

Enhancing NER Performance in Low-Resource Pakistani Languages using Cross-Lingual Data Augmentation

Toqeer Ehsan, Thamar Solorio

Department of Natural Language Processing,
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI),
Abu Dhabi, United Arab Emirates
{toqeer.ehsan, thamar.solorio}@mbzuai.ac.ae

Abstract

Named Entity Recognition (NER), a fundamental task in Natural Language Processing (NLP), has shown significant advancements for high-resource languages. However, due to a lack of annotated datasets and limited representation in Pre-trained Language Models (PLMs), it remains understudied and challenging for low-resource languages. To address these challenges, we propose a data augmentation technique that generates culturally plausible sentences and experiments on four low-resource Pakistani languages; Urdu, Shahr-mukhi, Sindhi, and Pashto. By fine-tuning multilingual masked Large Language Models (LLMs), our approach demonstrates significant improvements in NER performance for Shahr-mukhi and Pashto. We further explore the capability of generative LLMs for NER and data augmentation using few-shot learning.

1 Introduction

The performance of Named Entity Recognition (NER) in low-resource languages faces challenges due to the scarcity of annotated datasets and insufficient coverage in masked Large Language Models (LLMs) (Subedi et al., 2024). Causal LLMs, on the other hand, demonstrate their performance by achieving moderate scores for NER (Chen et al., 2023; Ye et al., 2023). These challenges make it difficult to develop effective NLP applications and highlight the need of focused effort to improve the applicability of these models on available datasets for low-resource languages.

Data augmentation approaches could be effective to enhance the NER datasets for low-resource languages. One such approach is the Easy Data Augmentation (EDA) (Wei and Zou, 2019), that offers simple and effective techniques, including synonym replacement, random insertion, random swap, and random deletion (Khalid et al., 2023; Liu and Cui, 2023; Litake et al., 2024). However, EDA

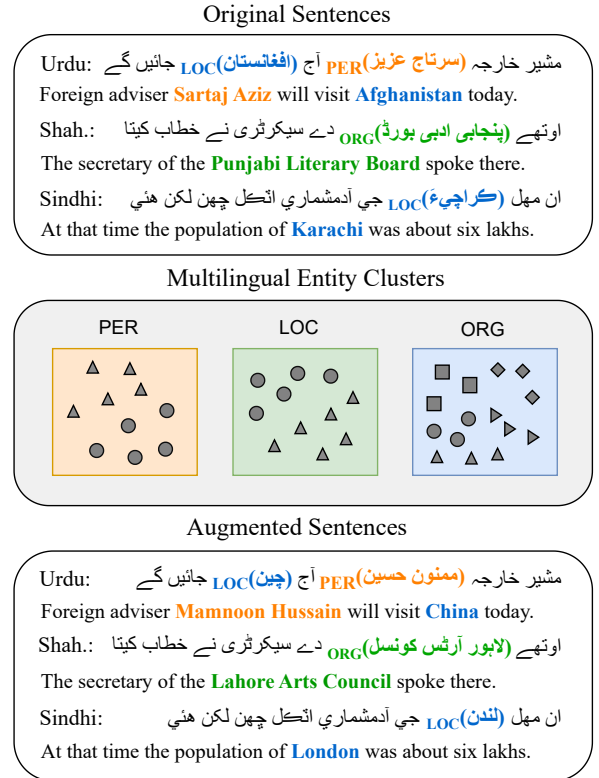


Figure 1: Examples of clustering-based data augmentation applied to three sample sentences. Entity mentions are represented in orange, blue and green colors.

can produce linguistically implausible text lacking verbal agreement based on gender and number. Additionally, EDA may produce out-of-context or offensive data for culturally sensitive content. This can affect the generalizability and learning of NER models. We aim to enhance NER performance for Pakistani low-resource languages by employing effective data augmentation as shown in Figure 1.

