ENHANCING NAMED ENTITY RECOGNITION ON HINER DATASET USING ADVANCED NLP TECHNIQUES

Harshvardhan Pardeshi*[1], Prof. Piyush Pratap Singh[2]

[1]MTech (Data Science), School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

[2]Professor, School of Computer and Systems Sciences, Jawaharlal Nehru University (JNU), New Delhi

*[1]Corresponding author mail ID: harshvardhan.researcher@gmail.com

Abstract

Named entity recognition (NER) is a natural language processing (NLP) categorization labelling job where the objective is to assign words to a predefined set of named entity classes. Although several NER models have been proposed thus far, experts have not yet discovered a workable solution for an inflectional language with limited resources, such as Hindi. . Usually, several manual techniques have been used for entity prediction, such as ambiguity, false positives, and reliance on lengthy annotated data. Nevertheless, these manual methods seem to be ineffectual and time-consuming for acquiring optimum outcomes and are unable to predict all the entity types in the Hindi language. To solve this issue, some researchers have concentrated on NER models. Conversely, it lacks speed and accuracy. Therefore, the present research uses advanced NLP models such as bidirectional encoder representations from transformers (BERT), Distil BERT and the robustly optimized BERT approach (RoBERTA) for effective entity prediction performance. It predicts entities such as person, location, organization and event in the Hindi language. The proposed research uses the Hindi named entity recognition (HiNER) dataset for the purpose of evaluation. The effectiveness of the present model is assessed via several evaluation metrics, such as the F1 score, recall, precision and accuracy, to assess the performance. Furthermore, the comparison of the proposed model reveals the effectiveness of the present research. In conclusion, the projected research envisioned contributing to the emerging NER models, thereby offering an effective entity prediction model for the needed users.

Keywords: Named entity recognition (NER), Natural language processing (NLP), Hindi language, Bidirectional encoder representations from transformers (BERT), Robustly optimized BERT approach (RoBERTA) and Distil BERT.

1. Introduction

Named entity recognition (NER) is a natural language processing (NLP) technique that concentrates on classifying and identifying entities in text, such as the names of locations, organizations, and people [1]. This system is significant for mining structured information for unstructured data text, improving search capabilities and contextual understanding [2]. NER systems play a vital role in processing Hindi text, facilitating better understanding and extraction of information from unstructured data. It helps in constructing knowledge graphs and is widely utilized across different domains, including finance [3] and healthcare [4, 5]. It performs effective data analysis and NLP [6] tasks. It has several drawbacks that could impact its effectiveness, including ambiguity in entity recognition, which relies on length-annotated data [7]. One of the major drawbacks of manual methods is that they are time-consuming and costly to produce, with issues such as informal language and scalability problems as the data volume increases, which affects the accuracy and performance [8]. Therefore, several studies have attempted to overcome the challenges associated with effective entity prediction; however, these studies lack accuracy and computational speed. Recently, research has utilized artificial intelligence (AI)-based technology for entity detection systems [9]. Hence, the advantages of AI and NLP offer various applications in entity prediction. The implementation of NLP in Hindi text prediction is a notable aspect of the platform.

The conventional research has developed 3 neural architectures called the bidirectional long short-term memory-convolutional neural network (BiLSTM-CNN), LSTM and LSTM-CNN on fast text and word2vec embeddings with 85–88% F1 scores. Among these models, the LSTM-CNN outperforms fast text embeddings and achieves an F1 score of 88.3% for Hindi Data [10]. Similarly, the prevailing model has deployed ML technology to guess the next word from the last word sequence in the Hindi language. The process decreases the number of keystrokes of the user by predicting upcoming words. In this ML technology, bidirectional encoder representations from transformers (BERT) model and masked language (ML) [11] representations have been utilized to identify the next word from the previous words. It has attained satisfactory performance [12]. Correspondingly, the prevailing method has revealed a large standard-abiding Hindi dataset. It comprises 2,220,856 tokens and 109,146 sentences, annotated along with 11 tags. The tag-set statistics in the dataset have a healthy distribution, especially for crucial classes such as location, organization and person. The dataset helps to attain an 88.78% F1 score for all tags and

a better F1 score while collapsing the tag set [13]. The conventional model presents an NER system for Indian languages such as Marathi and Hindi. Various variations of BERT, such as AlBERT, RoBERTa and baseBERT, are considered. Three available Marathi and Hindi datasets are evaluated. Xlm-RoBERTA [14] has achieved F1 scores of 75.90% and 83.4% on the IJCNLP 200 [11] and WikiAnn Hindi datasets, whereas the prevailing model MahaRoBERTA has achieved F1 scores of 64.34% and 88.90% on the IIT Bombay and WikiAnn Marathi datasets, respectively [11].

To solve these issues, the proposed model uses a certain set of procedures to improve the prediction performance through the identification of entities. They are Person, Location, Organization, and Event. The proposed model uses a publicly available dataset, namely, HiNER. Once the input data are loaded, it processes the preprocessing technique to perform tokenization and alignment. Then, the preprocessed data are divided into two parts at a ratio of 80:20, where 80% is used for the training process and 20% is used for the testing process. The training set is passed to the proposed advanced NLP models of BERT, Distil BERT and RoBERTA to perform entity prediction. Then, the testing data are tested by the prediction model and evaluated through performance metrics. The major contributions of the present NLP model are provided below.

- In the present research, an NLP model for entity prediction with the HiNER dataset is used to increase the computational complexity and accuracy.
- BERT, Distil BERT and RoBERTA are used to perform entity prediction effectively for the HiNER dataset.
- To estimate the effectiveness of the present model, several evaluation metrics, such as the F1 score, accuracy, precision and recall, are used.

1.1. Paper Organization

The flow of the present model is given here: section 1 provides an overview of the background of the proposed model. Section 2 reviews the conventional literature related to entity detection and problem identification. Section 3 precisely describes the proposed methodology. Furthermore, section 4 provides a table and graphical representation of the data analysis. Section 5 discusses the results of the proposed DL model with those of traditional methods. Finally, section 6 concludes the present model along with future research.

