

## BRIDGING DATA SCARCITY IN SINDHI NER USING MACHINE-LABELED CORPORA AND MULTILINGUAL TRANSFORMERS

Nazish Basir<sup>\*1</sup>, Muhammad Suleman Memon<sup>2</sup>, Mumtaz Qabulio<sup>3</sup>, Danish Nazir Arain<sup>4</sup>,  
Rafique Ahmed Vighio<sup>4</sup>

<sup>\*1</sup>Department of Information Technology, FET, University of Sindh

<sup>2</sup>Department of Information Technology, Dadu Campus, University of Sindh

<sup>3</sup>Department of Software Engineering, FET, University of Sindh

<sup>4</sup>Phd Scholar, Dr. AHS Bukhari Postgraduate Centre of ICT, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan

DOI: <https://doi.org/10.5281/zenodo.19464981>

**Keywords**

Named Entity Recognition, Sindhi language, multilingual transformers, weak supervision, pre-training, fine-tuning, low-resource languages.

**Article History**

Received: 29 December 2025

Accepted: 13 February 2026

Published: 28 February 2026

Copyright @Author

Corresponding Author: \*

Nazish Basir

**Abstract**

In low-resource languages, Named Entity Recognition (NER) possess many challenges and problems due to the shortage of high-quality annotated corpora and the unequal distribution of entity categories. In this study we inspect the effectiveness incorporating Machine-labeled data in Sindhi NER using two multilingual transformer models Multilingual BERT and XLM RoBERTa under two training settings, (i) direct fine-tuning on gold standard human-annotated data and (ii) machine-labeled pre-training followed by fine-tuning. The performances of models are assessed using entity-level precision, recall, and F1-score, together with learning-curve analysis and confusion matrices. The results indicates that machine-labeled pre-training improves recognition, particularly for mid-frequency and low-frequency entity categories. The XLM-RoBERTa outperform Multilingual BERT in both aggregated and entity-specific evaluations, and the pre-training increases the micro-F1 score for mBERT from 0.50 to 0.63 and for XLM-RoBERTa 0.72 to 0.79, compared to without the pre-training step. These findings indicate that large-scale weak supervision can mitigate data scarcity and improve contextual representation learning for Sindhi and other low-resource languages. The study provides a strong empirical baseline for Sindhi NER and highlights the practical value of machine-labeled pre-training for low-resource language processing.

**I. INTRODUCTION**

The Development of Natural Language Processing (NLP) systems for low-resource languages helps in promoting language preservation and allows their speakers to access digital information, technology, and online services. It also helps these communities participate more fully in the digital world. [1]. In Named Entity Recognition (NER), the lack of labeled datasets, limited language resources, and complex language structures make it difficult to develop and evaluate models [2].

Deep neural models have greatly improved Named Entity Recognition (NER) in languages with large datasets. However, progress in low-resource languages is still limited because of lack of data and linguistic challenges [3]. Models such as BiLSTM-CRF (which combine bidirectional recurrent networks with conditional random fields) perform well when large training datasets are available [4]. They can learn context effectively, but they need a lot of labeled data

to work well, which makes them less suitable for very low-resource languages.

Transformer-based pre-trained language models have greatly improved Natural Language Processing (NLP). BERT introduced deep bidirectional transformers trained on large datasets and achieved very strong results on many NLP tasks [5]. Its multilingual version, Multilingual BERT (mBERT), supports more than 100 languages and can transfer knowledge across languages [6]. Another model, XLM-RoBERTa (XLM-R), was trained on 2.5 TB of filtered CommonCrawl data and performs better than mBERT on many cross-lingual tasks, including Named Entity Recognition (NER) [7]. Both mBERT and XLM-R use shared subword vocabularies, which help the models transfer knowledge between different languages.

Sindhi presents a compelling case in this regard. It is an Indo-Aryan language spoken by millions, particularly in parts of South Asia, yet computational resources for it remain surprisingly limited. Its morphological complexity, agglutinative tendencies, free word order, absence of capitalization cues, and orthographic variations pose substantial challenges for automatic entity recognition. Existing work on Sindhi NER has primarily relied on small annotated datasets and conventional machine learning or neural sequence models [8], limiting the scalability and robustness of developed systems. There is also the absence of capitalization cues that English-based systems often rely on for identifying proper nouns. Similar challenges have been reported for other Pakistani languages such as Urdu, and multilingual transformer models have shown promising results [9], [10]. Yet, the performance of transformer-based models remains highly dependent on the availability of task-specific annotated data.

One possible way around the data problem is weak supervision. Instead of relying entirely on manually labeled corpora, models can first be trained on automatically generated or machine-labeled datasets. These datasets are often noisy, sometimes quite messy, but they can still expose the model to broad patterns [11], [12]. Such two-stage training strategies enable models to learn coarse entity patterns from large noisy corpora and refine them using smaller clean datasets. However, in practice the systematic evaluation of this approach for Sindhi NER using

modern multilingual transformers has not been sufficiently investigated.

In this paper, we conduct a comprehensive empirical study of Sindhi NER to examine that gap using two multilingual transformer architectures mBERT and XLM-R under both direct fine-tuning and weakly supervised pre-training pipelines. We evaluate four experimental configurations: (i) mBERT fine-tuned directly on a human-annotated Sindhi corpus, (ii) mBERT pre-trained on a machine-labeled corpus followed by fine-tuning, (iii) XLM-R fine-tuned directly on the human-annotated corpus, and (iv) XLM-R pre-trained on machine-labeled data and followed by fine-tuning. Comparing these configurations allows us to observe not only differences between model architectures but also the practical role of weak supervision in a low-resource environment.

**The main contributions of this work can be summarized as follows:**

- Development of a human-annotated Sindhi NER corpus containing newly defined named entity categories.
- Construction of a weakly supervised training framework that utilizes machine-labeled corpora for pre-training.
- A systematic empirical evaluation of multilingual transformer architectures under both direct and two-stage training strategies for Sindhi NER.

## II. LITERATURE REVIEW

Named Entity Recognition (NER) for the Sindhi language has increasingly attracted attention within the natural language processing (NLP) community due to the increased use of language on Social media and other platforms. Sindhi has several features that make automatic entity extraction difficult. These include complex word formation, flexible word order, no capitalization clues, and many dialect variations. In addition, the lack of large datasets and language tools makes it harder to build strong Named Entity Recognition (NER) systems. Previous research has explored rule-based, statistical, and neural methods. These studies have also worked on creating important language resources, such as annotated datasets and word embeddings, to support NLP research for Sindhi.