Four Urdu sentences are shown in Figure 2, illustrating the problem of implausibility. Urdu, Shahr-mukhi and Sindhi require verbal agreements, and augmenting entities from

- 1- (مومنہ) **PER** گورنمنٹ گرلز ہائی سکول میں پڑھتی ہے۔
(Momina)PER studies in Government Girls High School.
- 2- (چوہدری محمد سرور) **PER** آج صبح لاہور پہنچ گئے ہیں۔
(Chaudhry Muhammad Sarwar)PER has reached Lahore this morning.
- 3- (تحریک منہاج القرآن) **ORG** کا سالانہ عالمی مذہبی اجتماع کل ہو گا۔
The annual global religious gathering of **(Minhaj-ul-Quran Movement)ORG** will be held tomorrow.
- 4- کل (لاہور آرٹس کونسل) **ORG** نے انٹرنیشنل ڈانس ڈے کا پروگرام منعقد کیا
(Lahore Arts Council)ORG organized the program of International Dance Day yesterday.

Figure 2: Sample Urdu sentences for the analysis of EDA. Named entities are highlighted in bold.

the sentences 1 and 2, could result in disagreements. *Momina* (Nom.Fem.Sg) is a feminine name that has agreement with the verb *paRHti* (study.Hab.Fem.Sg), while *Chaudhry Muhammad Sarwar* (Nom.Masc.Sg) is a masculine name that agrees with the verb *gaE* (go.Past.Masc.Sg.Hon). Replacing these named entities can produce implausible text; for instance, the sentence *Chaudhry Muhammad Sarwar studies in Government Girls High School* would violate the verbal agreement rules of the language. The named entities in the last two sentences are considered opposites within the community, and replacing such named entities can produce text that is very offensive to the native community. The generated sentences remain grammatically correct but create contextual ambiguity.

We propose a cross-lingual data augmentation technique by clustering named entities as shown in Figure 1. This technique improves the quality of culturally sensitive content and grammar of the augmented text. We performed unsupervised entity clustering and entity replacement by aligning clusters for the source and candidate named entities of each type. NER experiments were conducted for low-resource settings as well as for entire datasets. We compared the results with EDA-based and generative augmentation methods for mono- and multilingual settings by fine-tuning the Glot500 (Imani et al., 2023) and XLM-RoBERTa (Conneau et al., 2019) models. Shahmukhi and Pashto datasets demonstrated significant improvements, producing F₁ scores of 88.06 and 88.29 with increases of 5.53 and 1.81 points, respectively.

Zero- or few-shot learning is relevant in low-resource scenarios where even augmented datasets are limited in size. We explore the capabilities of causal LLMs to perform NER and data augmenta-

tion for our low-resource languages using few-shot learning. The key contributions of this paper are as follows:

- We propose a novel cross-lingual augmentation technique that uses cluster dictionaries to produce culturally and linguistically plausible augmentations.
- We demonstrate the effectiveness of the proposed technique in multilingual NER experiments by utilizing cross-lingual representations.
- We provide insights into the potential of causal LLMs to perform NER and data augmentation for low-resource languages using few-shot learning.

2 Related Work

Manually annotated corpora are crucial for achieving state-of-the-art results in NER (Mayhew et al., 2023). Cross-lingual transfer also supports generalization and enhances the performance of models (Ding et al., 2024; Mo et al., 2024; Cotterell and Duh, 2024; Le et al., 2024; Hu et al., 2020). Data augmentation techniques enhance the size and learning capabilities of datasets for low-resource languages (Litake et al., 2024; Ye et al., 2024; Lancheros et al., 2024). For the task of NER, three data augmentation methods are mainly used; Easy Data Augmentation (EDA) (Wei and Zou, 2019) and its variants, translation-based methods and generative LLMs. EDA-based techniques demonstrate enhanced NER performance for low-resource languages (Litake et al., 2024). The data augmentation quality can be enhanced by using contextualized word embeddings (Torres et al., 2024) and cosine similarity (Bartolini et al., 2022).

Data augmentation based on back-translation has shown improvements for code-switched NER (Sabty et al., 2021). The translation-based data augmentation technique that performs cross-lingual entity augmentation also improves the performance of NER models (Liu et al., 2021; Lancheros et al., 2024; Chen et al., 2022).

