Leveraging Machine-Labeled Data and Cross-Lingual Transfer for NER in Urdu and Sindhi

Nazish Basir[1*], Dil Nawaz Hakro[2], Khalil-Ur-Rehman Khoumbati[3], Zeeshan Bhatti[4]

Abstract*: Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying entities within text. In low-resource languages like Urdu and Sindhi, due to limited annotated datasets and complex linguistic features such as rich morphology, agglutination, and the absence of capitalization cues the researchers face many challenges. In our study we are introducing a distinct approach that combines machine-labeled data generation with advanced multilingual transformer models to enhance the performance of low resource languages, with cross lingual transfer learning to improve NER performance in Sindhi. To the best of our knowledge, this is the first work exploring cross-lingual NER transfer from Urdu to Sindhi. We have also introduced two new entity types that include colors and foods, which have not been explored previously in Urdu and Sindhi language research. To reduce the need for extensive manual annotation, we used a bagging-based ensemble of Conditional Random Field (CRF) models, to generate high-confidence machine-labeled datasets. These models were trained on subsets of a smaller dataset, which were annotated by language experts. The machine-labeled data notably increased the volume of training data, which is essential for low-resource languages. We pre-trained two models, Multilingual BERT (mBERT) and XLM-RoBERTa on machine-labeled data and fine-tuned them on the human-annotated datasets. Our experiments demonstrated improvements in the performance of the Named Entity Recognition for both languages. Particularly, for Sindhi, the XLM-RoBERTa model's F1 score increased from 0.302 (without pre-training) to 0.681 after pre-training on a combined machine-labeled data of Urdu and Sindhi language, which is approximate increase of 125%. Our results show the effectiveness of incorporating machine-labeled data and cross-lingual knowledge transfer from Urdu to Sindhi language.

***Keywords:* Named Entity Recognition; NER; Urdu; Sindhi; Machine-Labeled Data; Cross-Lingual; Transfer Learning; mBERT; XLM-RoBERTa**

INTRODUCTION

Named Entity Recognition (NER) involves identifying and classifying key pieces of text into predefined categories like people, locations, organizations, numbers, dates, times, quantities, and so on [1]. Named entity recognition NER is an important part of natural language processing (NLP) and creating high-quality annotated datasets for low-resource languages like Urdu and Sindhi is very challenging [2][3]. Languages like this have complex structures, word forms, and grammar and this makes NER difficult for researchers.

[1-2-3-4]University of Sindh, Jamshoro
Country: Pakistan
Email: nazish.basir@usindh.edu.pk

They also lack important resources like gazetteers, annotated text datasets, and pre-trained models. Words in Urdu and Sindhi can be ambiguous and change meaning based on context, which adds more complexity in the task [4]. Not having annotated datasets and supporting tools like gazetteers is a major problem for NER in Urdu and Sindhi, especially due to challenges in writing, borrowed words from other languages, and different dialects. The researchers face even more problems when introducing new categories, which haven't been introduces in a languages before [5][6].

Our research addresses these challenges by applying a new method to generate machine-labeled data and by using advanced multilingual models to improve NER performance. We introduced new categories like colors and foods for advancing NER in Urdu and Sindhi. We have combined previously used methods like Conditional Random Fields (CRF) with modern multilingual transformer models, specifically Multilingual BERT (mBERT) [7] and XLM-RoBERTa [8] which are transformer-based models designed for multilingual NLP tasks. mBERT is an extension of BERT which is trained using masked language modeling (MLM) on Wikipedia data for 104 languages, enabling zero-shot cross-lingual transfer. XLM-R is an extension of RoBERTa which is pre-trained on 2.5TB of CommonCrawl data across 100 languages.

Our experiments is generating machine-labeled data using a bagging-based ensemble technique [9], which create several CRF models trained on different parts of a small, human-annotated dataset. To generate machine-labeled data, we first created a smaller carefully annotated dataset for both Sindhi and Urdu. This approach reduces the amount of manual annotation needed and making the process more efficient than previous methods which relied on larger annotated datasets.

Using this ensemble method of CRF, we created large machine-labeled datasets in Sindhi and Urdu, which we used as valuable datasets for pre-training NER models. Adding machine-generated data improves model accuracy, demonstrating the positive impact of machine-labeled data on NER performance. We experimented on both mBERT and XLM-RoBERTa to see how cross-lingual transfer learning and machine-labeled data affect the performance. Our experimental setups involved pre-training on machine-labeled datasets from both languages to evaluate improvements in NER in Sindhi and Urdu.

