz Akhtar, and Manish Shrivastava, Named Entity Recognition for Hindi-English Code-Mixed Social Media Text . https://aclanthology.org/W18-2405/

[5] Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora https://aclanthology.org/L18-1557/

[6] Mohd Zeeshan Ansari, Tanvir Ahmad, and Md Arshad Ali. 2018. Cross Script Hindi English NER Corpus from Wikipedia
Available: https://arxiv.org/abs/1810.03430

[7] Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. https://aclanthology.org/W14-3914/

[8] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection.
https://arxiv.org/abs/1805.11869

[9] Paras Tiwari, Sawan Rai, and C Ravindranath Chowdary. 2024. Large scale annotated dataset for code-mix abusive short noisy text.
https://dl.acm.org/doi/10.1007/s10579-023-09707-7

[10] Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named entity recognition for code-mixed indian corpus using meta embedding.
https://ieeexplore.ieee.org/document/9074379