The capabilities of causal LLMs are being explored for data augmentation (Evuru et al., 2024; Ye et al., 2024) and underlying NLP tasks such as NER (Naguib et al., 2024; Villena et al., 2024; Lu et al., 2024). Generative data augmentation techniques have demonstrated improvements (Evuru et al., 2024; Liu et al., 2022; Ye et al., 2024).

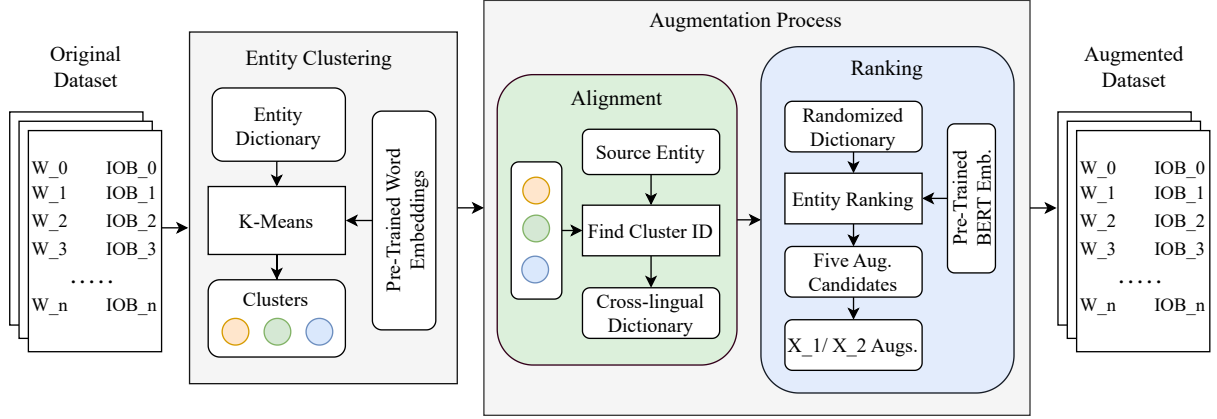


Figure 3: Cluster-based data augmentation process, which contains three phases. The entity clustering phase extracts unsupervised clusters for each entity type, alignment phase aligns cluster dictionaries with respect to the source (original) entities and the final phase ranks the source entity mentions with the best candidate. The original dataset corresponds to the manually annotated dataset, while the augmented dataset is the updated version obtained through the augmentation process.

Masking-based generative methods have produced better NER results by generating more plausible data augmentations (Song et al., 2024).

Causal LLMs are further employed to perform NER with zero- and few-shot learning (Naguib et al., 2024; Villena et al., 2024) as an alternative approach to data augmentation. These models are also progressing in various text domains (Lu et al., 2024; Monajatipoor et al., 2024). These advancements highlight the need to investigate the capabilities of these models for low-resource languages.

3 Cross-Lingual Data Augmentation

The languages selected in this work are topologically related and culturally similar. In terms of named entities, they share similar names, locations and organizations. Given these similarities, cross-lingual representation could be helpful in improving the performance of NER for the regional languages. Additionally, data augmentation techniques have shown improvements for low-resource languages, but EDA-based methods are blunt and may produce culturally offensive and/or ungrammatical sentences by replacing entities with the other entities of the same type without any additional semantic information. Out of 100 randomly selected sentences, 32 instances of verbal disagreements and three of sensitive religious named entities were found. These percentages estimate the occurrence of such issues in the augmented data. To address these issues, we propose a data augmentation technique that generates more sensible sentences and produces competitive NER performance

for the selected low-resource languages. The next section describes our proposed technique, followed by descriptions of EDA-based random replacement and generative approaches.

3.1 Cluster-based_{Aug}.

We propose a hybrid data augmentation technique inspired by EDA, combined with the application of unsupervised entity clustering. The technique consists of three phases; entity clustering, alignment, and ranking as illustrated by Figure 3.

