Image Augmented Hindi Named Entity Recognition (IA - HiNER)

Elevating Named Entity Recognition for Hindi Text Processing with Image Context

Aastha Pal IIT2021020, Suyash Deep IIT2021029, Anjali Singh IIT2021031, Jyothsna Niharika IIT2021055, , Jai Moryani IIT2021122

***Abstract*—This report outlines ongoing research on Hindi Named Entity Recognition (NER) augmented with additional information extracted from images. While existing efforts have primarily focused on NER in Hindi text alone, this project aims to incorporate supplementary insights from images to enhance NER accuracy. The methodology involves dataset creation, model development, exploration of image-text fusion mechanisms, as well as training and evaluation processes.**

***Index Terms*—Hindi, Named Entity Recognition, Textual Data, Visual Information, Image Captioning, XLM, MURIL, Dataset Creation, Model Development.**

I. Introduction

Social media platforms have evolved into hubs of user-generated content, providing valuable insights into the thoughts, perspectives, and preferences of individuals and communities. In this dynamic environment, it's crucial to extract meaningful information from social media posts containing both text and images. Named Entity Recognition (NER) is vital in natural language processing, identifying and classifying entities in text data. While existing NER systems in Hindi have made progress, they struggle with social media posts that include both textual and visual information.

Image-Augmented Hindi Named Entity Recognition (IA-HiNER) addresses this challenge by enhancing textual data with visual cues from images, improving entity recognition accuracy. While IA-HiNER has been explored in English, there's a gap in Hindi IA-HiNER research.

Our project methodology involves several stages: dataset creation, model development, exploration of fusion mechanisms, design of model architecture, and training and evaluation to facilitate Hindi Image-Augmented Named Entity Recognition (Hindi IA-HiNER). By incorporating both textual and visual cues, our approach aims to enhance the accuracy and robustness of entity recognition in Hindi text augmented with images. We commence by developing a user-friendly annotation website tailored for our institute, utilizing it for Hindi text annotation, and curating a diverse dataset comprising Hindi text paired with accompanying images. Subsequently, we leverage a combination of image models for feature extraction and transformer-based pretrained models such as XLM-RoBERTa and MuRIL for text feature extraction.

This research explores Hindi IA-HiNER and its advantages over traditional Hindi NER systems. By integrating both textual and visual data, Hindi IA-HiNER enhances entity recognition accuracy, enabling applications in social media analysis, sentiment analysis, recommendation systems, and trend detection within Hindi-speaking communities. Hindi IA-HiNER targets Hindi language processing, aiming to improve entity recognition accuracy in Hindi text augmented with images.

Additionally, we collected a comprehensive Hindi dataset using several new APIs and developed an annotation tool tailored for our project. This dataset, enriched with annotations, serves as a valuable resource for training and evaluating our Hindi IA-HiNER model.

II. Literature Review

A. Named Entity Recognition (NER):

NER has gained prominence in the research community for natural language processing (NLP) tasks. While traditional methods rely on manual annotation and statistical learning, recent trends favor deep learning approaches. These include LSTM-based architectures, with, Hammerton [2003] pioneering the use of LSTM-CRF for entity recognition. Later, Lample et al. [2016] proposed a BiLSTM-CRF model, enhancing sequence modeling capabilities. CNN-CRF structures were also successful in NER, as shown by Pinheiro and Collobert (2014). Moreover, Luo et al. [2018] improved entity recognition performance using BiLSTM-CRF with attention mechanisms. The introduction of BERT further enhanced NER effectiveness, with subsequent studies focusing on optimizing pre-trained models (Devlin et al. [2018], ; Jawahar et al. [2019]). Combining BERT with BiLSTM-CRF (Liu et al. [2019]; Luo et al. [2018]) or GRU resulted in improved NER results on datasets like CoNLL-2003. However, these methods are less effective for informal texts, prompting studies to incorporate additional resources like images for better performance.

While extensive research has been conducted in NER for languages like English and European languages, Indian languages, including Hindi, have faced challenges due to the lack of annotated data and adherence to annotation standards. The HiNER dataset, introduced by Murthy et al. [2022], fills a crucial void in the development of robust named entity

recognition (NER) systems for Hindi. With over 100,000 annotated sentences and adherence to standard annotation practices, HiNER provides a substantial resource for researchers working on Hindi NER. Its release marks a significant milestone in the field, enabling studies that evaluate and enhance NER models, including transformer-based architectures like XLM-Rlarge. This dataset has catalyzed advancements in Hindi NER research by addressing the longstanding challenge of limited annotated data in Indian languages.