2. Literature Review

The present section provides an analysis of various existing studies on entity prediction along with other techniques. For example, the conventional model has developed a DL architecture for NER that is based on the resource-scarce language Hindi. Conditional random field (CRF) [15] and Bi-LSTM are based on CNNs. The prevailing model has been evaluated on an NER-tagged standard corpus comprising 4478 Hindi sentences. Experimental outcomes have shown that the integration of prevailing models performed better than performing individual models did [16]. Similarly, MuRIL [17] has been utilized to extract CRF and deep features for named entity tagging by the prevailing model. It has been tested on 2 datasets: the Hindi dataset and the Bengali NER dataset. It achieves an 89.79% F1 score on the Hindi dataset. In the Bengali dataset-1 and dataset-2, 75.02% and 84.19% of the F1 scores, respectively, were obtained. Similarly, the prevailing model performs NER tasks on Hindi through incorporating a multilingual language model called the Multilingual Representation for the Indian Language (MuRIL). It has utilized the ICON 2013 Hindi NER dataset, which has attained 87.89%, 83.74% and 85.77% precision, recall and F1 scores, respectively [18]. Correspondingly, the prevailing model has presented an NER method for identifying entities in weather texts in the Hindi language. It is a twofold technique: one involves collecting data from forecasting weather websites, and the second involves applying an ML algorithm with Hindi language vector representation to gathered data for model training [19].

Similarly, the prevailing model has developed an NR system that is involved in DL and word embedding. It has developed a CNN and bidirectional gated recurrent unit (Bi-GRU) based on bilingual NER built on enhanced word embedding (EWE). It has utilized IJCNLP-08 NERSSEAL shared task corpora comprising an annotated dataset and one in Punjabi. The experimental results revealed 90.70%, 92.60% and 91.64% recall, precision and F1 score, respectively [20]. Concomitantly, the considered research presented a DL architecture based on CNN and BiLSTM networks. It has been estimated on NER-tagged standard corpus data that comprises 4478 Hindi sentences offered through TDIL, which have attained 61%, 56% and 58% precision, recall and F1 measures, respectively [21]. The traditional method has devised an architecture of a residual network called BiLSTM with fast text word embedding layers. It performs the labelling task on named entities in the dataset. It was released during ICJNLP 2008, which is part of a workshop on NER. It has achieved 69.5% and 96.8% F1 scores and accuracy and 81.9% and 60.4% precision and recall, respectively [22]. In contrast, the prevailing research has described the NER system for Hindi, which utilizes 2 methodologies. These include the context pattern-based

MENE (CP-MENE) and maximum entropy-based named-entity (BL-MENE) model frameworks. It has been focused on the Hindi Health Data (HHD) corpus, which has achieved better performance [23].

Similarly, the prevailing work has set a DL architecture to solve the issue of identifying the named entities in Hindi phrases. It performs recursive BiLSTM [24], which includes a denoising autoencoder with conditioning logic. Experiments have been conducted to analyse individual batch sizes and word embedding behavior, which is significant for DL model training and results in better performance [25]. Correspondingly, the prevailing model employs a DL NR model to recognize the named entities in Hindi, bilingual Hindi, Punjabi and bilingual Punjabi texts. The experimental outcomes have shown that NER system development can extract named entities not only from the Punjabi and Hindi datasets but also from bilingual texts [26]. Similarly, conventional methods address numerous features utilized for named entity recognition. Indian languages such as Malayalam, Telugu, and Tamil and Indo-Aryan languages such as Bengali, Punjabi, Marathi and Hindi are utilized in this prevailing model. It uses a corpus dataset from the FIRE 2013 NER shared task and cross-lingual information access (CLIA), which achieve 80.12% and 83.1% precision and recall in the Hindi language, respectively [27]. Similarly, the existing model has developed a Hindi word classification model, which has suggested several methods based on collective techniques such as AdaBoost, random forest (RF), decision tree (DT) and forecasts. It has evaluated imbalanced and balanced datasets. Compared with other ML models, the prevailing BagBoost model provides accuracies, precisions, recalls and F1 scores of 68%, 67%, 66% and 65%, respectively [6].

2.1. Problem identification

This section describes several conventional methods that have been limited by their ability to predict entities that have several limitations.

- In the prevailing research, the sizes of the testing and training datasets are small, and further enhancement is needed for better performance [13, 21].
- The accuracy and F1 score are the significant metrics used to measure the efficiency of the model. However, several existing studies lack accuracy and F1 scores [10, 18, 20-22].

3. Proposed Methodology

Named entity recognition (NER) is a natural language processing (NLP) technique that concentrates on classifying and identifying entities in text, such as the names of locations, organizations, and people. Therefore, the prediction of entities includes ambiguity and false positives, which often lead to delays and are considered a time-consuming process. To overcome this issue, several traditional models have focused on NER models; however, they lack accuracy, speed and computation. To overcome this issue, the proposed model employs advanced NLP techniques such as BERT, Distil BERT and RoBERTA to perform entity prediction. Figure 1 displays the overall design.

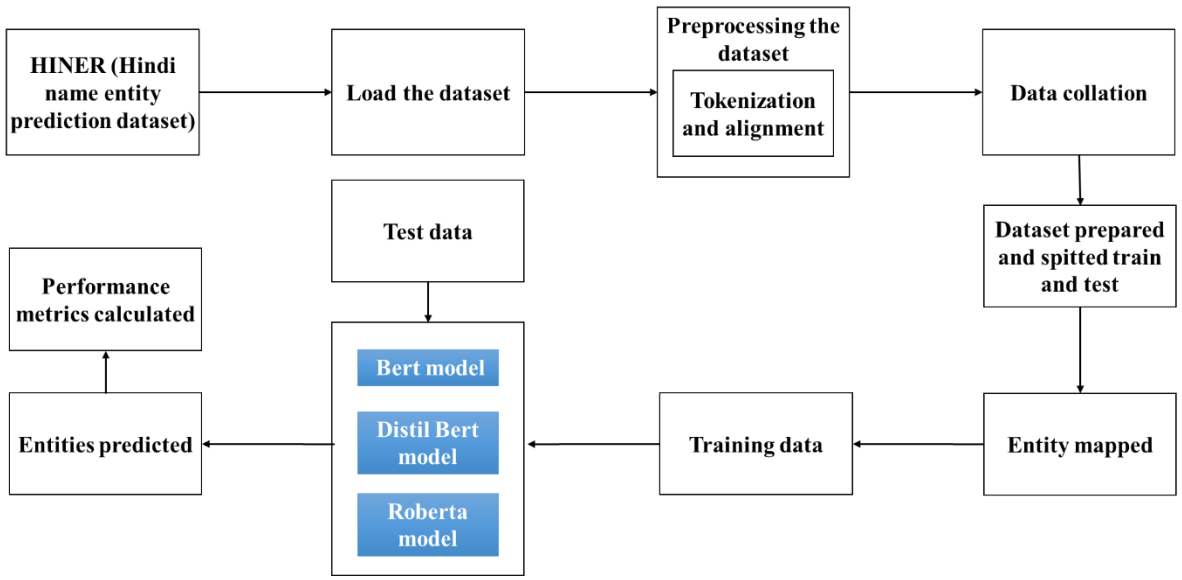

Figure 1. Overall Method

Figure 1 shows the overall method of the proposed prediction model. It comprises data loading, pre-processing of data and data collation, splitting of data and, eventually, a prediction process with BERT, Distil BERT and

RoBERTA models and performance metrics to find the results. The following sections discuss the functions used in the present model.