Cross-lingual transfer [10] in Named Entity Recognition (NER) enables knowledge transfer from high-resource to low-resource languages, and can reduce the need for large, annotated datasets. We want to observe how adding machine-labeled data affects model performance, especially for Sindhi. We also want to explore cross-lingual effects a suitable technique for low resource languages by pre-training models on machine-labeled data from both Sindhi and Urdu, taking advantage of the similarities between these two languages to enhance NER performance. This work

seeks to lay a strong foundations for developing NER capabilities in low-resource languages, demonstrating the effectiveness of machine-labeled data and cross-lingual knowledge transfer, opening a door for multilingual NLP research.

LITERATURE REVIEW

Early approaches to Urdu NER mainly relied on rule-based and statistical methods. Riaz [11] developed a rule-based NER system for Urdu that utilized handcrafted linguistic rules to identify entities, the method achieved good accuracy, but required manual effort to craft the rules. Singh et al. [12] proposed a rule-based Urdu NER system using a dataset of 6,000 person names and IJCNLP-08's twelve NE categories, addressing tagging and segmentation challenges, with plans to enhance it using POS tagging and expanded gazetteers. Similarly, Jahangir et al. [4] explored the use of n-gram models combined with gazetteers for Urdu NER. These statistical models showed some success, but they were limited by the small size of available annotated corpora and struggled with complexity of the Urdu language's morphology.

Researchers started developing machine learning algorithms with linguistic rules in order to get beyond the drawbacks of strictly rule-based or statistical approaches. Riaz et al. [13] used Maximum Entropy models that included linguistic cues and contextual features to show that well-crafted feature sets might improve accuracy, but these models still needed a lot of annotated data for training, which was not easily accessible for Urdu. A hybrid method that integrated n-gram models with language rules and gazetteers is proposed by Naz et al. [14], which demonstrated the ability of hybrid systems to managing the morphological complexity and ambiguity inherent in Urdu by achieving notable gains in precision and recall. As deep learning progressed, it open new possibilities for NER. Deep Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks and their bidirectional variations (BiLSTM), were proposed by Khan et al. [15] to capture contextual information and long-range dependencies in text. Their models performed better than conventional machine learning models by combining word embeddings and part-of-speech tags, demonstrated the efficiency of deep learning in managing intricate language patterns. Kazi et al. [15] tested with different embeddings, including FastText and multilingual BERT (mBERT), and improved BiLSTM models with different experiments. The F1-score for Urdu NER was considerably raise by their method, demonstrating the advantages of deep architectures and pretrained embeddings. The KPU-NE corpus, created by Malik [17], greatly expanded the quantity of annotated Urdu text accessible for NER study. Ahmed et al. [18] addressed the lack of labeled dataset by using data augmentation approach and utilizing BERT embeddings. They achieved state-of-the-art results by optimizing pretrained BERT models using supplemented datasets, highlighting the importance of data diversity and volume in model training. Bahad et al. [19] showed that multilingual transformers might provide competitive performance, especially when adjusted with language-specific data, They did fine-tuning on XLM-RoBERTa for Indian languages, including Urdu.

The recent researches has investigated transformer-based architectures pretrained on multilingual data. Large-scale pretraining on multilingual corpora helps these transformer-based models capture cross-lingual representations, which are useful for low-resource languages. One of the most important areas of recent research has been addressing the lack of annotated data. Training datasets have been artificially expanded through the use of data augmentation approaches, such as Contextual Word Embeddings Augmentation (CWEA), which was utilized by Anam et al. [20]. These techniques improve model generalization and lessen overfitting to sparse training data by producing a variety of utterances and circumstances.

Similarly, research on Sindhi Named Entity Recognition has advanced a bit but not as much as Urdu Language. Ali et al. [21], identifies important obstacles such the absence of capitalization cues, the rich morphological structure, and the lack of annotated datasets. Early works, such as those by Hakro et al. [22], developed foundational rule-based systems, achieving high accuracy. Jumani et al. [23] extended these efforts using gazetteer and rule-based tagging, achieving 98.71% accuracy. Significant progress was made with SiNER by Ali et al. [24], introducing a 1.35M token dataset annotated with BIO tagging, enabling models like BiLSTM-CRF to achieve an F1-score of 89.16%. Subsequent advancements by Ali et al. [25][26] which integrated context-aware neural models and multitasking frameworks using self-attention and adversarial learning, achieving F1-scores exceeding 91%. More recently, Ali et al. [27] evaluated Sindhi word embeddings, showcasing Skip-Gram as the most effective, surpassing other embeddings like GloVe and fastText. Emerging trends indicate a shift from rule-based methods to neural approaches, leveraging embeddings and multitasking. However, gaps remain in exploring transformer-based architectures and cross-lingual learning with languages like Urdu, presenting opportunities for further research.