B. Multimodal Named Entity Recognition (MNER):

In response to the rise of image-text formats on social media, recent research has shifted toward multimodal NER. Moon et al. [2018] introduced the concept of MNER for short social media posts, specifically targeting platforms like Snapchat characterized by short text accompanied by images. Their work addressed the challenge of inconsistent syntax and limited textual context by proposing a novel approach that incorporates both textual and visual information. They presented the SnapCaptions dataset, which consists of annotated named entities in Snapchat image-caption pairs. By augmenting traditional NER models with a deep image network and a modality-attention mechanism, Moon et al. demonstrated significant performance improvements over text-only models, highlighting the importance of leveraging visual context in MNER tasks. Lu et al. [2018] presented a visual attention model to find the image regions related to the content of the text. The attention weights of the image regions were computed by a linear projection of the sum of the text query vector and regional visual representations. The extracted visual context features were incorporated into the word-level outputs of the biLSTM model. Zhang et al. [2018] designed an adaptive co-attention network (ACN) layer, which was between the LSTM and CRF layers. The ACN contained a gated multimodal fusion module to learn a fusion vector of the visual and linguistic features. The author designed a filtration gate to determine whether the fusion feature was helpful in improving the tagging accuracy of each token. The output score of the filtration gate was computed by a sigmoid activation function.

Pretrained multimodal BERT models, including VL-BERT (Su et al. [2019]), ViLBERT (Lu et al. [2018]), VisualBERT (Chen et al. [2020]), UNITER (Chen et al. [2020]), LXMERT (Tan and Bansal [2019]), and Unicoder-VL (Chen et al. [2020]), extend the success of BERT in natural language processing to incorporate visual information. These models vary in architecture, visual representations, and pretraining tasks. However, they may face challenges in handling irrelevant visual features and limited object recognition categories. In the context of tweets, the assumption of highly related text-image pairs in caption datasets may not hold true. Pretrained multimodal BERT models integrate textual and visual information through fusion mechanisms.

In tweets and other social media content, a significant proportion of text-image pairs may be irrelevant, meaning the image does not add to the meaning of the text. To address these limitations, Su et al. [2019] proposed RpBERT, a Text-image Relation Propagation-based BERT Model for Multimodal NER. The key innovation of RpBERT lies in its integration of text-image relation propagation into the BERT framework. By incorporating soft or hard gates to select visual clues based on their relevance to the text, RpBERT aims to mitigate the adverse effects of irrelevant visual information on model learning.

III. Research Gap

- Lack of Hindi Dataset Containing Both Image and Text: Currently, there is a dearth of datasets in Hindi that combine both image and text modalities. While there are existing datasets for NER in Hindi text, such as HiNER, datasets that incorporate both textual and visual information are scarce. This absence hinders the development and evaluation of IA-NER models specifically tailored for the Hindi language.
- Limited Exploration of Hindi IA-NER Compared to languages like English, the exploration of IA-NER for Hindi remains relatively unexplored. Most existing research in IA-NER focuses on languages with abundant resources and established datasets, leaving a gap in the exploration of Image Augmented NER techniques for Hindi.
- Lack of Publicly Available Data Annotation Tools for Hindi Another challenge is the lack of publicly available tools and resources for annotating data in Hindi. Annotated datasets are crucial for training and evaluating IA-NER models, but the absence of accessible annotation tools impedes the creation of new datasets and hampers research progress in Hindi IA-NER.

IV. Problem Formulation

A. Problem Identification

There is a significant gap in the development of Hindi NER models with images including. While extensive research exists in both Image augmented for English NER, no established solution currently addresses the task of recognizing named entities within both Hindi text and associated images. This lack of a robust Hindi Image related NER system hinders various applications like image captioning, information retrieval, and visual question answering in Hindi.