3.1. Dataset Description

The proposed model uses the Hindi named entity recognition (HiNER) dataset, which was created for conventional NLP tasks of NER for the Hindi language at the IIT Bombay and CFILT laboratories. The dataset is gathered through numerous government information websites, and these sentences are annotated manually as part of the data gathering technique. This repository comprises the HiNER published in 2022 at the Language Resources and Evaluation conference (LREC). The HiNER dataset is annotated manually through a single annotator with a long run time. Some of the examples are listed as follows.

Cultural Entities encompass a range of elements, including Festivals, which are the names of cultural celebrations such as "दीवाली" (Diwali) and "होली" (Holi). Additionally, Traditions fall under this category, representing specific customs or rituals like "शादी" (wedding) and "रक्षा बंधन" (Raksha Bandhan). Scientific Entities include Scientific Terms, which are the names of theories, laws, or scientific phenomena, for instance, "गुरुत्वाकर्षण" (gravity) and "अणु" (atom). Biological Entities also belong here, referring to the names of species or organisms such as "तितली" (butterfly) and "पेड़" (tree).Artistic Entities comprise Movies and TV Shows, which are the titles of films or television series, for example, "शोले" (Sholay) and "कुंडली भाग्य" (Kundali Bhagya). Books and Literature are also included, representing the titles of literary works like "गुनाहों का देवता" (Gunahon Ka Devta) and "चाणक्य नीति" (Chanakya Neeti).

Sports Entities consist of Sports Events, which are the names of specific sporting competitions or tournaments, such as "क्रिकेट वर्ल्ड कप" (Cricket World Cup) and "ओलंपिक" (Olympics). Furthermore, Athletes, referring to the names of famous sports personalities like "सचिन तेंदुलकर" (Sachin Tendulkar) and "पीवी सिंधु" (P.V. Sindhu), are part of this category. Geopolitical Entities include Regions and Territories, which are specific geographical areas not classified as standard locations, for example, "कश्मीर" (Kashmir) and "पश्चिम बंगाल" (West Bengal). Historical Sites, representing the names of significant historical places like "ताज महल" (Taj Mahal) and "कुतुब मीनार" (Qutub Minar), also fall under this category.

Technological Entities encompass Brands and Products, which are the names of technology companies or their offerings, such as "गूगल" (Google) and "सैमसंग" (Samsung). Software and Applications are also included, referring to the names of software programs or mobile apps like "व्हाट्सएप" (WhatsApp) and "फेसबुक" (Facebook). Economic Entities comprise Companies and Corporations, which are the names of businesses and organizations not classified under a general organization tag, for example, "टाटा" (Tata) and "इन्फोसिस" (Infosys). Financial Instruments, representing the names of stocks, bonds, or other financial tools like "बीएसई" (BSE) and "निफ्टी" (Nifty), are also part of this category.

The official dataset link is provided below.

Dataset Link: ***https://huggingface.co/datasets/cfilt/HiNER-original***

3.2. Data Preprocessing

Pre-processing the data is vital for converting the raw input data into the required data. The present research utilizes tokenization and alignment techniques, which are crucial in entity prediction models. The proposed models use word piece tokenization, which is split into subwords for handling complex words effectively and improves the context understanding in entity prediction performance. Fine-tuning of the labelled dataset permits them to adjust certain prediction tasks, increasing accuracy through pattern learning in word–entity associations. This combination of advanced alignment and tokenization methods enables the proposed models to predict efficiently within numerous contexts, making them potent tools in NLP applications.

3.3. Data splitting

In deep learning, this technique is used to eliminate the data overfitting issue. Essentially, DL uses the data splitting method to train the respective research where the training data are given to the proposed research for equipping the training stage parameters. After the training process, the test set data are measured to calculate the present model for handling the observations. In the present model, the original data are split into 2 sets at a ratio

of 80:20. Eighty percent of the new observations are utilized for training, and the remaining 20 percent of the data are applied for the testing process to calculate the performance of the present research.

3.4. Proposed Entities Prediction- Bert, Distil BERT and Roberta

To enhance entity annotation, the present model leverages the power of three state-of-the-art language models: BERT, Distil BERT, and RoBERTa. These models provide advanced capabilities that can considerably improve the quality of entity annotation.

3.4.1. BERT Model

The BERT model is a dominant language model that reforms the field of NER [1]. By leveraging its attention mechanisms and bidirectional context representation, BERT captures the dependencies and nuances within text, enabling highly accurate classification and identification of named entities. Furthermore, it considers both the right and left context of every word in a sentence, offering a more inclusive understanding of the entity's role and significance within the particular context. BERT is considered to first be pretrained on large amounts of unlabelled data to learn common language representations. Then, it can be fine-tuned on a moderately small amount of labelled data for particular NER tasks, making it highly effective and efficient. For the NER system, a token classification process is added on top of the BERT base architecture as in the data preprocessing technique. This allows the model to make predictions at the token level, classifying each token as a specific entity type (e.g., person, organization, location and event). Correspondingly, the application of the BERT model in Hindi NER involves fine-tuning to pretrain on labelled datasets specific to entity prediction tasks. The process typically comprises the following: Models learn contextual embeddings from large amounts of unlabelled text. The model parameters were adjusted via labelled data to optimize the performance of NER.

3.4.2. Distil BERT

The Distil BERT is a distilled version of the BERT model [2]. It is designed to be faster and smaller when maintaining a high level of performance. It is particularly effective for tasks such as NER, where the goal is to classify and identify entities into predefined categories such as persons, events, officialdoms and locations. The performance of Distil BERT is meticulously tied to the utilized HiNER dataset, which could struggle with texts that vary crucially from this training set, leading to potential misclassifications. The proposed Distil BERT model might address issues with complex entities or sentences that require deeper appropriate understanding beyond what was captured during training. Several processes are needed to fine-tune Distil BERT on the HiNER dataset for Hindi NER, including loading the HiNER dataset and tokenizing and aligning the input text to preprocess the data. Then, the trained proposed Distil BERT model is loaded and fine-tuned via the preprocessed HiNER dataset. Finally, the performance of the fine-tuned model is tested on a held-out test set.