Researchers have also explored leveraging resources from related languages through cross-lingual transfer and multilingual models. Jean [28] explored the how Cross-lingual transfer learning influences multilingual pretrained models like mBERT and XLM-R to enhance NLP performance in low-resource languages by transferring knowledge from high-resourced similar languages. Recent advancements in fine-tuning and domain adaptation techniques have further improved cross-lingual transfer learning effectiveness, enabling more equitable language technology development. Kadidam [29] integrates BERT, RoBERTa, CNN, and LSTM with language-specific MLPs to enhance cross-lingual Named Entity Recognition (NER) for German and Japanese, demonstrating improved contextual understanding and classification accuracy. His approach highlighted the importance of dataset analysis and iterative optimization in refining model performance across linguistic domains. Wang et al. [30] proposed a cross-lingual NER approach using an attention mechanism and adversarial training to transfer knowledge from high-resource to low-resource languages. They performed experiments on English-Chinese datasets demonstrate a significant improvement in NER performance. Wu et al. [31] and Li et al. [32] developed frameworks that combine model transfer and data transfer, leveraging unlabeled data in the target language through enhanced knowledge distillation and reinforcement learning for instance selection. Zhou et al. [33] proposed ConNER, which combines translation-based and dropout-based consistency training to reduce overfitting on source language data and improve adaptability in target languages. These advanced techniques demonstrate the ongoing efforts to overcome data scarcity and linguistic barriers in low-resource NER.

Several previous works have developed large datasets for Named Entity Recognition (NER) in Urdu and Sindhi, focusing on training models for these languages. For Urdu, notable datasets include Jahangir et al.[4] with 31,860 words, IJCNLP2008 [5] with 40,408 words, UNER [34] with 48,673 words, MK-PUCIT [35] with 652,852 words, and UNER-I [36] with 58,633 words. These datasets feature entity types such as Person, Location, Organization, Date, Time, Number, Designation, Abbreviation, Brand, Title Person, Title Object, Measures, and Terms. For Sindhi, the prominent datasets are SiNER [24] with 1,358,691 words and Hakro et al.'s [22] dataset with 29,749 named entities. These include categories like Person, Location, Organization, Date/Time, Number, Designation, Abbreviation, Brand, Title Person, Title Object, Measures, Terms, Geopolitical Entities, Buildings, Nationalities, Events, Languages, and Artworks. Together, these datasets comprehensively cover a range of entity types relevant to their respective languages.

Recognizing the linguistic similarities between Urdu and Sindhi, we investigate the impact of pre-training models on machine-labeled data from both languages. This cross-lingual transfer approach to demonstrates significant performance gains regarding the benefits of leveraging language families in low-resource NER. Our experiments assess various configurations, providing insights into how machine-labeled data and multilingual pre-training influence NER performance in low-resource languages.

DATA PREPREATION

To create a high-quality annotated dataset for Named Entity Recognition (NER) in Urdu and Sindhi, we started by collecting text from well-known Urdu and Sindhi online newspapers through web scraping. The data which was collected by scrapping was further cleaned by using automated python scripts. This method helped us collect a vast verity of news data which has the required entity types used in various contexts. To address script and encoding issues, the UTF-8 encoding was used as standard script for both languages, preserving the accurate representation of diacritics, ligatures, and special characters. The cleaned and tokenized data was stored in structured Excel files. Each file contained three primary columns: Sentence ID (a unique identifier for each sentence), Words (tokenized text segmented according to language-specific rules) and Labels (a column reserved for entity annotations, formatted using the Inside-Outside-Beginning [IOB] tagging schema). The labels column was pre-filled with "O" (Outside) by default, allowing annotators to focus on tagging relevant entities.

We recruit two language experts who were fluent in Urdu and Sindhi for annotating the datasets. They used a set of predefined entity categories specific to these languages, to carefully label the dataset. We employed the Inside-Outside-Beginning (IOB) tagging scheme, a widely accepted method in NER labeling tasks. This scheme helps clearly mark the boundaries of multi-token entities with:

- B- (Beginning): Used in start of an entity
- (Inside): Used in mid part of an entity
- (Outside): Used for tokens that do not belong to any named entity.

Table I provides details of entity category, along with the corresponding IOB tags, description of entity, and examples from both languages.

Table 1: Entity Types with IOB Tags, Descriptions, and Examples in Urdu and Sindhi