B. Proposed Solution

The proposed solution involves a novel approach to Hindi Image Augmented Named Entity Recognition (IA-HiNER) by integrating image captions with textual data. The solution is structured into several key components:

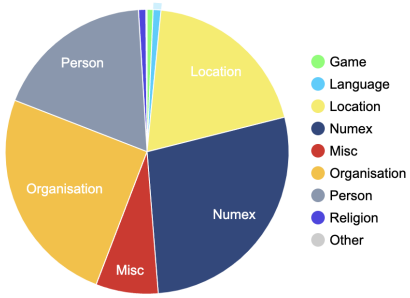
- **Dataset Generation:** - Utilize the News API to gather a diverse dataset of Hindi news articles paired with corresponding images. Adhere to the API's request limitation by strategically querying for relevant articles and images. - Extract captions from the collected images using image captioning techniques. This step enriches the dataset with textual descriptions of the images, augmenting the original textual data.

- **Data Annotation Framework:** - Develop a user-friendly web-based annotation framework tailored for annotating named entities in Hindi text. Enable annotators to mark entities in both the original text and the augmented text generated from image captions. - Implement features for efficient collaboration, version control, and quality assurance to streamline the annotation process.
- **Image Captioning** After gathering the dataset, we further enhanced it by performing image captioning using BLIP. This process generated descriptive textual captions for each image in the dataset, providing additional context for the corresponding textual data.
- **Model Training:** - Utilize pre-trained multilingual models such as XLM (Cross-lingual Language Model) and MuRIL (Multilingual Representations for Indian Languages) for training the Image Augmented Hindi NER model. - Fine-tune the pre-trained models on the annotated dataset, leveraging augmneted data. - Implement transfer learning techniques to adapt the models to the specific characteristics of Hindi text and named entities.
- **Evaluation and Performance Metrics:** - Evaluate the performance of the trained models using standard evaluation metrics for named entity recognition, including precision, recall, and F1-score. - Conduct thorough analyses to assess the impact of incorporating image captions on the IA-NER performance, comparing against baselines that utilize only textual data.
- **Iterative Improvement:** - Iterate on the proposed solution based on insights gained from model evaluations and user feedback. - Continuously refine the data collection, annotation, and model training processes to enhance the robustness and accuracy of the Hindi IA-NER system.

V. Proposed Methodology and Timeline

A. Methodology and Framework

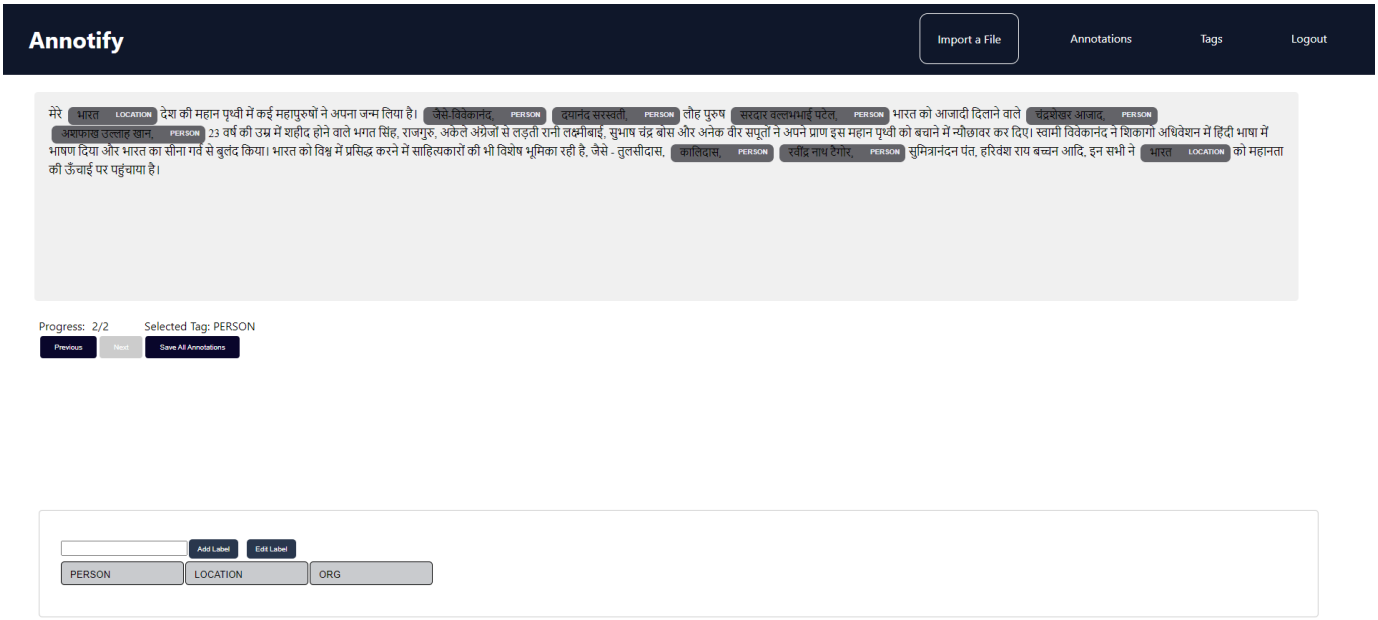
- **Annotation tool**
  To streamline the annotation process, we drew inspiration from the NER tool Textify which exists for only English language. This motivated us to create Annotify, our own solution tailored to handle both Hindi and English text. Leveraging modern technologies, we've utilized React for the frontend interface, making it intuitive and user-friendly. Meanwhile, Node.js with MongoDB powers the backend, ensuring robust performance and scalability. With Annotify, annotating named entities has never been easier or more efficient.