3.4.3. RoBERTa

RoBERTa is a powerful variant of the BERT model that excels in various NLP tasks, including NER [3]. When applied to the HiNER dataset, a comprehensive Hindi NER resource, it could crucially improve entity recognition capabilities. The dataset consists of over 2 million tokens annotated with 11 distinct entity tags, making it an outstanding training ground for fine-tuning RoBERTa. The proposed RoBERTa model improved the pretraining methodology—removing the next sentence prediction and employing dynamic masking—enables it to learn context more efficiently. Fine-tuning the RoBERTa model on the dataset involves loading the dataset utilizing libraries such as Hugging Face, preprocessing the data, and training the model on labelled samples. A technique that shows promise for improving Hindi NER applications is to use RoBERTa with the HiNER dataset to handle Hindi language data more accurately and efficiently. Figure 2 depicts the proposed architecture for effective entity prediction.

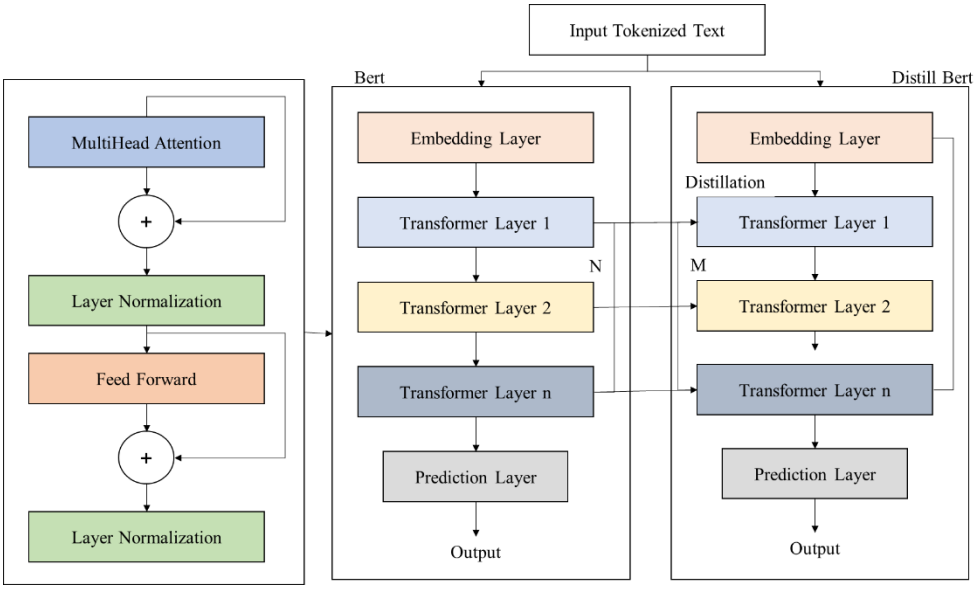

Figure 2. Proposed Architecture

Figure 2 exemplifies the proposed architecture of the present research.

The proposed Distil BERT model learns the context for each word by utilizing the transformer encoder. The contextual embeddings are produced by the transformer encoder via a self-attention mechanism. To represent the semantic information, the recovered contextual embeddings for every word are concatenated into a single vector. Pretraining is used to learn the representation via Distil BERT. To convert the input vectors into contextual embedding vectors—one per word—we use a multilayer bidirectional transformer encoder. As shown in Figure 2, Distil BERT uses knowledge distillation to reduce the BERT base model (bert-base-uncased) parameters by 40%, resulting in a 60% faster inference. The primary goal of distillation is to use a smaller model, such as Distil BERT, to approximate the whole output distributions of the BERT model. The accumulation is carried out locally utilizing the gradients from several minibatches prior to updating the parameters in each step. The dynamic masking employed during inference replaces the static masking utilized in BERT. Formally, an input is defined in equation (1):

$$Y = (y_1, y_2 \dots y_i, \dots \dots, y_l) \tag{1}$$

where $y_i \in R^{1\times d}$ is known as the embedding construction by adding the i^{th} token via the parallel token, position embedding and segment; d refers to the hidden layer's maximum embedding dimension; and l represents the maximum input sequence length. In layer j, the pair of short text embeddings is denoted in equation (2):

$$q^{(j)} = (q_1, q_2 \dots q_i, \dots, q_l) \tag{2}$$

where $q_i \in R^{1\times d}$ coincides with the parallel y_i dimension. $q^{(j)}$ is acquired through equations (3) to (10):

$$r = q^{j-1} \tag{3}$$

$$H_i(r) = softmax\left(\frac{(rp_i^R)(rp_i^Q)^T}{\sqrt{(\frac{d}{h})}}\right)(rp_i^V) \tag{4}$$

$$MultiHead(r) = Concat(H_1, \dots, H_i, \dots, H_h)p^O \tag{5}$$

$$LN(y_i) = \propto * \frac{y_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{6}$$

$$S = LN(r + MultiHead(r)) \tag{7}$$

$$RELU(y) = \{x\ ,\ x > 0\ 0,\ x \leq 0\} \tag{8}$$

$$FCFN(S) = ReLU(Sp_1 + b_1)p_2 + b_2 \tag{9}$$

$$q^{(j)} == LN(S + FCFN(S)) \quad (10)$$

where $p_i^R, p_i^Q and\ p_i^v \in R^{d\times\frac{d}{h}}, p^O \in R^{d\times d}, p_1 \in R^{d\times h-s}\ and\ p_2 \in R^{h-s\times d}$, $b_1 \in R^{1\times h-s}$ and $b_2 \in R^{1\times d}$ are the matrices of learnable parameters. h refers to the attention heads, and $h-s$ refers to the size of the hidden layer in the network. LN represents layer normalization, y_i represents i^{th} sample data of the input data, and $\sigma^2 and\ \mu$ represents the variance and mean of the feature vectors for each sample. ε represents the decimal, and $\beta\ and\ \propto$ represents the translation and scaling parameters, respectively. Likewise, table-1 shows the hyperparameter setting for the proposed frameworks.

Table-1 Hyperparameter and its value

Hyperparameter	Value
Learning Rate	1e-5
Batch Size	16
Number of Epochs	3
Weight Decay	0.01

The table-1 outlines the key hyperparameters used for training the models such as BERT, Distil and Roberta model and their corresponding values. A learning rate of 1e-5, or 0.00001, was employed, indicating a small step size during the optimization process, which can lead to finer adjustments in the model's weights and potentially better convergence. The batch size was set to 16, meaning the model processed 16 data samples at a time during each update step. The training was conducted for a limited duration of 3 epochs, representing the number of times the entire training dataset was passed through the model. Finally, a weight decay of 0.01 was applied as a regularization technique to prevent overfitting by penalizing large weight values. These hyperparameter choices collectively define the training configuration of the model.