Ali et al. [13] presented one of the earliest analyses of the challenges and opportunities linked with Sindhi NER, their study gave a comparison on rule-based, machine learning, and hybrid approaches and exposed the limitations of traditional techniques in managing complex word forms and different dialects of Sindhi. The authors also emphasized that the lack of publicly available datasets is a major problem. They suggest using cross-lingual transfer learning from related languages such as Urdu and Hindi, along with modern neural models is promising solutions for low-resource languages.

In start the implementations on NER primarily relied on handcrafted rules and pattern-matching strategies. Hakro et al. [14] developed a rule-based NER system, categorizing entities into 12 predefined classes using dictionaries and linguistic patterns. They evaluated on a corpus of with more than 200,000 words, and the system achieved an accuracy of 97%. In continuation on this work, Hakro et al. [15] solved word ambiguities by using context-based scoring and lookup tables, looking at surrounding words to correctly identify entities. Jumani et al. [16] also proposed a rule-based framework employing gazetteers and suffix heuristics to detect 10 entity types, on a small annotated corpus achieving 98.71% accuracy. While these systems reported high accuracy, their dependency on handcrafted rules limited the scalability and generalization across domains.

Resource creation has played a critical role in advancing Sindhi NLP. Dootio and Wagan [17] introduced a large Sindhi text corpus for sentiment analysis comprising 15,788 documents annotated with polarity labels, while not directly focused on NER, this enabled feature engineering using techniques such as TF-IDF and Document Term Matrices, supporting broader linguistic research.

A significant advancement in Sindhi NER research was the development of annotated benchmark datasets and utilization of neural models. Ali et al. [8] developed SiNER, a gold-standard corpus containing 1.35 million tokens from 1,338 news articles annotated with 11 entity categories using the BIO scheme. Baseline evaluations using CRF achieved F1-scores of 82.54% and BiLSTM-CRF achieved 89.16% , highlighting the better performance of neural sequence models and the effectiveness of FastText embeddings. Using same corpus, Ali et al. [18]

proposed a joint neural architecture for POS tagging and NER that integrated BiLSTM layers with adversarial transfer learning and self-attention mechanisms, improved contextual representations, achieving F1-scores of 96.18% for POS tagging and 91.42% for NER.

Complementary efforts by Ali et al. [19] focused on large-scale corpora and embedding generation and constructed a 61-million-word Sindhi corpus and trained word embeddings using GloVe, CBOW, and Skip-Gram models, with Skip-Gram demonstrating superior semantic representation and achieving an F1-score of 90.11% in downstream NER tasks. Likewise, Ali et al. [20] introduced the SiPOS corpus with 293,000 tokens annotated using Universal POS tags and demonstrated strong performance using BiLSTM-Attention-CRF architectures. Further enhancements were proposed by Ali et al. [21] through CaBiLSTM, a context-aware bidirectional model incorporating self-attention and CRF layers, which achieved an F1-score of 91.25% on the SiNER corpus.

Moving to neural approaches, a study by Memon et al. [22], compared LSTM and GRU architectures for Sindhi POS tagging and the results indicated that LSTM models capture morphological dependencies better and achieved a validation accuracy of 80.96%. In a recent study Ali et al. [23] evaluated Sindhi word embeddings, confirming the effectiveness of Skip-Gram embeddings with 95.32% POS tagging accuracy and a 90.11% NER F1-score. Sindhi NER has moved from rule-based methods to neural models, but problems like lack of data, complex word forms, and performance across different domains still exist. This motivates the use of multilingual and transformer-based models.

For sequence labeling tasks like Named Entity Recognition (NER) Conditional Random Fields (CRFs) have been widely applied to due to their ability to model label dependencies and account for contextual information across a sequence of tokens. CRF models predict consistent label sequences and have been effectively used for pseudo-labeling to improve automatically labeled data quality and to enhance the performance of downstream deep learning models. Kim et al. [24] utilized CRFs in a bootstrapping framework, where iterative application of a CRF on a small seed corpus significantly

improved machine-labelled biomedical corpora, achieving an F1 score of 71.87% after three iterations and effectively handling out-of-vocabulary and ambiguous terms. More recent studies continue to confirm the relevance of CRF-based structured prediction. Arslan [25] used a BiLSTM-CRF architecture with contextual embeddings for product entity extraction from unstructured Turkish text, showing continued efficacy of CRF layers in sequence labeling tasks when combined with neural models and word embeddings. Reporting high performance and highlighting the robustness of CRFs particularly in low-resource language settings, Thant et al. [26] introduced the myNER corpus for Burmese NER and evaluated CRF and BiLSTM-CRF models with fastText embeddings.

In cross-lingual and low-resource contexts, the pretrained multilingual transformer models such as multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) have become dominant for NER. At SemEval-2023 Task 2 on Multilingual Complex NER, fine-tuning XLM-R demonstrated strong multilingual representation capabilities across 12 diverse languages, establishing its effectiveness as a base transformer for complex NER scenarios [27]. Al-Duwais et al. [28] used mBERT and XLM-R on Arabic cross-lingual NER, showing that such models are powerful for cross-lingual transfer and generalization, especially when adapted to target languages. Additionally, transformer-based approaches with XLM-R significantly outperform traditional baselines, often achieving higher precision and recall scores relative to mBERT, on Urdu and Arabic social media NER tasks for handling morphological and syntactic complexity. Collectively, the approaches like CRF for structured pseudo-labels and the use of deep multilingual transformers like mBERT and XLM-R looks promising to achieve robust NER performance in low-resource research settings, such as that of Sindhi.

In our previous study [29], we investigated the use of CRF-generated machine-Labeled data combined with multilingual transformers (mBERT and XLM-RoBERTa) to improve NER in Urdu and Sindhi through cross-lingual transfer. Although the results demonstrated the usefulness of machine annotation, that study did not evaluate a Sindhi-centric setting with an expanded human-annotated corpus, detailed

entity-wise analysis, or confusion matrix evaluation. Furthermore, while transformer-based models achieve competitive performance, their effectiveness remains constrained by the availability of high-quality annotated data. Motivated by these observations, in the present work we extend this line of research by constructing larger annotated resources and systematically examining the impact of machine-labeled pre-training with multilingual transformers for Sindhi NER, aiming to provide a practical and data-efficient solution for low-resource settings.

### III. METHODOLOGY

This section presents the methodology adopted to develop the Sindhi Named Entity Recognition (NER) system. The overall development process followed a structured sequential pipeline consisting of the following stages:

- Data Gathering and Cleaning
- Human-Labeled Corpus Creation
- Machine-Labeled Corpus Creation
- Fine-Tuning on Human-Labeled Corpus
- Pre-Training on Machine-Labeled Corpus and Fine-Tuning on Human-Labeled Corpus
- Results and Evaluation

Each stage is elaborated in detail in the following subsections.