- **Data Collection**
  - Building a comprehensive Hindi dataset integrating both text and images posed challenges due to the absence of existing datasets meeting our criteria.
  - Web scraping proved ineffective due to varying website structures and anti-scraping measures like CAPTCHAs.
  - We turned to API-based data collection, leveraging the News API for efficient access to Hindi news articles.
  - The API's constraint of 100 requests at a time prompted manual gathering and multiple responses generation to meet dataset goals.

Data Split	Number of Samples
Training Data	4356
Validation Data	1090
Test Data	992
Total	6438

- **Data Cleaning**
  - Removing Meaningless Captions:
    - First, we identify and remove captions or text segments that seem meaningless or irrelevant to our dataset's purpose. This includes text containing only symbols, emojis, or non-linguistic characters, as well as phrases with excessive repetition or nonsensical content.
    - We utilize text processing techniques such as regular expressions or rule-based filtering to automatically detect and filter out such captions.
    - Manual inspection is also crucial to ensure the accuracy of the removal process, especially for ambiguous cases.
  - Translating English Words to Hindi:
    - Next, we convert English words or phrases found in the dataset into their Hindi equivalents to maintain consistency and coherence.
    - We use machine translation tools or libraries capable of translating English text to Hindi, ensuring that the translated text accurately reflects the intended meaning.
    - After performing batch translation of English words/phrases within the dataset, we validate the translated text for correctness and appropriateness, as machine translation may sometimes introduce errors.
    - We also consider domain-specific dictionaries or terminology lists to improve translation accuracy, particularly for specialized or technical terms.
- **Data Annotation**
  - Before annotating first we made a python code which took our .txt format file having image link and captions, then that code returned us a .xml file which had basically two columns in one we had image link and other had the corresponding caption of the image.
  - Next we took all the image links in another .xml file and implemented a code which downloads all the images give in the sheet in the increasing order of



NER Tool Screenshot

their row number and simultaneously we named the images with increasing image_id so that we can map the images with their captions.

- Our data annotation process began with an initial attempt at manual annotation through our website. However, the endeavor proved to be exceedingly time-consuming and impractical given the volume of data.
- In response to these challenges, we sought suggestions from panel members and explored semi-automated annotation tools as an alternative approach. Despite our efforts, existing semi-automated tools lacked robust language processing capabilities for Hindi, resulting in inaccuracies and errors during annotation.
- As a solution, we turned to the Hindi Named Entity Recognition (NER) model developed by IIT Bombay. While the model had a limitation of annotating only 120 (but we did 100 only at once for our convinience) sentences per iteration, we devised a strategy to process the dataset iteratively in batches of 120 sentences, ensuring efficient annotation.
- Throughout the process, we took measures to maintain consistency and coherence by converting English words or phrases into their Hindi equivalents. We relied on machine translation tools and libraries to facilitate accurate translation, and subsequently validated the translated text for correctness and appropriateness.
- Additionally, we considered domain-specific dictionaries or terminology lists to enhance translation accuracy, particularly for specialized or technical terms. This systematic approach enabled us to effectively annotate the dataset, laying the groundwork for further analysis and research in the realm of Hindi language processing.