4. Results and Discussion

This division analyses and deliberates the outcomes of the proposed system that is accomplished to ensure effective entity prediction in the Hindi language.

4.1. Exploratory Data Analysis

This part provides the EDA of the HiNER dataset used in the proposed model. It is used to understand and visualize the data in the HiNER dataset. Figure 3 represents various entities in the dataset of the proposed model.

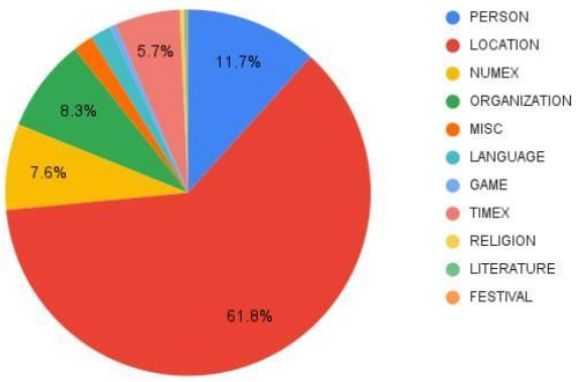

Figure 3. Various Entities in the Dataset

The above pie chart (Figure 3) gives a categorical breakdown of NER labels by percentage. The largest category is "LOCATION" (61.8%), meaning that the dataset predominantly contains locations as named entities. Other notable categories include "PERSON" (11.7%), "NUMEX" (7.6%), and "ORGANIZATION" (8.3%), while the

rest of the categories make up smaller portions. This analysis suggests a dataset with an imbalance in NER labels, which is dominated primarily by "location" entities and a large number of shorter sentences.

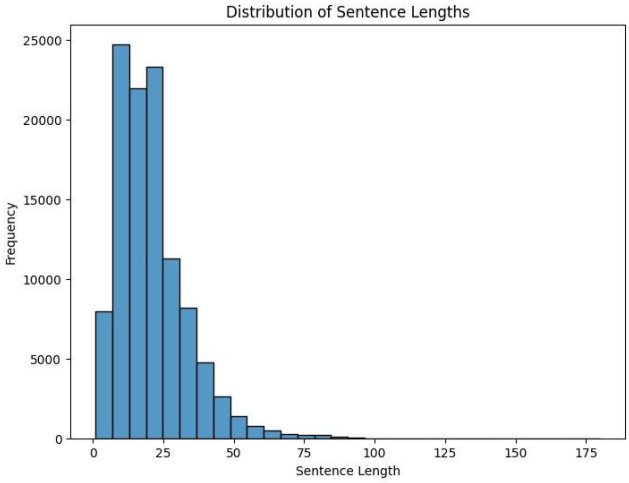

Figure 4. Distribution of sentence lengths

Figure 4 displays the distribution of sentence lengths in the dataset. Most sentences have lengths between 10 and 25 words, with the highest frequency being approximately 15–20 words. As sentence length increases beyond 50 words, the frequency decreases sharply, indicating that long sentences are rare in the utilized dataset.

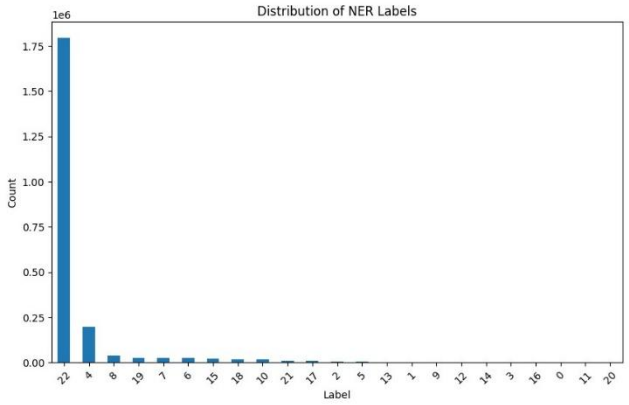

Figure 5. Distribution of NER Labels

Figure 5 signifies the distribution of various NER labels by their frequencies. The most frequent label (Label 22) dominates the dataset, suggesting that this entity type occurs much more often than the others. A few other labels (such as 8, 4, and 19) appear significantly less common but still have perceptible representations. The rest of the labels occur very infrequently. This distribution designates an imbalance in the entity types, with a majority of one or a few labels.

4.2. Performance Metrics

Performance metrics are primarily used for detecting the efficiency of the projected research by utilizing various metrics, such as the recall rate, precision, accuracy and F1 score value.

1. Recall Metric

The recall metric is used to analyse the percentage of data that are appropriately predicted in the present model. The method is expressed in equation (11):

$$Recall = \frac{True\_Pos}{False\_Neg+True\_Pos} \tag{11}$$

2. Accuracy

Accuracy is a significant metric utilized to examine the number of estimates that show the efficiency of the proposed model. The accuracy formula is demonstrated in equation (12),

$$Accuracy\ Metric = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

FP, TN, TP, FN are false positive, true negative, true positive and false negative.

3. F1 score

The F1 score is a metric that is utilized to evaluate the predictions in the respective model, which are generated for the positive class. The f1 measure is calculated through equation (13):

$$F1\ Score\ Metric = \ 2 * \frac{R*P}{R+P} \quad (13)$$

R and P refer to the recall and precision, respectively.

4. Precision

Precision is a metric that is presented as the covariance unit of the method. It is analysed via exact predictable cases. (TP)The entire group of cases that were considered $precisely$ $(TP + FP)$. It encompasses the producibility and repeatability of the capital. Equation (14) describes the precision formula.

$$Precision\ Metric = \frac{TP}{FP+TP} \quad (14)$$

4.3 Performance Analysis

This section includes an internal comparison of the proposed advanced NLP models such as BERT, Distil BERT and the RoBERTa model. The table. Figure 1 and Figure 6 compare the RoBERTa model on the basis of the HiNER data entities.