#### A. Data Gathering and Cleaning

To construct the corpus for NER in Sindhi, textual data were collected from two primary sources the news websites such as BBC Sindhi, Kawish News and others and Sindhi-language Wikipedia pages. These platforms provided diverse content covering politics, culture, economics, and current affairs and structured textual material rich in named entities, including locations, organizations, historical personalities, and public institutions. Web-scraping tools such as BeautifulSoup and Selenium were used to automate the data collection process.

Preprocessing was done to remove advertisements, navigation links, and duplicate values. The text was converted to UTF-8 to keep the writing consistent and readable. More reprocessing was done for Sindhi, including handling its right-to-left script, normalizing punctuation, fixing spelling variations, and splitting text into tokens correctly. Many python scripts were implemented to apply automation in the process.

**B. Human-Labeled Corpus Creation**

Two local Sindhi language experts performed the manual annotation process following a detailed labeling guidelines and the Inside-Outside-Beginning (IOB) tagging scheme. To simplify the process for annotators Excel sheets were provided with validation applied to avoid mistyping and to ensure data integrity.

For annotation quality and consistency, inter-annotator disagreements were systematically reviewed, and final labels were assigned through consensus discussions. The final human-annotated corpus comprises 1,122 sentences, 33,808 tokens, and 3,814 named entities, providing a solid benchmark for evaluating Sindhi NER systems.

The annotated dataset encompasses both conventional named entity categories and language-specific entities relevant to the Sindhi context. Throughout this paper, entity categories are referred to using abbreviated labels for clarity and consistency, including COL (Color), DAT (Date/Time), EVT (Event), FOOD (Food), LOC (Location), MEA (Measure), NORP (Nationalities, Religious or Ethnic Groups), NUM (Number), ORG (Organization), PER (Person), POS (Position), PRD (Product), and SPT (Sports). A detailed description of these entity categories is presented in Table I, while their statistical distribution across the corpus is illustrated in Fig. 1.

**TABLE I**

DESCRIPTION OF NAMED ENTITY CATEGORIES IN THE SINDHI ANNOTATED CORPUS, INCLUDING TAG TYPES, DEFINITIONS, AND EXAMPLE ENTITIES WITH ENGLISH TRANSLATIONS.

| Entity Type  | Tag Types         | Description                                         | Sindhi Examples (with English Translation)                   |
|--------------|-------------------|-----------------------------------------------------|--------------------------------------------------------------|
| Color        | B-COL,<br>I-COL   | Names of colors                                     | ڳاڙهو (Red)<br>نيرو (Blue)                                   |
| Date/Time    | B-DAT,<br>I-DAT   | Specific dates, times, or periods.                  | ۱۴ آگسٽ (14 August)<br>سومر (Monday)                         |
| Event        | B-EVT,<br>I-EVT   | Named events, festivals, or historical occasions.   | عید الفطر (Eid al-Fitr)<br>عرس شاهه لطيف (Urs of Shah Latif) |
| Food         | B-FOOD,<br>I-FOOD | Names of food items or dishes.                      | برياني (Biryani)<br>ڪباب (Kebab)                             |
| Location     | B-LOC,<br>I-LOC   | Geographical locations                              | ڪراچي (Karachi)<br>پاڪستان (Pakistan)                        |
| Measure      | B-MEA,<br>I-MEA   | Units of measurement                                | ڪلوگرام (Kilogram)<br>ڪلاڪ (Hour)                            |
| NORP         | B-NORP,<br>I-NORP | Nationalities, religious, or ethnic groups.         | سنڌي (Sindhi)<br>مسلمان (Muslim)                             |
| Number       | B-NUM,<br>I-NUM   | Numerical values and ordinal references.            | ڏهه (Ten)<br>پهريون (First)                                  |
| Organization | B-ORG,<br>I-ORG   | Names of organizations, institutions, or companies. | سنڌ يونيورسٽي (University of Sindh)<br>يونيليور (Unilever)   |
| Person       | B-PER,<br>I-PER   | Names of individuals mentioned in text.             | بلاول (Bilawal)<br>بينظير ڀٽو (Benazir Bhutto)               |
| Position     | B-POS,<br>I-POS   | Titles or positions associated with individuals.    | وزير اعظم (Prime Minister)<br>چيف جسٽس (Chief Justice)       |
| Product      | B-PRD,<br>I-PRD   | Brands or specific products.                        | آئي فون (iPhone)<br>ڪوڪا ڪولا (Coca-Cola)                    |
| Sports       | B-SPT,<br>I-SPT   | Names of sports or specific games.                  | ڪرڪيٽ (Cricket)<br>هاڪي (Hockey)                             |

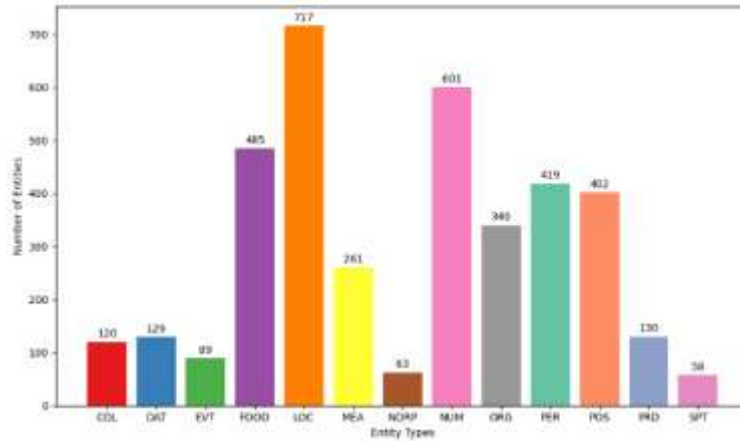


Fig. 1. Distribution of named entity categories across the annotated Sindhi corpus.

### C. Machine-Labeled Corpus Creation

The machine-labeled corpus was generated by firstly extracting large amount of data through web scrapping and carefully cleaning following the same approached used for Human-annotated corpus. The pipeline followed is illustrated in Fig. 2 having multiple steps, including bootstrap sampling, CRF-based sequence labeling, confidence-based filtering, and ensemble voting.

At first, semantic word representations computed using FastText embeddings were projected into a lower-dimensional space using Principal Component Analysis (PCA) [31] to reduce noise and computational complexity. Lexical and contextual features, such as word length, affixes, digits, punctuation indicators, character n-grams, and neighboring tokens, were then extracted for each token and Conditional Random Fields (CRF) were trained to model sequential dependencies in entity labeling. An ensemble bagging approach was adopted. Given an ensemble of  $M$  models  $\{f_1, f_2, \dots, f_M\}$ , each trained on a bootstrap sample, the predicted label  $\hat{y}_i$  for token  $w_i$  was determined via majority voting:

$$\hat{y}_i = \arg \max_{y \in Y} \sum_{m=1}^M I[f_m(w_i) = y] \quad (1)$$

Here,  $\hat{y}_i$  is the predicted label,  $Y$  is the set of possible labels,  $f_m(w_i)$  is the prediction of the  $m$ -th model, and  $I[\cdot]$  is the indicator function. Tokens with prediction confidence below 0.8 were assigned the neutral class  $O$ . This method generated a high-quality machine-labeled corpus suitable for pre-training.