- **Model Development**

  The development of the complete model involves a meticulous process integrating image captioning using the BlipImageCaptioning architecture and named entity recognition (NER) utilizing the XLM-Roberta model.

  1) Image Captioning Model:
     a) Initialization and Configuration:
        The model initialization involves loading the BlipImageCaptioning architecture and configuring it to utilize GPU acceleration if available, ensuring efficient processing. In the absence of a GPU, the model falls back to CPU processing.
     b) Checkpoint and Output File Handling:
        To maintain continuity and facilitate recovery from interruptions, the model manages checkpoint files to store information about the progress of caption generation. Additionally, an output file is maintained to store the resulting captions, ensuring seamless data handling throughout the process.
     c) Data Processing and Image Captioning:
        i) Feature Extraction:
           Visual features are extracted from images using Convolutional Neural Networks (CNNs) integrated into the BlipImageCaptioning architecture. This process captures intricate visual details essential for generating accurate textual descriptions.
        ii) Caption Generation:
           A) Conditional Captioning:
              The model generates captions conditioned on a predefined prompt, such as ”a photography of,” providing contextual cues for generating relevant descriptions.

B) Unconditional Captioning:
Captions are also generated without conditioning, allowing the model to autonomously generate descriptions solely based on the visual content of the images.

iii) Multilingual Translation:
After caption generation, both conditional and unconditional captions are translated into Hindi using the Googletrans library, facilitating cross-linguistic analysis and interpretation.

d) Merging Captions with Annotated Text:
The generated captions, both conditioned and unconditional, are merged with the original annotated text to create a comprehensive dataset. Furthermore, each word in the unconditional captions is annotated with the "B-X" tag, indicating the presence of named entities, thus enriching the dataset for subsequent analysis.

e) Writing Merged Data to Output File:
The merged dataset, comprising image IDs, annotated text, conditional captions, and unconditional captions, is serialized and written to an output file. This file serves as a repository of processed data for further analysis and model refinement.

2) XLM-Roberta Large for Named Entity Recognition (NER) It utilizes a translation-based model architecture, learning shared representations across multiple languages during pretraining.

a) Data Preparation:
Extract image IDs and annotated text from the dataset and load the images for processing. To perform Image Captioning, employ a pre-trained Vision Encoder-Decoder model and generate captions for each image. Annotation process involves augmenting original text with generated captions. Annotate each word in captions for entity recognition (B-X) and store annotated data for further analysis or use.

b) Data Splitting:
The training data is split into training and validation sets using the train_test_split function. This facilitates model evaluation during the training phase, ensuring robust performance on unseen data.

c) Tokenization and Encoding:
Sentences are tokenized using the XLM-Roberta tokenizer, and labels are converted to numerical IDs using label2id mappings. The data is then encoded, establishing alignment between tokens and NER tags, crucial for model training.

d) Model Training: The XLM-Roberta model is initialized for token classification, with model hyperparameters meticulously set. Training commences on the encoded training data, with validation conducted on the validation set to monitor model performance and prevent overfitting.

3) MuRIL Model for Named Entity Recognition (NER):

a) Model Initialization and Configuration:
The MuRIL model, developed by Google Research, is optimized for multilingual text processing tasks like Named Entity Recognition (NER). Pre-trained with extensive language representations, MuRIL is fine-tuned for NER tasks across various languages.

b) Data Preparation and Annotation:
Prior to training, the dataset undergoes meticulous annotation. Each word is labeled with its entity type (e.g., person, organization) to create ground truth data for model training.

c) Data Preparation and Annotation:
Prior to training, the dataset undergoes meticulous annotation. Each word is labeled with its entity type (e.g., person, organization) to create ground truth data for model training.

d) Data Splitting:
The annotated dataset is divided into training and validation sets. This ensures robust model evaluation and prevents overfitting by allowing the model to generalize to unseen data.

e) Tokenization and Encoding:
Text data is tokenized into subwords and encoded into numerical representations suitable for model input. The MuRIL tokenizer handles multilingual text, ensuring accurate alignment between tokens and entity labels.

f) Model Training:
MuRIL is trained using supervised learning techniques on the encoded training data. During training, the model learns to predict entity labels for each token in the input sentences, optimizing its parameters to minimize the loss function.