Entity Type	Tag Types	Description	Urdu Examples	Sindhi Examples
Person	*B-PER, I-PER*	*Names of individuals, often denoting human actors in text.*	*علی، فاطمہ، عمران خان*	*علي، فاطمه، آصف زرداري*
Location	*B-LOC, I-LOC*	*Geographical locations, including cities, countries, and landmarks.*	*کراچی، پاکستان، اسلام آباد*	*ڪراچي، پاڪستان، حيدرآباد*
Organization	*B-ORG, I-ORG*	*Names of organizations, institutions, or companies.*	*یونیسیف، اقوام متحدہ، پی ٹی وی*	*يونيسيف، سنڌ ٽيليويزن*
Position	*B-POS, I-POS*	*Titles or positions associated with people, often job-related.*	*وزیر اعظم، چیف جسٹس، صدر*	*وزيراعظم، چيف جسٽس، صدر*
Product	*B-PRD, I-PRD*	*Names of products, including brands or specific items.*	*سام سنگ موبائل، ٹویوٹا کار*	*ٽيوٽا ڪار، سامسنگ ٽي وي*
Date/Time	*B-DAT, I-DAT*	*References to specific dates, times, or periods.*	*12 مارچ 2023، صبح 9 بجے*	*12 مارچ 2023، صبح 9 وڳي*
Event	*B-EVT, I-EVT*	*Names of events, including holidays, festivals, or specific occasions.*	*عید، یوم پاکستان، کرسمس*	*عيد، پاڪستان ڏي، ڪرسمس*
Color	*B-COL, I-COL*	*Color names, often used in descriptions.*	*سبز، سرخ، نیلا*	*سائو، ڳاڑهو، نيرو*
Food	*B-FOOD, I-FOOD*	*Names of food items, dishes, or culinary products.*	*بریانی، روٹی، کباب*	*برياني، مانيء، ڪباب*
NORP	*B-NORP, I-NORP*	*Nationalities and religious groups.*	*پاکستانی، مسلمان، ہندو*	*پاڪستاني، مسلمان، هندو*
Number	*B-NUM, I-NUM*	*Numerical values, including quantities or ordinal references.*	*دو، پانچ، سترہ*	*ٻه، پنج، ستره*
Measure	*B-MEA, I-MEA*	*Units of measurement, such as distance, weight, or time duration.*	*کلو، میٹر، گھنٹہ*	*ڪلو، ميٽر، ڪلاڪ*

We applied the IOB format across 12 different entity categories, including traditional entities like people and locations, as our newly introduced categories such as colors and foods. The reason for adding color and food is that they were appearing many times in news data and we felt the need to enhance the traditional NER in Urdu and Sindhi. The IOB tagging method across both datasets ensures reliable annotation and maintain consistency of data formats which is very important for effective cross-lingual model training and evaluation. The final datasets were saved in CSV formats.

MACHINE-LABELED DATA GENERATION

After the creation of a human-annotated dataset of high-quality for Named Entity Recognition (NER) in Urdu and Sindhi, we used a machine-labeled data generation technique using Conditional Random Fields (CRF) with a bagging-based ensemble approach. The process involves training multiple CRF models on random subsets of the labeled data, generating token-level predictions on unlabeled data, and applying a voting mechanism with confidence filtering. Figure 1 illustrates the workflow of our CRF bagging and voting approach for machine-labeled data generation, detailing each stage from data preparation through to the final aggregation of labels.

The machine-labeled data generation process follows these main steps:

1) Feature Extraction: We extract a set of token-level features to represent each word, including basic attributes (e.g., token length, presence of digits), character-level n-grams (prefixes and suffixes), FastText embeddings (reduced with PCA), and contextual information from surrounding words. These features enable the CRF models to effectively learn relationships between tokens and their corresponding labels.

2) Data Sub setting with Bagging: We prepared multiple subsets of the human-annotated dataset. Using bagging, each CRF model was trained on a unique random subset, representing approximately 80% of the original data. This subset sampling technique ensures diversity among models by exposing each model to slightly different data distributions, which reduces overfitting and improves generalization across models.

3) CRF Model Training on Subsets: Each subset, while capturing a similar distribution to the full dataset, has a unique selection of samples. This diversity encourages each CRF model to learn slightly different decision boundaries, which collectively contribute to a more robust ensemble when the models are later combined. For each subset, we trained a CRF model using the extracted token-level features. Each CRF model is trained independently, generating a mapping between the token-level features and the entity labels. The training process is configured with slightly varied regularization parameters (e.g., c1 and c2) to introduce additional variance, allowing each model to have a unique bias in learning.

4) Token-Level Prediction on Unlabeled Data: Once trained, each CRF model independently makes token-level predictions on the unlabeled dataset. Given an unlabeled sentence, each model predicts an entity label for every token in that sentence based on its learned patterns.

5) Voting and Confidence Filtering: To determine the final label for each token, we applied a voting process across the predictions from all CRF models. For each token position, we aggregate the predicted labels from each model and count how often each label appears. The label with the highest frequency (majority vote) is selected as the candidate for the final label. This process ensures that the label chosen is the one that the majority of models agree upon, which tends to increase reliability.

6) Confidence Filtering: We apply a confidence threshold by using the rule that if the most frequent label meets or exceeds a predefined threshold (e.g., 80% of the models), it is selected as the final label for that token. For instance, if 8 out of 10 models agree on a label, this label would be assigned since it meets an 80% confidence threshold. If no label meets this threshold, the token is labeled with a other tag ('O'), which means it is not an entity.