The resulting machine-labeled Sindhi corpus contains 532,321 tokens across 17,179 sentences, with a total of 22,016 named entity annotations distributed across multiple categories. These include 3,417 person names (PER), 6,072 locations (LOC), 2,090 organizations (ORG), 1,451 dates (DAT), 3,763 numerical expressions (NUM), 76 events (EVT), 1,186 measurements (MEA), 161 nationalities/religious or ethnic groups (NORP), 3,267 positions or job titles (POS), 84 products (PRD), 33 colors (COL), 387 food items (FOOD), and 29 sports-related terms (SPT). This semi-automatic annotation greatly reduces manual effort, creating a large, diverse corpus that enables models to learn richer contextual representations before fine-tuning on the human-annotated dataset.

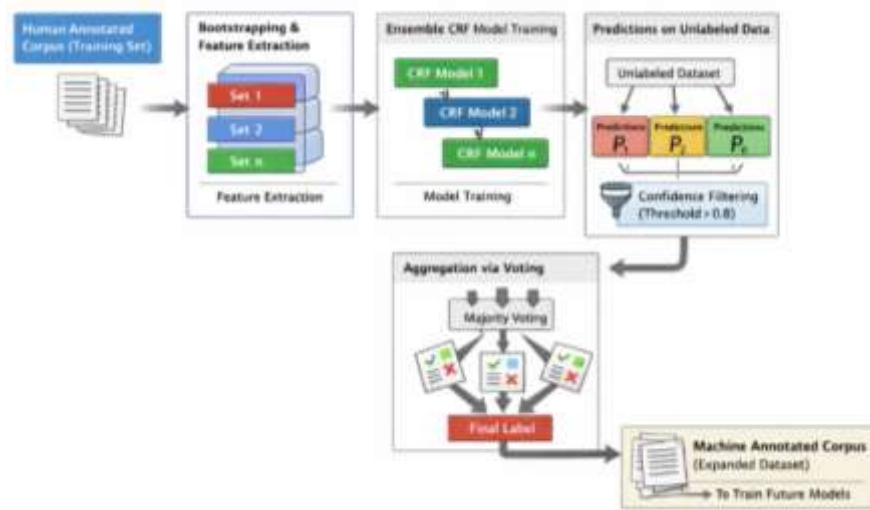


Fig. 2. Workflow of the proposed machine-labeling system: from bootstrapping human-annotated data to generating an expanded machine-labeled corpus through ensemble learning and confidence filtering.

#### IV. EXPERIMENTAL SETUP

This section presents the experimental setup adopted for the study. To investigate the effectiveness of

transformer-based models for Sindhi Named Entity Recognition (NER), two experimental pipelines were designed.

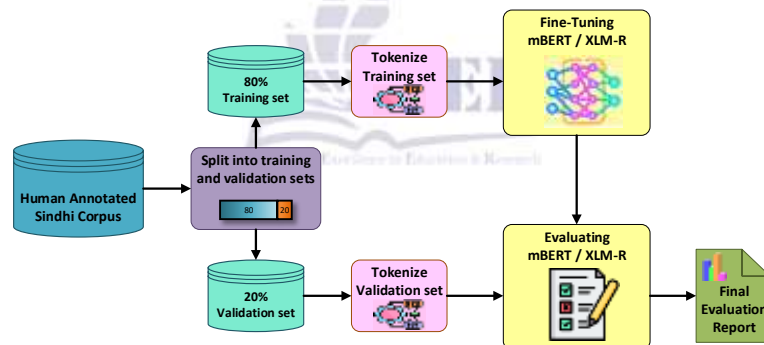


Fig. 3. Pipeline for direct fine-tuning of multilingual transformer models on the human-labeled Sindhi NER corpus.

The first pipeline in fig 3 performs direct fine-tuning of multilingual transformer models (mBERT and XLM-R) on the human-annotated Sindhi corpus,

where the dataset is split into training and validation sets, tokenized, and then used to train and evaluate the models.

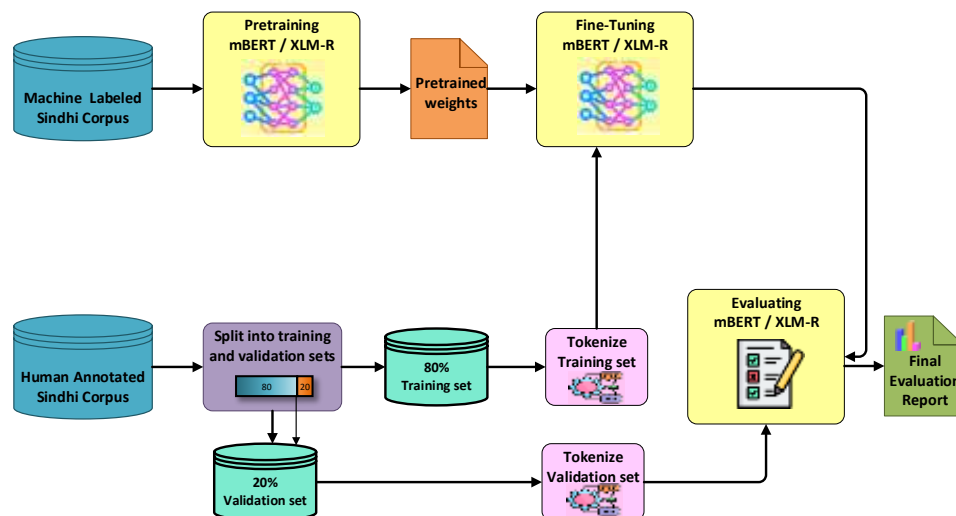


Fig. 4. Pipeline for pre-training and fine-tuning multilingual transformer models using combined machine-labeled and human-annotated Sindhi NER corpora.

The second pipeline in fig 4 introduces an additional pre-training stage on a large machine-labeled Sindhi corpus, allowing the models to first learn general linguistic and entity patterns from weakly annotated data. The pre-trained weights are then used to initialize the models before fine-tuning on the human-annotated dataset.