Table 1. Comparison of RoBERTa

Entity	Precision metric	Recall metric	F1-Score metric
Game	0.8	0.84	0.82
Language	0.9	0.92	0.91
Literature	0.72	0.53	0.61
Location	0.94	0.93	0.94
Misc	0.65	0.7	0.68
Numex	0.72	0.7	0.71
Organization	0.73	0.8	0.77
Festival	0.72	0.69	0.73
Person	0.79	0.82	0.75
Religion	0.66	0.86	0.75
Timex	0.8	0.84	0.82

Both Table 1 and Figure 6 illustrate the comparison of the RoBERTa model on various entities on the HiNER dataset. The RoBERTa model attained precision values of 0.8, 0.9, 0.72, 0.94, 0.65, 0.72, 0.73, 0.72, 0.79, 0.66 and 0.8. Similarly, it attained 0.84, 0.92, 0.53, 0.93, 0.7, 0.7, 0.8, 0.69, 0.82, 0.86 and 0.84 of recall and 0.82, 0.91, 0.61, 0.94, 0.68, 0.71, 0.77, 0.73, 0.75, 0.75 and 0.82 of F1 score on Game, Language, Literature, Location, Misc, Numex, Organization, Festival, Person, Religion and Timex, respectively.

Table 2. Comparison of Distil BERT

Entity	Precision metric	Recall metric	F1-Score metric
Game	0.8	0.84	0.82
Language	0.89	0.9	0.89

Literature	0.5	0.31	0.39
Location	0.91	0.91	0.92
Misc	0.64	0.65	0.68
Numex	0.66	0.63	0.64
Organization	0.64	0.65	0.64
Festival	0.71	0.7	0.73
Person	0.72	0.7	0.71
Religion	0.69	0.61	0.65
Timex	0.73	0.75	0.74

Both Table 2 and Figure 7 depict the comparison of the Distil BERT model. The F1 scores were 0.8, 0.89, 0.5, 0.91, 0.64, 0.66, 0.64, 0.71, 0.72, 0.69 and 0.73 for precision; 0.84, 0.9, 0.31, 0.91, 0.65, 0.63, 0.65, 0.7, 0.7, 0.61, and 0.75 for recall; and 0.82, 0.89, 0.39, 0.92, 0.68, 0.64, 0.64, 0.73, 0.71, 0.65 and 0.74 for game, language, literature, location, Misc, Numex, Organization, Festival, Person, Religion and Timex, respectively.

Table 3. Comparison of BERT

Entity	Precision metric	Recall metric	F1-Score metric
Game	0.57	0.73	0.64
Language	0.91	0.92	0.91
Literature	0.63	0.67	0.65
Location	0.95	0.94	0.95
Misc	0.66	0.77	0.71
Numex	0.72	0.69	0.71
Organization	0.77	0.82	0.79
Festival	0.43	0.17	0.25
Person	0.83	0.86	0.85
Religion	0.67	0.87	0.76
Timex	0.83	0.84	0.83

Figure 8 and Table 3 show the comparison of the BERT model in the HiNER dataset. The F1 scores were 0.64, 0.91, 0.65, 0.95, 0.71, 0.71, 0.71, 0.25, 0.85, 0.76 and 0.83; 0.73, 0.92, 0.67, 0.94, 0.77, 0.69, 0.82, 0.17, 0.86, 0.87 and 0.84 for recall; and 0.57, 0.91, 0.63, 0.95, 0.66, 0.72, 0.77, 0.43, 0.83, 0.67 and 0.83 for precision on Game, Language, Literature, Location, Misc, Numex, Organization, Festival, Person, Religion and Timex, respectively. Table 4 describes the overall comparison of the proposed NLP models.

Table 4. Overall Comparison

Model	Accuracy Metric	Precision metric	Recall metric	F1-Score metric
BERT	0.97	0.89	0.89	0.9
RoBERTa	0.96	0.87	0.88	0.88
Distil BERT	0.95	0.83	0.82	0.83

Table 4 depicts the comparison. The proposed BERT model achieves better results than other proposed models, such as RoBERTa and Distil BERT. The BERT model achieved accuracies, precisions, recalls and F1 scores of 0.97, 0.89, 0.89 and 0.9, respectively. The accuracy, precision, and recall and F1 scores of the RoBERTa and Distil BERT models were 0.96, 0.87, 0.88 and 0.88, respectively, and the accuracy, precision, recall and F1 score were 0.95, 0.83, 0.82 and 0.83, respectively.

Figure 9 shows an overall comparison of all 3 models, such as BERT, Distil BERT and RoBERTa. Among these models, BERT achieves better results than the other 2 models do. It attained 1% and 2% accuracy, 2% and 6% precision, 1% and 7% recall and 2% and 6% F1 scores greater than those of the RoBERTa and Distil BERT models, respectively. This shows that the proposed BERT model is efficient in the entity prediction system.

4.4 Comparative Analysis

Comparative analysis is demonstrated in the present section. Table shows the comparative analysis of different models.

Table 5. Comparative Analysis [4]

Models	Precision	Recall	F1 score
Existing Model	0.87	0.83	0.85
BERT	0.89	0.89	0.90
RoBERTa	0.87	0.88	0.88
Distil BERT	0.83	0.82	0.83

This table-5 presents the performance of four different language models MuRIL, BERT, RoBERTa, and DistilBERT based on their precision, recall, and F1 score. BERT demonstrates the strongest overall performance with the highest precision (0.89), recall (0.89), and F1 score (0.90). RoBERTa also shows competitive results with a precision of 0.87, recall of 0.88, and an F1 score of 0.88. MuRIL, a model specifically designed for Indian languages, achieves a precision of 0.87, recall of 0.83, and an F1 score of 0.85. DistilBERT, a version of BERT, exhibits the lowest scores among the four models, with a precision of 0.83, recall of 0.82, and an F1 score of 0.83.

Table 6. Comparative Analysis [5]

Models	Precision	Recall	F1 score
RNN–baseline	0.57	0.50	0.53
RNN-CNN	0.63	0.46	0.53
RNN–CRF	0.65	0.50	0.57
RNN–CNN–CRF	0.54	0.56	0.55
LSTM–baseline	0.64	0.48	0.55
LSTM–CNN	0.63	0.50	0.56
LSTM-CRF	0.61	0.56	0.58
LSTM–CNN–CRF	0.56	0.58	0.57
Bi-LSTM–baseline	0.66	0.54	0.60
Bi-LSTM–CNN	0.70	0.53	0.60
Bi-LSTM–CRF	0.67	0.58	0.62
Bi-LSTM–CNN–CRF	0.73	0.57	0.65
BERT	0.89	0.89	0.90
RoBERTa	0.87	0.88	0.88
Distil BERT	0.83	0.82	0.83

Likewise table-6 compares the performance of various neural network architectures, including RNN, LSTM, and Bi-LSTM variants with and without CNN and CRF layers, against transformer-based models like BERT, RoBERTa, and DistilBERT on the HiNER dataset. Among the sequential models, Bi-LSTM-CNN-CRF achieves the highest F1 score of 0.65, demonstrating the benefit of combining bidirectional long short-term memory networks with convolutional neural networks and a conditional random field layer. However, the transformer-based models significantly outperform these architectures, with BERT achieving the top scores across all metrics (0.89 precision, 0.89 recall, and 0.90 F1 score), followed closely by RoBERTa (0.87 precision, 0.88 recall, and 0.88 F1 score) and DistilBERT (0.83 precision, 0.82 recall, and 0.83 F1 score). This indicates that for the HiNER task, transformer models exhibit a substantial advantage over traditional recurrent neural network-based approaches.