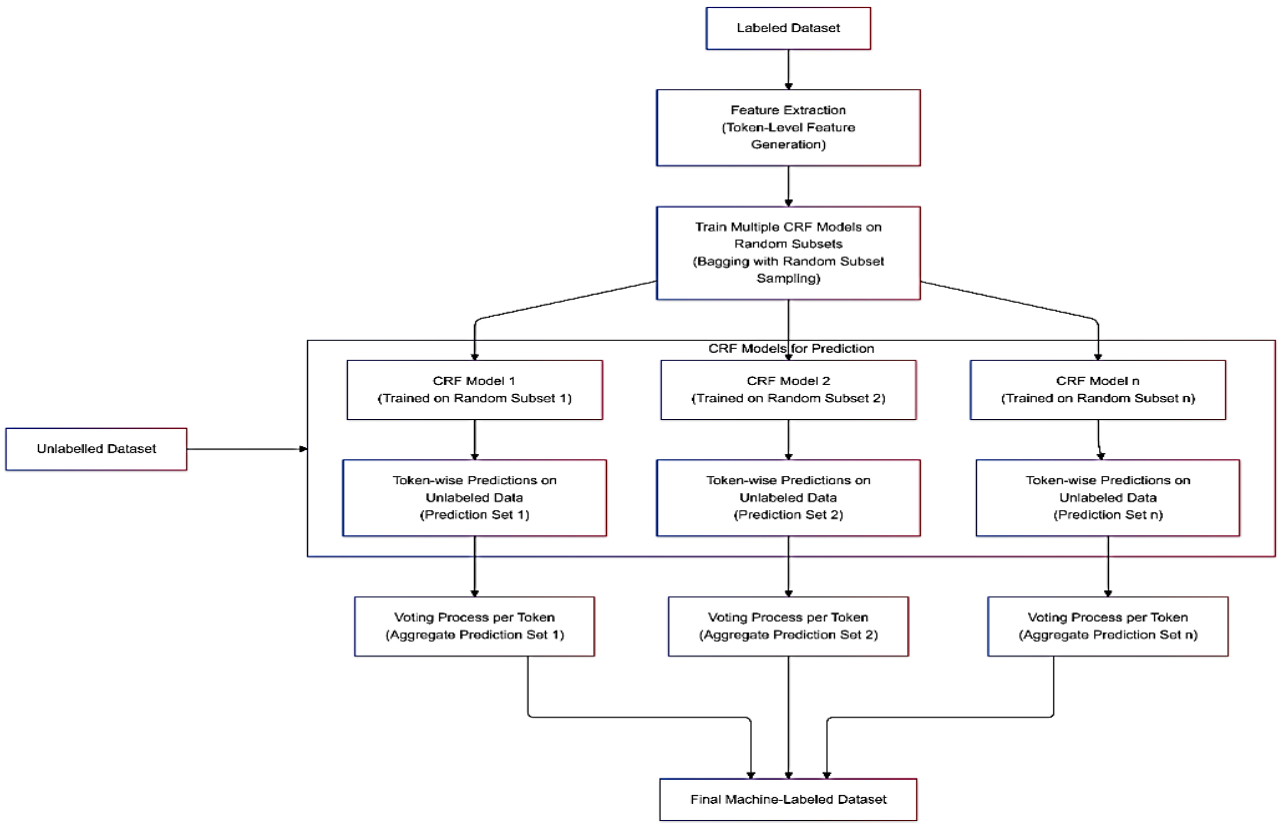

Figure 1: Workflow of CRF bagging and voting approach for machine-labeled data generation

7) Deciding on Ties and Ambiguities: In cases where two or more labels have same frequencies but do not meet the confidence threshold, the other tag ('O') is assigned to minimize the risk of incorrect labeling. In this way we can avoid ambiguous labels and ensures that only high-confidence predictions are added in the machine-labeled dataset.

All these steps together make the Final Machine-Labeled Dataset, which improves our training data with reliable and high-quality labels. The bagging and voting methods use the combined predictions of many CRF models. This way, we reduce the biases or mistakes from individual models and create a more consistent and accurate machine-labeled dataset for Named Entity Recognition in Urdu and Sindhi.

To give an overview of the data used in both the labeled and machine-labeled stages, Table 2 shows a comparison of the four datasets: Human Labeled Urdu (HLU), Human Labeled Sindhi (HLS), Machine Labeled Urdu (MLU), and Machine Labeled Sindhi (MLS). This table lists the total number of sentences, words, and non-entity tokens ("O") for each dataset, along with the counts for each type of entity.