#### A. Hardware and Software Configuration

All experiments were conducted in the Google Colab environment using an NVIDIA Tesla T4 GPU with 16 GB VRAM, an Intel Xeon CPU, and 32 GB of RAM. The implementation utilized Python 3.10 and PyTorch, with Hugging Face's transformers library for pre-trained transformer models. Additional libraries included seqeval for NER evaluation, pandas for dataset handling, scikit-learn for train-test splitting and confusion matrices, and matplotlib and seaborn for visualization.

#### B. Datasets

Both the machine-labeled corpus and The human-labeled corpus used same format and tagging scheme. Both files were in csv format having columns; sentence\_id, word and labels. Sentences were grouped by sentence\_id, preserving word-level sequences and entity labels. The human-labeled dataset was split into 80% training and 20% validation.

#### C. Tokenization and Dataset Preparation

Tokenization was performed using BertTokenizerFast for BERT and XLM-R tokenizer (xlm-roberta-base) for XLM-R. A custom PyTorch Dataset class aligned tokenized subwords with original word-level labels, masking subword tokens with -100 to exclude them from loss computation. Sequences were padded or truncated to a maximum length of 512 tokens. Separate DataLoaders were prepared for pre-training and fine-tuning, with a batch size of 8.

#### D. Model Architectures

Two transformer-based models were evaluated:

##### 1. BERT-based model:

BertForTokenClassification initialized with bert-base-multilingual-cased. It outputs token-level predictions for the combined label set.

##### 2. XLM-R-based model:

XLMRobertaForTokenClassification initialized with xlm-roberta-base. This model is designed for cross-lingual performance and can leverage multilingual representations to improve Sindhi NER.

Both models were trained on GPU when available. The number of output labels was set to the total unique label set, ensuring compatibility with both pre-training and fine-tuning datasets.

### E. Pre-training Configuration

For Pre-training both models used cross-entropy loss with label smoothing (0.1) to reduce the impact of noisy labels. AdamW optimizer was applied with a learning rate of  $1 \times 10^{-5}$  and weight decay 0.01. A cosine learning rate scheduler with 10% warmup steps was used. Training lasted 5 epochs with gradient accumulation over 2 steps. The pre-trained models were saved and later used for fine-tuning.

### F. Fine-tuning Configuration

For Fine-tuning both BERT and XLM-R used AdamW optimizer with a learning rate of  $3 \times 10^{-5}$  and a cosine scheduler with 10% warmup. Training was performed for 5 epochs using token-level cross-entropy loss, ignoring masked subword tokens. Learning and validation losses were monitored to ensure convergence and detect overfitting or underfitting.

### G. Evaluation Metrics

Evaluation included token-level metrics (precision, recall, F1-score) using Seqeval and entity-level metrics computed via custom logic that included "O" labels. Entity-level confusion matrices and learning curves

were generated to visualize model performance and detect any training instability.

### H. Reproducibility

Reproducibility was ensured through fixed random seed (42), deterministic tokenization, and consistent hyperparameters. Both models mBERT and XLM-R followed identical pre-training and fine-tuning pipelines, allowing direct comparison of monolingual versus cross-lingual capabilities for Sindhi NER.

## V. RESULTS AND DISCUSSION

This section presents the experimental results, the evaluation focuses on comparing monolingual fine-tuning and machine-labeled pre-training followed by fine-tuning for both mBERT and XLM-R architectures.

### A. Overall Model Performance

Micro-averaged metrics are used as the main evaluation method because they combine the results from all entity instances in the dataset. This gives a fair and overall measurement of how well the model performs on the entire dataset. Table II presents the overall micro-averaged evaluation results for all experimental configurations.

TABLE II

OVERALL MICRO-AVERAGED PERFORMANCE OF mBERT AND XLM-R FOR SINDHI NER.

| Model Configuration              | Precision | Recall | F1-Score |
|----------------------------------|-----------|--------|----------|
| mBERT (Fine-tune Only)           | 0.53      | 0.46   | 0.50     |
| mBERT (Pre-trained + Fine-tuned) | 0.65      | 0.61   | 0.63     |
| XLM-R (Fine-tune Only)           | 0.69      | 0.75   | 0.72     |
| XLM-R (Pre-trained + Fine-tuned) | 0.77      | 0.81   | 0.79     |

The results clearly demonstrate that machine-labeled pre-training considerably improves performance for both architectures, pre-training increases the F1-score for mBERT from 0.50 to 0.63 and for XLM-R from 0.72 to 0.79, indicating enhanced generalization capability.

In both configurations, XLM-R with pre-training achieves the highest performance, obtaining a micro-F1 score of 0.79. The results implies that multilingual contextual representations combined with machine-labeled pre-training provide stronger entity categorization and better boundary detection for Sindhi NER.

Overall, the results confirm that machine-labeled pre-training benefits to performance gains, and that XLM-R is better model for low-resource Sindhi NER compared to mBERT.

### B. Training Dynamics Analysis

Training and validation loss curves were analyzed to understand how the models learned during training. Fig 5 and Fig 6 show the learning behavior of mBERT and XLM-R with and without pre-training.

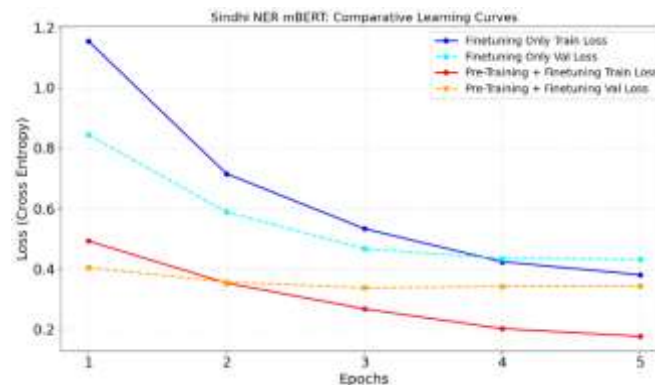


Fig. 5. Learning curves of mBERT for fine-tuning only and pre-training + fine-tuning configurations (training and validation loss across epochs).

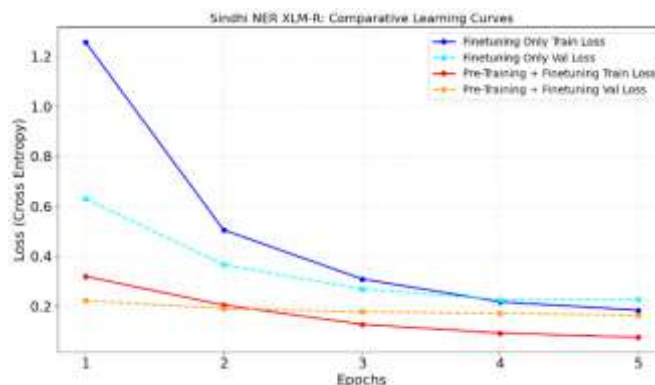


Fig. 6. Learning curves of XLM-R for fine-tuning only and pre-training + fine-tuning configurations (training and validation loss across epochs).