Entity Clustering Named entities were clustered using context-free word embeddings from pre-trained models (Grave et al., 2018; Tehseen et al., 2023), where each word has a single embedding regardless of its context which are helpful in clustering process. We employed the K-Means clustering algorithm to cluster entities based on their embeddings and cosine similarity. While clustering is an unsupervised method, we interpreted these clustering representing specific categories for each entity type. To evaluate the effectiveness of the approach, we manually assessed the unsupervised clustering of 200 entities for each entity type in Urdu. The person and location types were categorized into two clusters; masculine and feminine for persons, and country/continent and city/places for locations. In contrast, named entities from the organization type were grouped into ten clusters; entertainment, financial, health/education, justice/govt, news, politics, religious, water/electricity, abbreviations and miscellaneous. The accuracies for

correctly clustered named entities were 86.0% for persons, 87.5% for locations, and 84.5% for organizations, as determined through manual evaluation. The K-means clustering approach was implemented using NLTK’s *KMeansClusterer* to categorize named entity embeddings into distinct groups. The clustering process utilized cosine distance as the similarity metric, ensuring that entities with similar vector representations were grouped together effectively. To enhance the stability and robustness of the clustering process, we performed 25 repetitions. For clustering, separate dictionaries of unique named entities were created based on the splits of annotated training sets.

We achieve a single feature vector by averaging the vectors for each token in an entity of the location and organization types. However, person names have a specific pattern in Pakistani culture. The first name usually belongs to the individual, followed by a family name. A feminine first name is typically followed by a masculine name, that could be the name of the father, tribe, caste, or creed. For instance, in the entity mention *Madiha Khalid*, *Madiha* is the feminine name followed by the masculine name *Khalid*. Similarly, many names, particularly masculine names, begin with a title representing a designation, tribe, caste, or creed. We prepared a list of these titles to filter them out and used first names to obtain feature vectors. This approach improved the performance of clustering.

Alignment The prepared clusters are aligned between the source and candidate entity mentions. The source entity refers to the original entity mention in the dataset, while the candidate entity is the one selected to replace the source entity. In the alignment phase, the cluster ID of the source entity is determined by looking it up in the manually identified clusters. A dictionary containing unique named entities from the corresponding cluster is then passed to the next phase.

Ranking The ranking procedure is performed in two steps. In the first step, an entity is selected from a randomized cluster dictionary by computing highest cosine similarity with respect to the source entity mention. Unlike the clustering process, contextualized word embeddings from Glot500-base, which has data coverage of all our selected languages, are used to select similar candidate entities. This step generates five augmented sentences for each original sentence. In the second step, micro F_1 score is computed for augmented sentences to

assess their plausibility, using Glot500-base model fine-tuned on multilingual datasets. This pretrained model automatically validates each generated candidate. The tokens of each augmented sentence are fed into the model to predict the named entities. The sentence with the highest F_1 score is selected to be a part of the augmented dataset. The F_1 score is computed by treating the model output as the predicted annotation, while the manually annotated named entities in the augmented sentence serve as a reference in the process. We further prepared multiple augmented datasets by including one sentence with the highest score (X_1), two sentences with top two scores (X_2) and all augmented sentences with an F_1 score of 1.0.

3.2 Random Replacement (EDA-RR_{Aug.})

The random replacement data augmentation is a straightforward approach which is based on EDA methods (Wei and Zou, 2019). The augmentation process has two steps; 1) take all sentences in the training data with labeled named entities, 2) for each entity mention in a sentence, replace it with a named entity of the same type. The second step continues until all entity mentions in a sentence are replaced randomly. As a result, a new augmented dataset is produced, which is added to the training set to enhance its size and diversity. This method is simple and efficient to implement, but it may produce contextually implausible text that could be incorrect or offensive to the community.

3.3 Generative_{Aug.}

To add the contextual information in the data augmentation, we performed generative data augmentation using LLaMA3 (Touvron et al., 2023) with few-shot learning. The approach is similar to the entity-level augmentation proposed by Ye et al. (2024). We employed instruction-finetuned version of LLaMA3 (LLaMA3-8B-Instruct). We selected LLaMA3 due to its open-access nature and strong few-shot learning capabilities. LLaMA3 has been trained on a diverse multilingual corpus, but its direct exposure to Pakistani languages is limited. However, Urdu is a widely spoken language with significant online resources, LLaMA3 demonstrates moderate performance in generating Urdu text. We constructed a prompt by providing three examples containing each entity type and instructed the model to replace entity mentions with similar entities. The augmentation was performed for low-resource training sets due to

time and resource constraints. The prompt that we used for data augmentation is given below:

You are an expert in augmenting data for named entities for Urdu language. The input contains the ORIGINAL TEXT followed by the AUGMENTED TEXT. Perform augmentation by replacing named entities with new entities of the same type and return the AUGMENTED TEXT. Three examples are given for your reference:

EXAMPLE 1:

ORIGINAL TEXT:

AUGMENTED TEXT:

4 Languages and Datasets

Pakistan is home to many widely spoken languages, each with unique linguistic characteristics and cultural significance. Urdu is the national language of Pakistan that has 232 million speakers worldwide. Shahmukhi (Punjabi), Sindhi, and Pashto have 67, 30 and 40 million speakers, respectively (Eberhard and Fennig, 2024). These languages pose several challenges for the task of NER, such as absence of capitalization, contextual ambiguity, flexible word-order, and agglutinating nature (Khalid et al., 2023; Ehsan and Hussain, 2021; Ahmed et al., 2024). The statistics of the selected datasets are shown in Table 1. Despite the larger sample sizes in Shahmukhi, Sindhi, and Urdu datasets, they face limited domain coverage, incomplete NER labels, low sentence-to-entity ratio, and noisy annotations, underscoring their low-resource status. The MK-PUCIT, Shahmukhi and SiNER datasets were released without validation sets; therefore, we used 10% of the train sets for validation.

Urdu: Being in the *Vital* category (Eberhard and Fennig, 2024), Urdu is relatively resource-rich compared to the regional languages. Several NER datasets are available for Urdu with different data annotations and sizes (Khana et al., 2016; Hussain, 2008; Jahangir et al., 2012; Malik, 2017). However, we experimented with Urdu-Wikiann (Rahimi et al., 2019; Lovenia et al., 2024) and MK-PUCIT (Kanwal et al., 2019), which are larger datasets annotated with coarse-grained named entities; person, location and organization.

Shahmukhi: There is only one NER dataset available for Shahmukhi, which has been annotated using person, location, and organization types

Lang./Dataset	Type	Train	Test	Val.
Urdu / Urdu-Wikiann	PER	6,839	363	340
	LOC	6,891	334	352
	LOC	6,891	334	352
	ORG	6,759	323	327
	# Sents.	20,000	1,000	1,000
Urdu / MK-PUCIT	PER	11,965	5,215	–
	LOC	23,880	8,380	–
	ORG	8,665	3,014	–
	# Sents.	24,080	16,609	–
Punjabi / Shahmukhi	PER	4,655	1,957	–
	LOC	1,855	648	–
	ORG	538	236	–
	# Sents.	13,547	5,821	–
Sindhi / SiNER	PER	12,894	5,564	–
	LOC	2,769	630	–
	ORG	1,331	891	–
	# Sents.	31,870	7,418	–
Pashto / Pashto-Wikiann	PER	32	28	39
	LOC	37	45	45
	ORG	43	38	33
	# Sents.	100	100	100

Table 1: Type-wise statistics of the datasets for Urdu, Shahmukhi, Sindhi and Pashto.

(Ahmad et al., 2020). The quality of the dataset was further enhanced by using the BIO annotation scheme (Tehseen et al., 2023). The dataset contained some annotation inconsistencies. To ensure the validity of our NER results, we manually reviewed and corrected the annotations in one thousand sentences from the test set. While this review process was conducted to enhance the reliability of our evaluation.

Sindhi: Ali et al. (2020) released the first large annotated dataset for the Sindhi language called SiNER. We experimented with three coarse-grained entity types to make it compatible with the other datasets.

Pashto: Pashto lacks in fundamental language processing tools (Eberhard and Fennig, 2024). We used the Pashto dataset from Wikiann (Rahimi et al., 2019) that contains 100 sentences for train, test and validation sets. Since the dataset was automatically annotated and exhibited some annotation inconsistencies, we reviewed the test set manually to ensure valid NER results.