The machine-labeled datasets (MLU and MLS) contain many more words, and sentence counts due to automated expansion with minimal human effort. Entity distribution displays same patterns as observed in the human-labeled datasets. This comparison highlights that the increased data volume provided by machine labeling looks promising and can enhance NER performance through transfer learning and cross-lingual analysis.

Table 2: Comparison of Human Labeled and Machine Labeled Datasets

Characteristics	HLS	HLU	MLS	MLU
Total Sentences	*1009*	*2128*	*18620*	*17265*
Total Tokens	*34315*	*47046*	*576052*	*557480*
Non-Entity ('O')	*28625*	*37006*	*499046*	*495293*
LOC	*732*	*600*	*9966*	*7284*
NUM	*608*	*481*	*6845*	*8342*
FOOD	*485*	*716*	*397*	*527*
PER	*436*	*514*	*6817*	*4900*
POS	*415*	*435*	*7692*	*5253*
ORG	*346*	*311*	*6132*	*4828*
MEA	*262*	*169*	*2441*	*2647*
DAT	*138*	*457*	*3135*	*2080*
PRD	*130*	*502*	*247*	*468*
COL	*120*	*558*	*13*	*40*
EVT	*98*	*621*	*104*	*904*
NORP	*52*	*138*	*292*	*2555*

EXPERIMENT SETUP

The are two powerful multilingual models, firstly Multilingual BERT (mBERT) and secondly XLM-RoBERTa. We used them both to improve NER for the low-resource languages Sindhi and Urdu. The model mBERT, pre-trained on 104 languages, is designed to capture universal language representations, making it suitable for multilingual tasks. The model XLM-RoBERTa, a more recent model trained on a larger Common Crawl dataset across 100 languages, has shown strong performance in multilingual contexts, particularly for cross-lingual tasks. Our experiments were conducted on a T4 GPU with high RAM provided by Google Colab. We experimented with multilingual BERT (mBERT) and XLM-RoBERTa, setting a learning rate of 5e-5 and optimizing with the AdamW optimizer. Both pre-training and fine-tuning phases were run for 3 epochs with a batch size of 16, and the maximum sequence length was set to 128 tokens. A warmup ratio of 0.06 was applied, allowing for a gradual increase in the learning rate at the start of training. Fine-tuning included evaluation at each epoch to monitor model performance. Each model variant's outputs were saved to designated directories. Machine-labeled datasets created with a CRF-based approach using FastText embeddings were used for initial pre-training, followed by fine-tuning on human-labeled datasets (HLU for Urdu and HLS for Sindhi) with an 80-20 split for training and evaluation.

RESULTS AND DISCUSSIONS

Table 3 presents the precision, recall, and F1 scores for mBERT and XLM-RoBERTa when fine-tuned on the human-annotated datasets—Human Labeled Sindhi (HLS) and Human Labeled Urdu (HLU)—following pre-training on various configurations of machine-labeled data. The machine-labeled data, created through our ensemble labeling approach, proved essential in achieving substantial performance gains.

In Sindhi, we observed significant improvements with XLM-RoBERTa, which consistently outperformed mBERT in all configurations. Notably, XLM-RoBERTa's F1 score increased from 0.302 (without pre-training) to 0.681 after pre-training on both Machine Labeled Sindhi (MLS) and Machine Labeled Urdu (MLU), representing an increase of approximately 125%. This substantial improvement underscores the effectiveness of incorporating machine-labeled data and cross-lingual knowledge transfer from Urdu.

Also, mBERT's performance for Sindhi got much better, with its F1 score going up from 0.402 (without pre-training) to 0.497 after pre-training on both MLS and MLU. Even though mBERT didn't do as well as XLM-RoBERTa, these results show that the model improved from the extra machine-labeled data and cross-lingual transfer, helping it work better for Sindhi NER.

For Urdu, pre-training on machine-labeled datasets made big improvements for both models. mBERT got its highest F1 score of 0.720 after pre-training on both MLS and MLU, compared to 0.681 without any pre-training—an improvement of about 5.7%. XLM-RoBERTa got an F1 score of 0.713 with MLS and MLU pre-training, better than 0.656 without pre-training. These results suggest that while both models work well on Urdu, mBERT might be a bit better, maybe because it fits better with Urdu's language features.

The improvements we saw are likely because of several things. First, the machine-labeled datasets gave us a lot more training data, which is very important for languages like Sindhi and Urdu that don't have much data. Pre-training on this data let the models see more kinds of entities, contexts, and language details, making them work better on the human-annotated datasets. Second, the cross-lingual transfer—where knowledge from Urdu helped Sindhi NER and the other way around—was helpful. Since Urdu and Sindhi share language features, pre-training on both machine-labeled datasets let the models use shared patterns and entity representations.