For mBERT without pre-training, the training loss decreased from 1.1547 to 0.3806, and the validation loss decreased from 0.8439 to 0.4324. It is observed that the validation loss stopped improving after the fourth epoch, showing limited generalization. When pre-training was used, mBERT started with a lower training loss of 0.4925 and achieved a lower validation loss of 0.3372. This shows that pre-training helped the model learn faster and reduce overfitting. XLM-R showed faster and more stable learning than mBERT. Without pre-training, the training loss decreased from 1.2575 to 0.1829, and validation loss decreased from 0.6299 to 0.2264, showing good generalization. With pre-training, XLM-R achieved the best results, reaching the lowest validation loss of 0.1633 and showing very stable learning. Overall, XLM-R worked better than mBERT, learning faster, having lower validation loss, and generalizing better. Pre-training helped both models perform better.

### C. Impact of Pre-training on Convergence

It has been observed that across both architectures, machine-labeled pre-training leads to:

- Lower initial fine-tuning loss
- Faster convergence within early epochs
- Improved validation stability
- Reduced risk of overfitting

The effect is particularly notable for mBERT, where pre-training substantially narrows the train-validation gap. Although XLM-R already exhibits strong convergence without pre-training, additional pre-training further enhances stability and lowers validation loss.

Overall, training dynamics analysis confirms that machine-labeled pre-training improves optimization efficiency and generalization, especially for low-resource Sindhi NER.

### D. Entity-wise Evaluation (Strict IOB)

Strict IOB evaluation was conducted for each named entity type, with precision, recall, and F1-score to provide entity-wise performance analysis and detailed understanding of the model's strengths and weaknesses. The results are visualized through heatmaps in Figs. 7-10 to highlight performance variations across categories and models.

#### 1. mBERT (Fine-Tune Only)

The entity-wise results for BERT (mBERT) trained only with fine-tuning are shown in Fig. 7. The model

performed well on common entities, with an F1-score of 0.80 for NUM and 0.79 for MEA. It also recognized PER entities reasonably well with an F1-score of 0.54. But, the model failed to detect some rare or complex entities such as EVT, NORP, and SPT with an F1-score of 0.00. Other entities like PRD and DAT had low performance with F1-scores below 0.30. Because of this imbalance the macro F1-score was 0.35 even though the micro F1-score was 0.50. These results show that fine-tuning alone is not enough to recognize rare or complex entities in the Sindhi Corpus.

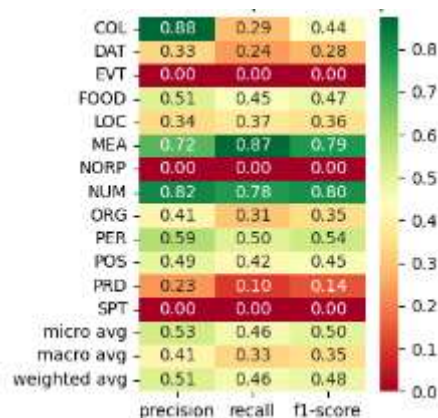


Fig. 7. Entity-wise precision, recall, and F1-score for mBERT (Fine-tune Only).

#### 2. XLM-R (Fine-Tune Only)

Figure 8 shows the entity-wise results of XLM-R in the fine-tuning-only setting. Compared to BERT (mBERT), XLM-R performs better on most entity types. The model achieved high F1-scores for NUM (0.89), MEA (0.87), and PER (0.83). It also improved recognition of LOC (0.72) and ORG (0.63). The

macro F1-score of 0.55 shows better performance across different entity classes. However, rare entities such as NORP and SPT were still not recognized (F1 = 0.00). This indicates that fine-tuning alone is not enough to learn very low-frequency entities. Overall, XLM-R performs better because of its stronger multilingual representations.

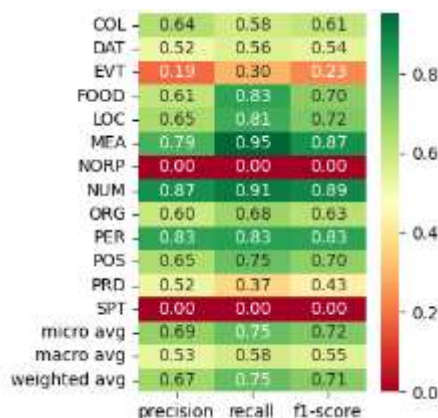


Fig. 8. Entity-wise precision, recall, and F1-score for XLM-R (Fine-tune Only).

### 3. mBERT (Pre-training + Fine-Tuning)

Figure 9 shows the entity-wise results of BERT (mBERT) after machine-labeled pre-training followed by fine-tuning on the Sindhi dataset. Pre-training improved the model's performance on many entity types. The F1-score for PER increased from 0.54 to 0.70, and DAT improved greatly from 0.28 to 0.70. Other entities such as POS (0.68), PRD (0.49), and

SPT (0.30) also showed better results. However, some entities like NORP were still not recognized. The macro F1-score increased from 0.35 to 0.52, showing that machine-labeled pre-training helps the model learn better representations and improves recognition of several entity types, especially those with moderate frequency.

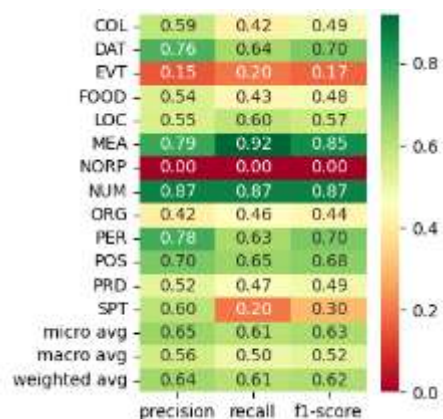


Fig. 9. Entity-wise precision, recall, and F1-score for mBERT (Pre-trained + Fine-tune).

### 4. XLM-R (Pre-training + Fine-Tuning)

Figure 10 shows the entity-wise performance of XLM-R with machine-labeled pre-training followed by fine-tuning. This setup achieved the best results across most entity types. High F1-scores were observed for NUM (0.92), MEA (0.90), LOC (0.83), and PER (0.84). Medium-frequency entities such as ORG (0.77) and POS (0.71) were also recognized well.

Importantly, rare entities like NORP (0.29) and SPT (0.52) which were not detected before, are now recognized. The macro F1-score reached 0.69, and the weighted F1-score was 0.79, showing balanced performance across entity types. These results show that combining pre-training with fine-tuning helps the model learn better representations and improves recognition of both common and rare entities.