4.5 Ablation Study

Table 7. Ablation study for BERT model

Hyperparameter Change	Precision	Recall	F1 Score
Learning Rate = 0.001	0.86	0.85	0.88
Learning Rate = 0.0001	0.87	0.86	0.87
Learning Rate = 0.00001	0.89	0.89	0.90
Batch Size = 8	0.86	0.81	0.83
Batch Size = 16	0.87	0.82	0.84
Batch Size = 32	0.89	0.89	0.90

The ablation study on the BERT model reveals the impact of learning rate and batch size on performance metrics. Decreasing the learning rate from 0.001 to 0.00001 generally improved precision, recall, and F1 score, with the lowest learning rate yielding the best results (0.89, 0.89, 0.90 respectively), suggesting finer weight updates enhance learning. Similarly, increasing the batch size from 8 to 32 showed a positive trend in performance, with a batch size of 32 achieving the highest precision (0.89), recall (0.89), and F1 score (0.90), indicating more stable gradient estimation aids training. Notably, an F1 score of 0.90 was achieved with both the lowest learning rate and the largest batch size tested, highlighting the critical role of hyperparameter tuning in optimizing the BERT model's performance for the entity recognition.

Table-8 Ablation study for Distil BERT model

Hyperparameter Change	Precision	Recall	F1 Score
Learning Rate = 0.001	0.79	0.80	0.82
Learning Rate = 0.0001	0.80	0.81	0.81
Learning Rate = 0.00001	0.87	0.88	0.88
Batch Size = 8	0.81	0.81	0.82
Batch Size = 16	0.87	0.88	0.88
Batch Size = 32	0.80	0.76	0.79

This ablation study on hyperparameter tuning for entity recognition using the HiNER dataset in Table 8 reveals the impact of varying learning rates and batch sizes on the model's performance, as measured by precision, recall, and F1 score. When examining the learning rate, a clear trend emerges: decreasing the learning rate generally leads to improved performance. The lowest learning rate tested, 0.00001, yielded the highest scores across all metrics, achieving a precision of 0.87, a recall of 0.88, and an F1 score of 0.88. This suggests that for effective entity recognition on the HiNER dataset with this model, a smaller learning rate allows for more stable and precise weight updates during training. Regarding batch size, the results indicate that a batch size of 16 resulted in the best performance, matching the F1 score achieved with the lowest learning rate (0.88 for precision and recall as well). Smaller (8) and larger (32) batch sizes led to comparatively lower F1 scores, suggesting that a batch size of 16 provides a good balance for learning the patterns in the HiNER dataset for this specific entity recognition task. The study underscores the importance of careful hyperparameter tuning, as different values can significantly influence the model's ability to accurately identify and classify entities within the HiNER dataset.

Table-9 Ablation study for RoBERTa model

Hyperparameter Change	Precision	Recall	F1 Score
Learning Rate = 0.001	0.79	0.80	0.81
Learning Rate = 0.0001	0.82	0.80	0.77
Learning Rate = 0.00001	0.83	0.82	0.83
Batch Size = 8	0.82	0.80	0.79
Batch Size = 16	0.83	0.82	0.83
Batch Size = 32	0.80	0.79	0.78

This ablation study in table 9 explores the impact of varying learning rates and batch sizes on a model's performance, as evaluated by precision, recall, and F1 score. Examining the learning rate, the results suggest that a smaller learning rate generally leads to better performance, with the learning rate of 0.00001 achieving the highest precision (0.83), recall (0.82), and F1 score (0.83). While the initial decrease from 0.001 to 0.0001 showed a slight improvement in precision but a dip in recall and F1 score, further reducing the learning rate to 0.00001 consistently yielded the best metrics. When considering the batch size, a size of 16 also resulted in the highest precision (0.83), recall (0.82), and F1 score (0.83), matching the performance obtained with the smallest learning rate. Smaller (8) and larger (32) batch sizes led to lower F1 scores, indicating that a batch size of 16 strikes a favorable balance for this particular task and dataset. Overall, the study highlights the importance of hyperparameter tuning, as different configurations of learning rate and batch size can significantly influence the model's ability to generalize and perform effectively. The optimal performance appears to be achieved with a smaller learning rate (0.00001) and a moderate batch size (16).

5. Conclusion

The prediction of entities includes ambiguity and false positives, which often lead to delays and is considered a time-consuming process. Hence, effective entity prediction is important for ensuring that the entities are in various languages. However, providing accuracy by human professionals is considered to be limited and time-consuming. To solve this challenge, the current research on the BERT, Distil BERT and RoBERTa models is focused on overcoming these limitations and improving the prediction accuracy of the NER system. It predicts entities such as people, events, organizations and locations. The proposed research utilized the HiNER dataset to demonstrate its effectiveness. For the HiNER dataset, the proposed BERT method achieves accuracies, precisions, recalls and F1 scores of 0.97, 0.89, 0.89 and 0.89, respectively, which are optimal results among the other proposed models, such as the proposed RoBERTa and proposed Distil BERT. The accuracy, precision, and recall and F1 score of the RoBERTa and Distil BERT models were 0.96, 0.87, 0.88 and 0.88, respectively, and the F1 score, recall, precision and accuracy were 0.83, 0.82, 0.83 and 0.95, respectively. The comparative performance results indicate that the proposed BERT model outperforms the other models. In the future, the present approach could be applied to various entity prediction datasets for prediction purposes to improve the efficacy of NER mechanisms in various domain languages. These findings could assist in further research on effective NER models.

Declarations

Funding Support

There is no funding support for this study.

Conflict of Interest

There is no conflict of interest.

Data Availability Statement

There is no research data outside the submitted manuscript file.

REFERENCES

[1] A. Goyal, V. Gupta, M. J. C. Kumar, and Systems, "Deep learning-based named entity recognition system using hybrid embedding," vol. 55, no. 2, pp. 279-301, 2024.