Table 3: Performance Comparison of mBERT and XLM-RoBERTa on Sindhi and Urdu NER Tasks

Language	Model	Pre-training	Fine-tuning	Precision	Recall	F1 Score
Sindhi	*mBERT*	*None*	*HLS*	*0.659*	*0.289*	*0.402*
		MLS		*0.671*	*0.386*	*0.490*
		MLU		*0.665*	*0.358*	*0.466*
		MLS + MLU		*0.690*	*0.389*	***0.497***
	XLM-RoBERTa	*None*	*HLS*	*0.430*	*0.233*	*0.302*
		MLS		*0.708*	*0.609*	*0.655*
		MLU		*0.713*	*0.576*	*0.637*
		MLS + MLU		*0.729*	*0.639*	***0.681***
Urdu	*mBERT*	*None*	*HLU*	*0.708*	*0.656*	*0.681*
		MLS		*0.752*	*0.687*	*0.718*
		MLU		*0.721*	*0.653*	*0.685*
		MLS + MLU		*0.757*	*0.687*	*0.720*
	XLM-RoBERTa	*None*	*HLS*	*0.686*	*0.628*	***0.656***
		MLS		*0.738*	*0.688*	*0.712*
		MLU		*0.745*	*0.654*	*0.696*
		MLS + MLU		*0.737*	*0.691*	***0.713***

Also, XLM-RoBERTa's design, made for cross-lingual tasks, probably helped it do better in Sindhi, which had less labeled data. For Urdu, mBERT's slightly better performance suggests that its multilingual training might match more closely with Urdu's grammar and meaning, helping it understand language details better.

Adding machine-labeled data made a big change, especially for Sindhi, where XLM-RoBERTa's performance more than doubled with the extra pre-training. These findings show that machine-labeled data and cross-lingual transfer greatly improve NER abilities in both languages, showing the potential of these methods to improve NLP tasks in settings where there is not much data.

CONCLUSION

In this study, we worked on the challenges of Named Entity Recognition (NER) in languages with little data, like Urdu and Sindhi. We made a new method that combines machine-made labeled data with advanced multilingual models. We added new categories like colors and foods to NER for these languages. By using a bagging-based group of Conditional Random Field (CRF) models, we created high-confidence machine-labeled datasets. This increased the training data a lot without the effort of manual labeling. Our experiments showed that pre-training Multilingual BERT (mBERT) and XLM-RoBERTa on the machine-labeled data, and then fine-tuning on the human-annotated datasets, improved NER performance in both Urdu and Sindhi. Adding machine-labeled data led to big gains, especially for Sindhi, where XLM-RoBERTa's F1 score increased by about 125%. These results highlight how effective it is to use machine-labeled data and cross-language knowledge transfer to improve NER in low-resource languages. Although our models did not achieved very high F1 score but still showed the increase in F1 score specially in Sindhi. The cross-language transfer between Urdu and Sindhi shows the advantage of using similarities between related languages to improve NLP tasks. Our method not only deals with the problem of not having enough data but also sets the stage for future research in multilingual and low-resource language processing.

FUTURE WORK

In future this methodology can be applied to other low-resource Pakistani languages, especially those within the same language family or with similar linguistic features. To validate the generalizability of our approach experiments on more languages can be performed. In future we can extend the entity categories and add more diverse and domain-specific entities, which could further enhance the utility of NER systems in various applications. To improve the F1 score the other transformer-based models, such as multilingual T5 or mBART, can be used. Also experimenting with different fine-tuning strategies, such as continual learning and domain adaptation techniques, may improve model robustness and adaptability to new data. A detailed error analysis can be conducted to identify common misclassifications. New methods can be developed interpret model decisions so, more reliable NER system can be created.

REFERENCES

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investigationes, vol. 30, no. 1, pp. 3–26, 2007.

[2] F. Ullah, A. Gelbukh, M. T. Zamir, E. M. F. Riverón, and G. Sidorov, “Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu,” *Computers*,

vol. 13, no. 10, Oct. 2024, doi: 10.3390/computers13100258.

[3] S. Naz, A. I. Umar, S. H. Shirazi, S. A. Khan, I. Ahmed, and A. A. Khan, “Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language,” Research Journal of Applied Sciences, Engineering and Technology, vol. 8, no. 10, pp. 1272–1278, 2014.

[4] F. Jahangir, M. W. Anwar, U. I. Bajwa, and X. Wang, "N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language," in Proceedings of the 2012 Conference on Language Resources and Evaluation (LREC), 2012.

[5] A. K. Singh, "Named entity recognition for South and South East Asian languages: taking stock," in Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, 2008.

[6] S. N. Khan et al., "Urdu Word Segmentation using Machine Learning Approaches," International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 377–385, 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT 2019, 2019, pp. 4171–4186.

[8] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of ACL 2020*, 2020, pp. 8440–8451.