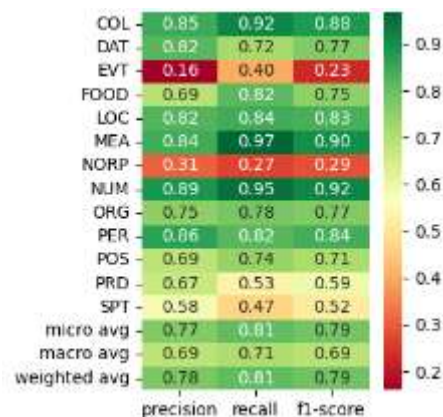


Fig. 10. Entity-wise precision, recall, and F1-score for XLM-R (Pre-trained + Fine-tune).

### E. Key Observations

In both architectures, machine-labeled pre-training consistently improves entity-level performance specially for mid and low frequency entities. XLM-R performs better than mBERT in all experiments, proving stronger multilingual representations. Common entities such as NUM, MEA, and PER are better recognized across models, but challenging categories like NORP, EVT, and SPT remained a problem. Pre-training enables detection of previously

unseen entities and producing more balanced performance reflected in macro F1-scores.

### F. Confusion Matrix Analysis

The confusion matrices were generated for each configuration to further analyze the performance of the NER models at the entity level, which allow a detailed examination of how frequently each entity type is correctly classified and which entities are most often confused with others.

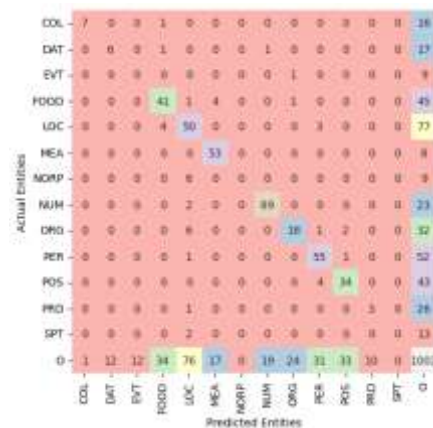


Fig. 11. Entity-level confusion matrix for mBERT fine-tuned only.

In Fig 11, mBERT fine-tuned only, the model performs quite well on frequent entities such as NUM and MEA, achieving high counts along the diagonal but, some entities such as EVT, NORP, and SPT are

rarely predicted correctly, often being misclassified as O (non-entity) or other frequent categories like LOC and ORG.

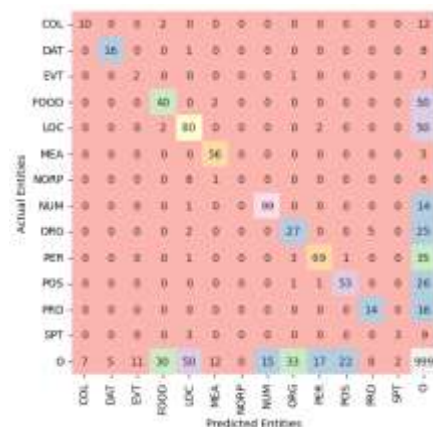


Fig. 12. Entity-level confusion matrix for mBERT pre-trained and fine-tuned.

In Fig 12, mBERT after pre-training on machine-labeled data, shows improved classification. The counts along the diagonal increased for entities like LOC, NUM, and PER, while off-diagonal misclassifications decrease. Still, rare entities such as

NORP experience confusion with more frequent categories, showing that pre-training benefits overall recognition but does not fully resolve low-resource entity issues.

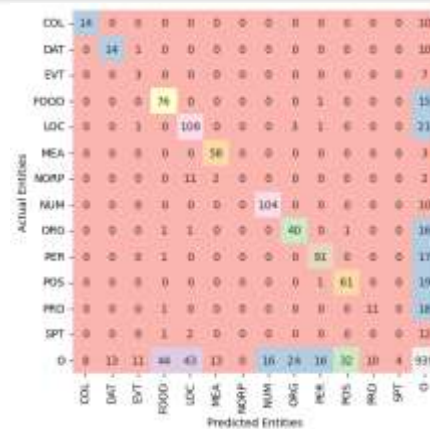


Fig. 13. Entity-level confusion matrix for XLM-R fine-tuned only.

In Fig 13, XLM-R fine-tuned only, demonstrates stronger performance compared to mBERT. Entities such as LOC, NUM, PER, and MEA exhibit very high correct prediction counts, while misclassifications are

reduced. Nevertheless, rare entities such as SPT and EVT are still occasionally misclassified, although to a lesser extent than mBERT.

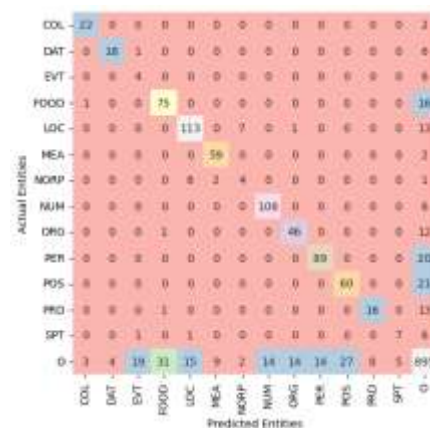


Fig. 14. Entity-level confusion matrix for XLM-R pre-trained and fine-tuned.

Finally, in Fig 14, XLM-R with pre-training and fine-tuning achieves the best results. Most entities show high diagonal counts with minimal off-diagonal misclassifications. Pre-training helps the model better distinguish between similar entities such as LOC and ORG, and rare entities like NORP and SPT also see improvements.

In general, the confusion matrices, confirms that XLM-R with pre-training and fine-tuning provides the most accurate and balanced entity-level predictions for Sindhi NER.

## VI. CONCLUSION

This study examined the use of transformer-based models for Named Entity Recognition (NER) in

Sindhi, a low-resource South Asian language. Two multilingual models, mBERT and XLM-R, were tested in two settings: (i) fine-tuning using only human-labeled data, and (ii) pre-training on machine-labeled data followed by fine-tuning. Results from learning curves, heatmaps, and confusion matrices display that machine-labeled pre-training clearly improves model performance, especially for rare entity types. Among the models, XLM-R with pre-training and fine-tuning achieved the best results, overtaking mBERT in both overall and entity-level performance. The experiments also revealed challenges that are common in low-resource NER. Rare entities such as NORP, EVT, and SPT were sometimes misclassified, and similar entities like LOC and ORG were occasionally confused.

However, using machine-labeled data helped reduce the problem of limited training data and improved the model's ability to generalize. In conclusion, this work provides a baseline for low-resource NER and shows that machine-labeled pre-training significantly improves transformer-based models, supporting future research.