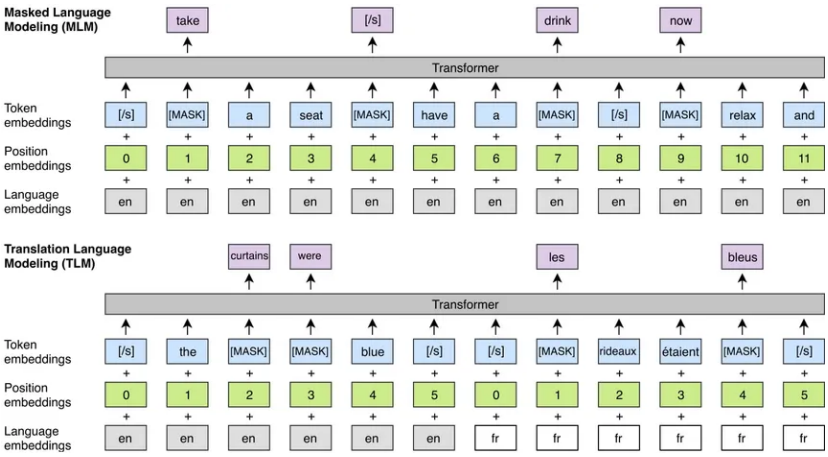
- **Model Architecture**

1) Input Data:
The architecture starts with an input data placeholder, representing the dataset for training the model, which includes images and their corresponding Hindi descriptions or labels.

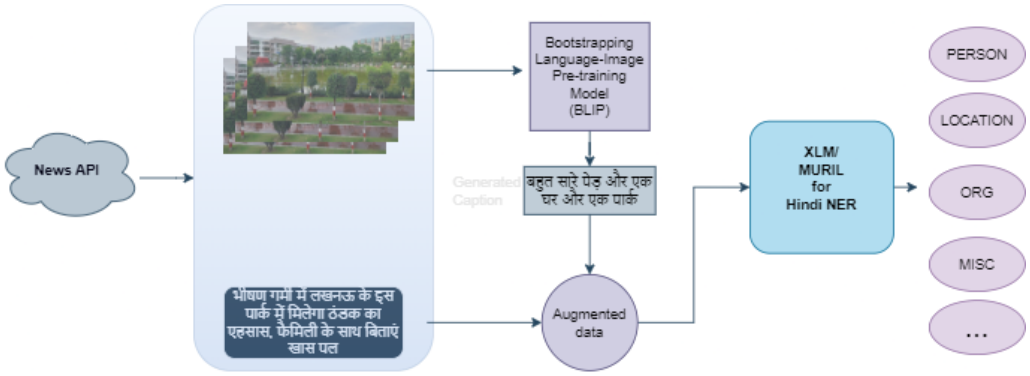
2) Data Augmentation:
The 'Augmented data' node indicates that the input data is enhanced to improve the model's performance, potentially by adding variations to the training data.

3) BLIP:
The 'Bootstrapped Language-Image Pre-training Model (BLIP)' is a central component. It is pre-trained on both language and image data, enabling the model to understand the context of an image alongside textual information.

4) NER Categories:



XLM-Architecture



Architecture of the Image Augmented Hindi NER Model

The 'PERSON', 'LOCATION', 'ORG', and 'MISC' boxes are categories for Named Entity Recognition (NER), allowing the model to identify and categorize entities within the text.

5) XLM-ROBERTa/MURIL:
The XLM-ROBERTa/MURIL for Hindi NER' component is designed for Hindi Named Entity Recognition, utilizing a pre-trained model pre-trained for Indian languages and fine-tuned for NER in Hindi.

This architecture is specifically designed to process data that includes both visual and textual Hindi elements, with a focus on accurately identifying named entities within the text.