[2] S. Bahad, P. Mishra, K. Arora, R. C. Balabantaray, D. M. Sharma, and P. J. a. p. a. Krishnamurthy, "Fine-tuning Pretrained Named Entity Recognition Models For Indian Languages," 2024.

[3] A. Toprak and M. J. T. J. o. S. Turan, "Enhanced Named Entity Recognition algorithm for financial document verification," vol. 79, no. 17, pp. 19431-19451, 2023.

[4] A. J. J. o. A. A. i. H. M. Janowski, "Natural language processing techniques for clinical text analysis in healthcare," vol. 7, no. 1, pp. 51-76, 2023.

[5] S. Kumar, G. Soni, and A. Sharan, "Deep Learning for Extracting Biomedical Entities from COVID-19 Dataset: A Case Study," in *Text Mining Approaches for Biomedical Data*: Springer, 2024, pp. 283-297.

[6] P. Vats, N. Sharma, and D. K. J. E. E. T. o. S. I. S. Sharma, "A Novel Ensemble Model for Complex Entities Identification in Low Resource Language," vol. 11, no. 4, 2024.

[7] M. Sabane, A. Ranade, O. Litake, P. Patil, R. Joshi, and D. J. a. p. a. Kadam, "Enhancing Low Resource NER Using Assisting Language And Transfer Learning," 2023.

[8] R. Shelke, D. S. J. T. E. Thakore, and Management, "A survey on various methods used in named entity recognition for hindi language," 2020.

[9] K. A. Lima *et al.*, "A novel data and model centric artificial intelligence based approach in developing high-performance named entity recognition for bengali language," vol. 18, no. 9, p. e0287818, 2023.

[10] S. Menon, J. Sanjanasri, B. Premjith, K. J. A. i. D. S. Soman, and C. Technologies, "Named Entity Recognition Using Deep Learning and BERT for Tamil and Hindi Languages," p. 395.

[11] O. Litake, M. Sabane, P. Patil, A. Ranade, and R. J. I. Joshi, "Mono Versus Multilingual BERT: A Case Study in Hindi and Marathi Named Entity Recognition," p. 607.

[12] S. Agarwal, Sukritin, A. Sharma, and A. Mishra, "Next Word Prediction Using Hindi Language," in *Ambient Communications and Computer Systems: Proceedings of RACCCS 2021*: Springer, 2022, pp. 99-108.

[13] R. Murthy, P. Bhattacharjee, R. Sharnagat, J. Khatri, D. Kanojia, and P. J. a. p. a. Bhattacharyya, "Hiner: A large hindi named entity recognition dataset," 2022.

[14] A. A. Choure, R. B. Adhao, and V. K. Pachghare, "NER in Hindi Language Using Transformer Model: XLM-Roberta."

[15] H. Gandhi and V. J. P. C. S. Attar, "Extracting aspect terms using CRF and bi-LSTM models," vol. 167, pp. 2486-2495, 2020.

[16] R. Sharma, S. Morwal, B. Agarwal, R. Chandra, M. S. J. N. C. Khan, and Applications, "A deep neural network-based model for named entity recognition for Hindi language," vol. 32, no. 20, pp. 16191-16203, 2020.

[17] S. Khanuja *et al.*, "Muril: Multilingual representations for indian languages," 2021.

[18] R. Sharma, S. Morwal, B. J. C. S. Agarwal, and Language, "Named entity recognition using neural language model and CRF for Hindi language," vol. 74, p. 101356, 2022.

[19] G. Yadav, S. S. Rathore, and D. J. I. J. o. S. I. Sadhya, "Named entity recognition for weather domain text in Hindi," vol. 6, no. 2, pp. 178-187, 2021.

[20] A. Goyal, V. Gupta, and M. J. K.-B. S. Kumar, "A deep learning-based bilingual Hindi and Punjabi named entity recognition system using enhanced word embeddings," vol. 234, p. 107601, 2021.

[21] R. Sharma, S. Morwal, B. J. I. J. o. C. I. Agarwal, and N. Intelligence, "Entity-extraction using hybrid deep-learning approach for Hindi text," vol. 15, no. 3, pp. 1-11, 2021.

[22] R. Shelke and D. J. I. J. o. N. L. C. Thakore, "A novel approach for named entity recognition on Hindi language using residual bilstm network," vol. 9, no. 2, pp. 1-8, 2020.

[23] A. Jain, D. Yadav, A. Arora, and D. K. J. C. S. Tayal, "Named-Entity Recognition for Hindi language using context pattern-based maximum entropy," vol. 23, 2022.

[24] R. Shelke and S. J. J. Vanjale, "Recursive LSTM for the Classification of Named Entity Recognition for Hindi Language," vol. 27, no. 4, pp. 679-684, 2022.

[25] R. Shelke and S. J. R. d. I. A. Vanjale, "Deep Named Entity Recognition in Hindi Using Neural Networks," vol. 36, no. 4, 2022.

[26] A. Goyal, V. Gupta, and M. J. I. J. o. R. Kumar, "Recurrent neural network-based model for named entity recognition with improved word embeddings," vol. 69, no. 10, pp. 6970-6976, 2023.

[27] C. Malarkodi and S. L. Devi, "A Deeper Study on Features for Named Entity Recognition," in *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 66.

[1] Z. Hu, W. Hou, X. J. N. C. Liu, and Applications, "Deep learning for named entity recognition: a survey," vol. 36, no. 16, pp. 8995-9022, 2024.
[2] D. Zheng, J. Li, Y. Yang, Y. Wang, and P. C.-I. J. A. S. Pang, "MicroBERT: Distilling MoE-Based Knowledge from BERT into a Lighter Model," vol. 14, no. 14, p. 6171, 2024.
[3] G. Bharathi Mohan, R. Prasanna Kumar, K. Krishna Jayanth, and S. J. S. C. S. Doss, "Telugu Language Analysis with XLM-RoBERTa: Enhancing Parts of Speech Tagging for Effective Natural Language Processing," vol. 6, no. 2, p. 106, 2025.
[4] R. Sharma, S. Morwal, B. J. C. S. Agarwal, and Language, "Named entity recognition using neural language model and CRF for Hindi language," vol. 74, p. 101356, 2022.
[5] R. Sharma, S. Morwal, B. Agarwal, R. Chandra, M. S. J. N. C. Khan, and Applications, "A deep neural network-based model for named entity recognition for Hindi language," *Neural Computing Applications,* vol. 32, no. 20, pp. 16191-16203, 2020.