5 Experimental Setup

We conducted NER experiments designed to improve performance for low-resource languages, where supervised models often struggle due to limited annotated datasets. This research addresses three key questions; 1) How effective are data augmentation techniques to enhance NER for low-

resource languages? 2) Do cross-lingual data representations improve NER performance in multilingual settings? 3) How does few-shot learning compare to fully supervised models as an alternative to data augmentation? We hypothesize that cross-lingual representations, combined with multilingual datasets improve NER results for topologically related and culturally similar languages.

5.1 NER Models and Architectures

For our NER experiments, we employed two pre-trained multilingual masked language models: Glot500-base (Imani et al., 2023) and XLM-RoBERTa-large (Conneau et al., 2019).

- Glot500-base supports over 500 languages and is based on RoBERTa’s (Conneau et al., 2019) architecture. It uses transformer-based contextualized token embeddings and is particularly designed for low-resource languages like Urdu, Shahmukhi, Sindhi, and Pashto.
- XLM-RoBERTa-large is another transformer-based multilingual models that supports 100 languages, including Urdu, Sindhi, and Pashto. It is pre-trained on massive multilingual text corpora using masked language modeling (MLM) objectives.

To fine-tune these models for NER, we added a token classification layer on the top of the final transformer layer which receives the hidden states from the last layer of the model and computes the multi-class probability distribution over the entity classes for each token. This setup classifies tokens into person, location and organization categories.

We fine-tuned both models on mono- and multilingual datasets to investigate their performance for NER for low-resource setting by including 100, 200, 500 and 1000 train samples. Additionally, we experimented with the data augmentation techniques to further improve NER performance for low-resource languages.

5.2 Few-Shot Learning with Causal Models

While the primary focus of this paper is on data augmentation techniques to enhance NER performance in low-resource languages, we also explore few-shot learning as an alternative approach. Although various causal LLMs have recently been evaluated for the task of NER, they still struggle to compete with state-of-the-art supervised models (Naguib et al., 2024; Villena et al., 2024; Lu et al.,

2024). This raises a research question; how well do these models perform in low-resource languages?

We performed NER by using a few-shot learning approach by prompting LLaMA3-8B-Instruct¹ and Mistral-7B-Instruct-v0.3² which are instruction tuned. LLaMA3-8B is trained on 15 trillion tokens with a context length of 8K. Mistral-7B also has the same context length but its training size is not disclosed. We created a prompt, similar to GenerativeAug., describing details of the task by providing three examples for each language (Appendix E). The inputs and outputs were formatted as sequences of texts and NER labels. For erroneous outputs, the number of labels matching the number of tokens in the input was selected for evaluation. We evaluated the performance of both causal models on 1,000 sentences from each dataset.

6 Results and Discussion

We use micro F-scores to ensure a balanced evaluation of NER performance across all entity types. Table 2 presents Micro-F₁ score for low-resource NER experiments using monolingual and multilingual data settings. The training sets contain 100, 200, 500 and 1,000 samples for each dataset. In the multilingual settings, we combined training samples from all selected languages (Urdu, Shahmukhi, Sindhi, and Pashto). To maintain balanced representation, we ensured that each language contributed an equal number of samples in low-resource scenarios. The results are presented from fine-tuned Glot500-base and XLM-RoBERTa-large models. Similarly, Table 3 shows NER results for the entire datasets. The training samples in all augmented datasets were doubled in one iteration, and the NER results are presented after this iteration. Further analysis from multiple iterations is presented in the Appendix C.

Our data augmentation technique improved NER results for low-resource languages by reducing the generation of grammatically implausible and culturally offensive content. The augmentation technique helps maintain semantics and cultural appropriateness, that highly impacted the model performance. The model trained on the augmented datasets demonstrated higher generalizability due to less exposure to the contextually implausible in-