- **Training and Evaluation**
  - Train the model on the prepared dataset, using appropriate optimization algorithms and monitoring training progress.
  - Evaluate the model performance on a held-out test set using metrics like precision, recall, and F1-score for named entity recognition.
  - Consider incorporating techniques like dropout and data augmentation to improve model generalizability.
- **Considerations**
  - **Language-specific challenges:** Address potential issues like limited availability of pre-trained Hindi language models and the complex morphology of the Hindi language.
  - **Computational resources:** Consider the computational requirements of training and deploying the model, potentially exploring techniques like model compression or efficient hardware utilization.
  - **Domain adaptation:** If the target domain has specific characteristics, fine-tune the model on additional domain-specific data to improve performance.

VI. Results

Utilizing contextualized word representations, such as those presented by Conneau et al., 2020 and Khanuja et al., 2021, has revolutionized Natural Language Processing (NLP) tasks. In our study, we focus on evaluating the NER performance of two prominent models: XLM-R_{large} and MuRIL. XLM-R_{large}, an extension of the XLM-R_{base} model developed by Facebook AI, is particularly notable for its robustness in handling multilingual data and its ability to capture nuanced contextual information across diverse languages. MuRIL is particularly notable for its pre-training on 17 Indian languages and their transliterated counterparts, making it well-suited for multilingual NLP tasks. We conduct thorough hyperparameter tuning for both models, optimizing batch size and learning rate to maximize F-Score on the development set. Due to GPU memory constraints, we limit our batch size variations to $\{8, 16, 32\}$. Likewise, we explore learning rates within the range

{9e-5, 2e-5, 1e-6, 5e-5, 2e-5} to identify the most effective rate.

Learning Rate	9e-5
Number of Epochs	10
Batch Size	20
Min Batch Size	4
Average Micro F1 Score	0.7412
Learning Rate	2e-5
Number of Epochs	30
Batch Size	16
Min Batch Size	4
Average Micro F1 Score	0.82422
Learning Rate	1e-6
Number of Epochs	20
Batch Size	16
Min Batch Size	4
Average Micro F1 Score	0.6104
Learning Rate	5e-5
Number of Epochs	12
Batch Size	16
Min Batch Size	4
Average Micro F1 Score	0.7282
Learning Rate	2e-5
Number of Epochs	14
Batch Size	16
Min Batch Size	4
Average Micro F1 Score	0.9237

TABLE I
SUMMARY OF HYPERPARAMETERS WITH RESPECTIVE AVERAGE MACRO F1 SCORES ON XLM-R_{LARGE}

Our experiments encompass two dataset variations: one comprising all 11 NER tags on augmented and unaugmented data. Across five runs, we meticulously report our best F1-Score.

Entity	**XLM-R_{large}**	**MuRIL**
Others	90.00	81.17
Festival	00.00	0.00
Game	44.00	40.88
Language	00.00	00.00
Literature	00.00	00.00
Location	68.00	60.89
NUMEX	64.10	57.53
Organization	72.76	63.21
Person	71.00	61.89
Religion	44.00	47.43
TIMEX	5.63	55.34
Micro	80.44	78.11
Macro	36.97	31.46
Weighted	79.25	76.06

TABLE II
TEST-SET F1-SCORE OF XLM-R_{LARGE} AND MURIL ON UNAUGMENTED DATA

From the comparison presented in Table 1 and Table 2, it's evident that XLM-R_{large} demonstrates the highest performance among the models evaluated, followed closely by MuRIL. Interestingly, across both tables, the Festival entity consistently exhibits the lowest performance across all models. This suggests that XLM-R_{large} and MuRIL are more adept at recognizing various entities within the dataset compared to other models, while the Festival entity poses a particular challenge for all models evaluated.

Entity	**XLM-R_{large}**	**MuRIL**
Others	91.00	89.17
Festival	00.00	00.00
Game	58.00	41.88
Language	22.00	00.00
Literature	57.69	00.00
Location	69.86	66.81
NUMEX	66.00	56.31
Organization	72.76	69.26
Person	76.14	68.60
Religion	36.00	50.43
TIMEX	60.00	60.17
Micro	92.00	89.11
Macro	48.97	41.46
Weighted	91.00	88.06

TABLE III
TEST-SET F1-SCORE OF XLM-R_{LARGE} AND MURIL ON AUGMENTED DATA

VII. DISCUSSIONS

The data provides a comprehensive overview of IA-HiNER across various entity types. Firstly, in the "Others" category, which likely represents non-entity tokens, the system demonstrates a high accuracy rate, correctly identifying 8043 instances while only mislabeling 795 tokens. Interestingly, no festivals are identified as named entities according to the data, suggesting potential challenges in recognizing this specific entity type. Conversely, the system performs reasonably well in identifying game-related entities ("B-GAME"), with a relatively low number of incorrect tags (14 out of 45).