[9] J. Kim, Y. Ko, and J. Seo, "Construction of Machine-Labeled Data for Improving Named Entity Recognition by Transfer Learning," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6006–6015.

[10] J. Wang, "Cross-lingual Transfer Learning for Low-Resource Natural Language Processing Tasks," Ph.D. dissertation, Karlsruhe Institute of Technology, 2021.

[11] K. Riaz, “Rule-based Named Entity Recognition in Urdu,” Association for Computational Linguistics, 2010.

[12] U. Singh, V. Goyal, and G. S. Lehal, “Named Entity Recognition System for Urdu,” In Proceedings of COLING, 2012, pp. 2507-2518

[13] F. Riaz, M. W. Anwar, and H. Muqades, "Maximum entropy based Urdu named entity recognition," in 2020 International Conference on Engineering and Emerging Technologies (ICEET), Feb. 2020, pp. 1–5.

[14] S. Naz, A. I. Umar, and M. I. Razzak, "A hybrid approach for NER system for scarce resourced language-Urdu: Integrating n-gram with rules and gazetteers," Mehran University Research Journal of Engineering & Technology, vol. 34, no. 4, pp. 349–358, Oct. 2015.

[15] W. Khan, A. Daud, F. Alotaibi, N. Aljohani, and S. Arafat, "Deep recurrent neural networks with word embeddings for Urdu named entity recognition," ETRI Journal, vol. 42, no. 1, pp. 90–100, Feb. 2020.

[16] S. Kazi, M. Rahim, and S. Khoja, "A deep learning approach to building a framework for Urdu POS and NER," Intelligence (AI), vol. 31, pp. 68, Jan. 2023.

[17] M. K. Malik, "Urdu Named Entity Recognition and Classification System Using Artificial Neural Network," International Journal of Advanced Computer Science and Applications, vol. 8, no. 9, pp. 139–144, 2017.

[18] 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.

[19] S. Bahad, P. Mishra, K. Arora, R. C. Balabantaray, D. M. Sharma, and P. Krishnamurthy, "Fine-tuning Pre-trained Named Entity Recognition Models For Indian Languages," arXiv preprint arXiv:2405.04829, May 2024.

[20] R. Anam, M. W. Anwar, M. H. Jamal, U. I. Bajwa, I. D. Diez, E. S. Alvarado, E. S. Flores, and I. Ashraf, "A deep learning approach for Named Entity Recognition in Urdu language," Plos One, vol. 19, no. 3, Mar. 2024, Art. no. e0300725.

[21] 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.

[22] D. N. Hakro, M. A. Hakro, and I. A. Lashari, "Sindhi Named Entity Recognition (SNER)," Government: Research Journal of Political Science, vol. 5, Jan. 2017.

[23] A. K. Jumani, M. A. Memon, F. H. Khoso, A. A. Sanjrani, and S. Soomro, "Named entity recognition system for Sindhi language," in Emerging Technologies in Computing: First International Conference, iCETiC 2018, London, UK, August 23–24, 2018, Proceedings 1, Springer International Publishing, 2018, pp. 237–246.

[24] 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, May 2020, pp. 2953–2961.

[25] W. Ali, J. Kumar, Z. Xu, R. Kumar, and Y. Ren, "Context-Aware Bidirectional Neural Model for Sindhi

Named Entity Recognition," Applied Sciences, vol. 11, no. 19, p. 9038, Sep. 2021.

[26] 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.

[27] 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, Aug. 2024.

[28] G. Jean, “Cross-Lingual Transfer Learning for Low-Resource NLP Tasks: Leveraging Multilingual Pretrained Models,” 2023. [Online]. Available: https://www.researchgate.net/publication/387003964

[29] V. V. Kadidam, “Cross Lingual Named Entity Recognition using Deep Learning,” 2024.

[30] H. Wang, L. Zhou, J. Duan, and L. He, “Cross-Lingual Named Entity Recognition Based on Attention and Adversarial Training,” *Applied Sciences (Switzerland)*, vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042548.

[31] Q. Wu, Z. Lin, B. F. Karlsson, B. Huang, and J. G. Lou, "Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data," arXiv preprint arXiv:2007.07683, Jul. 2020.

[32] Z. Li, C. Hu, X. Guo, J. Chen, W. Qin, and R. Zhang, "An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), May 2022, pp. 170–179.

[33] R. Zhou, X. Li, L. Bing, E. Cambria, L. Si, and C. Miao, "ConNER: Consistency training for cross-lingual named entity recognition," arXiv preprint arXiv:2211.09394, Nov. 2022.

[34] 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.

[35] 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. H. Saeed and S. Qureshi, "Improving Urdu NER by Combining Machine Learning and Rule-Based Approaches," in Proceedings of the 2017 International Conference on Asian Language Processing (IALP), Nov. 2017, pp. 34–37.