For future work, several directions are suggested, mainly an active learning with human feedback could improve the quality of machine-labeled data. The model could be tested on different domains and different text sources such as news, legal documents, or social media. Adding morphological and syntactic features may also help because Sindhi has complex linguistic variations. This framework could also be extended to other Pakistani languages such as Pashto and Panjabi, helping build multilingual NER systems for national languages.

#### REFERENCES

- [1] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," *Natural Language Processing*, vol. 31, no. 2, pp. 183–197, 2025.
- [2] Z. Zhang, "Named entity recognition: challenges in document annotation, gazetteer construction and disambiguation," University of Sheffield, 2013.
- [3] M. O. Raza, N. A. Mahoto, A. Shaikh, N. Pathan, H. Alshahrani, and M. A. Elmagzoub, "A Machine Learning Approach of Text Classification for High-and Low-Resource Languages," *Comput. Intell.*, vol. 41, no. 4, p. e70114, 2025.
- [4] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 260–270.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers), 2019, pp. 4171–4186.
- [6] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jun. 2019, pp. 4996–5001. [Online]. Available: <http://arxiv.org/abs/1906.01502>
- [7] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440–8451.
- [8] W. Ali, J. Lu, and Z. Xu, "SiNER: A Large Dataset for Sindhi Named Entity Recognition," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 2953–2961. [Online]. Available: <https://sindhyat.com/database/SindhiRomanDictionary/>
- [9] K. Aziz et al., "Advancing Urdu named entity recognition: deep learning for aspect targeting," *Complex & Intelligent Systems*, vol. 11, no. 12, p. 489, 2025.
- [10] A. Ahmed, D. Huang, and S. Y. Arafat, "Enriching Urdu NER with BERT Embedding, Data Augmentation, and Hybrid Encoder-CNN Architecture," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 4, Apr. 2024, doi: 10.1145/3648362.
- [11] J. Shang, L. Liu, X. Gu, X. Ren, T. Ren, and J. Han, "Learning named entity tagger using domain-specific dictionary," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2054–2064.
- [12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.
- [13] W. Ali, A. Kehar, and H. Shaikh, "Towards Sindhi Named Entity Recognition: Challenges and Opportunities," in *1st National Conference on Trends and Innovations in Information Technology*, 2015. [Online]. Available: <https://www.mustgo.com/worldlanguages/sindhi/>

- [14] D. Nawaz Hakro, M. Ahmed Hakro, and I. Ali Lashari, "SINDHI NAMED ENTITY RECOGNITION (SNER)," *The Government: Research Journal of Political Science*, vol. 5, 2017.
- [15] D. Nawaz, S. A. Awan, Z. A. Bhutto, M. Memon, and M. Hameed, "Handling Ambiguities in Sindhi Named Entity Recognition (SNER)," *Sindh University Research Journal-SURJ (Science Series)*, vol. 49, no. 3, pp. 513–516, 2017.
- [16] A. K. Jumani, M. A. Memon, F. H. Khoso, A. A. Sanjrani, and S. Soomro, "Named entity recognition system for Sindhi language," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICTS*, Springer Verlag, 2018, pp. 237–246. doi: 10.1007/978-3-319-95450-9\_20.
- [17] M. A. Dootio and A. I. Wagan, "Development of Sindhi text corpus," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 468–475, May 2021, doi: 10.1016/j.jksuci.2019.02.002.
- [18] W. Ali, R. Kumar, Y. Dai, J. Kumar, and S. Tumrani, "Neural Joint Model for Part-of-Speech Tagging and Entity Extraction," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Feb. 2021, pp. 239–245. doi: 10.1145/3457682.3457718.
- [19] W. Ali, J. Kumar, J. Lu, and Z. Xu, "Word Embedding based New Corpus for Low-resourced Language: Sindhi," *arXiv preprint arXiv:1911.12579*, Nov. 2019, [Online]. Available: <http://arxiv.org/abs/1911.12579>
- [20] W. Ali, Z. Xu, and J. Kumar, "SiPOS: A Benchmark Dataset for Sindhi Part-of-Speech Tagging," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2021, pp. 22–30. doi: 10.26615/issn.2603-2821.2021\_004.
- [21] W. Ali, J. Kumar, Z. Xu, R. Kumar, and Y. Ren, "Context-aware bidirectional neural model for sindhi named entity recognition," *Applied Sciences (Switzerland)*, vol. 11, no. 19, Oct. 2021, doi: 10.3390/app11199038.
- [22] A. A. Memon, S. Hina, A. K. Kazi, and S. Ahmed, "Parts-of-speech tagger for Sindhi language using deep neural network architecture," *Mehran University Research Journal of Engineering and Technology*, vol. 43, no. 3, p. 47, Jul. 2024, doi: 10.22581/muet1982.2768.
- [23] W. Ali, S. Tumrani, J. Kumar, and T. R. Soomro, "An Evaluation of Sindhi Word Embedding in Semantic Analogies and Downstream Tasks," *arXiv preprint arXiv:2408.15720*, 2024, [Online]. Available: <http://arxiv.org/abs/2408.15720>
- [24] J. Kim, Y. Ko, and J. Seo, "Construction of Machine-Labeled Data for Improving Named Entity Recognition by Transfer Learning," *IEEE Access*, vol. 8, pp. 59684–59693, 2020, doi: 10.1109/ACCESS.2020.2981361.
- [25] D. Soomro, S. Kumar, A. A. Shaikh, and H. Raj, "Bio-NER: Biomedical Named Entity Recognition using Rule-Based and Statistical Learners," 2017. [Online]. Available: [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)
- [26] K. L. Thant, K. Nongpong, Y. K. Thu, T. Aung, K. H. Wai, and T. M. Oo, "myner: Contextualized burmese named entity recognition with bidirectional lstm and fasttext embeddings via joint training with pos tagging," in *2025 IEEE International Conference on Cybernetics and Innovations (ICCI)*, 2025, pp. 1–6.
- [27] R. Mehta and V. Varma, "Llm-rm at semeval-2023 task 2: Multilingual complex ner using xlm-roberta," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023, pp. 453–456.
- [28] M. Al-Duwais, H. Al-Khalifa, and A. Al-Salman, "A Benchmark Evaluation of Multilingual Large Language Models for Arabic Cross-Lingual Named-Entity Recognition," *Electronics (Switzerland)*, vol. 13, no. 17, Sep. 2024, doi: 10.3390/electronics13173574.
- [29] N. Basir, D. N. Hakro, K. U. R. Khoubati, and Z. Bhatti, "Leveraging machine-labeled data and cross-lingual transfer for NER in Urdu and sindhi," *J. Inf. Commun. Technol.-(JICT)*, vol. 19, pp. 1–8, 2025.