However, there are some discrepancies in identifying language-related entities ("B-LANGUAGE"), with 13 incorrect tags and 6 instances missed entirely. On the other hand, literature-related entities ("B-LITERATURE") are identified accurately without any incorrect tags, indicating strong performance in this category. For location entities ("B-LOCATION"), the system demonstrates a decent performance, with 396 correct tags and 162 incorrect tags. Overall, while the system shows varying levels of accuracy across different entity types, further analysis may be required to understand the underlying reasons for these discrepancies and improve the system's performance.

Now, we report the following categories of errors listed.

Label	**Description**
Correct	The gold annotations and the system predictions are the same
Incorrect	The system prediction and the gold annotation don't match
Missing	The system prediction classifies an entity as not a named entity

TABLE IV
SHORT DESCRIPTION OF THE CATEGORIES OF ERRORS

VIII. FUTURE ADVANCEMENTS

1) Expansion of Dataset: Explore diverse sources of Hindi text data beyond news articles, including social media posts, to create a more comprehensive dataset reflecting diverse language usage and context.

Tag	Correct	Incorrect	Missed
Others	8043	795	0
FESTIVAL	0	0	0
GAME	45	14	0
LANGUAGE	23	13	6
LITERATURE	9	0	0
LOCATION	396	162	0
NUMEX	569	334	0
ORGANIZATION	550	184	0
PERSON	402	181	0
RELIGION	3	2	0
TIMEX	239	60	0

TABLE V
DETAILED RESULTS ON IA-HINER DATA FROM XLM-R_{LARGE}

2) Advanced Multimodal Techniques: Investigate innovative methods to leverage both textual and visual information effectively for Named Entity Recognition (NER), aiming to enhance accuracy and robustness.
3) Multimodal Fusion Strategies: Experiment with advanced fusion techniques to integrate information from images and text optimally, prioritizing the grounding of entities detected in images with their corresponding textual representations.
4) Grounded Named Entity: Refine semantic alignment techniques to establish a direct mapping between entities in images and their textual counterparts, improving the accuracy and contextual understanding of entity recognition across multimodal data.

REFERENCES

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faizaan Ahmed, Zhe Gan, Yu Cheng, and Jing Liu. Uniter: Universal imagetext representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv [Preprint]*, 2018. doi: 10.48550/arXiv.1810.04805. URL https://arxiv.org/abs/1810.04805.

James Hammerton. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175, Edmonton, AB, 2003.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, Florence, 2019.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv [Preprint]*, 2016. doi: 10.48550/arXiv.1603.01360.

Yubo Liu, Fandong Meng, Jinchao Zhang, Jiajun Xu, Yu Chen, and Jie Zhou. GCDT: A global context enhanced deep transition architecture for sequence labeling. *arXiv [Preprint]*, 2019. doi: 10.18653/v1/P19-1233. URL https://arxiv.org/abs/1906.02437.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1990–1999, Melbourne, VIC, 2018. Association for Computational Linguistics.

Lingyun Luo, Zhihao Yang, Peng Yang, Yue Zhang, Lu Wang, Hongfei Lin, et al. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34:1381–1388, 2018. doi: 10.1093/bioinformatics/btx761.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity recognition for short social media posts. In *Proc. of NAACL HLT*, pages 852–860, 2018. URL https://aclanthology.org/N18-1078.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. Hiner: A large hindi named entity recognition dataset. *arXiv preprint arXiv:2204.13743*, page 8, 2022. doi: 10.48550/arXiv.2204.13743. URL https://arxiv.org/abs/2204.13743.

Weijie Su, Xiaoyang Zhu, Yue Cao, Bin Li, Lei Lu, Fan Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114, 2019.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. *Proc. AAAI Conf. Artificial Intelligence*, 32 (1), Apr. 2018. doi: 10.1609/aaai.v32i1.11962. URL https://ojs.aaai.org/index.php/AAAI/article/view/11962.