¹<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

²<https://huggingface.co/mistralai/Mistral-7B-v0.3>

Monolingual Settings		Glot500-base				XLM-RoBERTa-large			
Dataset	Augmentation	100	200	500	1000	100	200	500	1000
Urdu-Wikiann	Original dataset	70.93	77.23	83.51	80.24	72.77	71.21	84.21	87.21
	Generative _{Aug.}	77.13	79.29	84.24	86.81	79.66	83.85	85.01	85.50
	EDA-RR _{Aug.}	74.87	77.42	84.57	85.87	71.75	80.27	82.79	85.84
	Cluster-based _{Aug.}	76.62	81.00	83.78	85.31	75.24	80.79	84.30	85.97
Shahmukhi	Original dataset	59.62	65.27	71.92	75.44	53.67	59.44	70.65	75.58
	Generative _{Aug.}	53.83	62.45	69.85	74.89	58.81	58.64	66.96	74.68
	EDA-RR _{Aug.}	58.44	63.98	70.34	73.87	51.95	64.75	72.40	75.32
	Cluster-based _{Aug.}	60.78	68.03	73.17	77.11	59.61	65.89	74.19	77.40
SiNER	Original dataset	62.25	69.61	75.82	80.27	73.63	78.16	81.22	82.80
	Generative _{Aug.}	53.76	60.64	69.09	73.76	64.12	71.58	73.81	77.09
	EDA-RR _{Aug.}	64.69	72.29	73.50	72.65	75.40	75.22	80.01	83.00
	Cluster-based _{Aug.}	65.64	71.17	76.88	79.46	74.27	75.96	81.60	84.48
Pashto-Wikiann	Original dataset	32.86	—	—	—	44.24	—	—	—
	Generative _{Aug.}	45.66	—	—	—	45.26	—	—	—
	EDA-RR _{Aug.}	45.75	—	—	—	48.92	—	—	—
	Cluster-based _{Aug.}	48.54	—	—	—	50.00	—	—	—
Multilingual Settings		Glot500-base				XLM-RoBERTa-large			
Dataset	Augmentation	100	200	500	1000	100	200	500	1000
Urdu-Wikiann	Original dataset	73.12	74.82	84.45	84.90	63.32	78.25	82.33	80.97
	Generative _{Aug.}	78.92	79.67	84.16	85.77	77.78	80.93	82.07	85.43
	EDA-RR _{Aug.}	72.75	77.43	83.04	83.98	79.13	78.70	81.10	82.02
	Cluster-based _{Aug.}	76.83	82.25	84.35	85.34	78.60	79.49	81.14	83.21
Shahmukhi	Original dataset	65.33	69.21	75.84	79.38	56.87	67.85	72.27	76.66
	Generative _{Aug.}	64.32	68.45	74.46	77.45	65.69	69.23	74.68	78.21
	EDA-RR _{Aug.}	66.23	69.34	74.48	78.32	65.90	69.44	75.86	77.26
	Cluster-based _{Aug.}	68.88	73.47	77.51	80.01	67.83	71.38	73.79	78.90
SiNER	Original dataset	62.35	67.78	73.85	78.83	67.37	74.02	76.42	79.23
	Generative _{Aug.}	58.53	65.30	71.78	73.59	56.76	68.78	75.19	76.96
	EDA-RR _{Aug.}	64.72	69.71	74.30	77.84	69.79	74.48	76.06	79.77
	Cluster-based _{Aug.}	66.76	73.22	76.26	79.61	71.99	75.24	78.40	80.45
Pashto-Wikiann	Original dataset	62.26	67.68	73.68	78.58	67.01	73.79	76.22	78.96
	Generative _{Aug.}	58.51	65.19	71.62	73.43	65.68	68.66	74.98	76.73
	EDA-RR _{Aug.}	64.66	69.53	74.12	77.59	69.60	74.32	75.86	79.53
	Cluster-based _{Aug.}	66.63	73.12	76.05	79.35	71.78	74.98	78.17	80.21

Table 2: Micro-F₁ scores of fine-tuned multilingual Glot500-base and XLM-RoBERTa-large models for NER in low-resource settings. The results of the cluster-based augmentation are compared against the original training set, generative augmentation from LLaMa3 (Generative_{Aug.}) and EDA - Random Replacement (EDA-RR_{Aug.}).

formation. This confirms that grammatically and contextually inappropriate data can degrade the model performance by introducing noise and reducing its ability to generalize effectively. The following paragraphs present a comparison of data augmentation techniques for each dataset.

Urdu-Wikiann The Urdu-Wikiann dataset demonstrates inconsistent performance for different augmentation techniques, which is caused by three main reasons. First, Urdu is a resource-rich language compared to the other three regional languages and fine-tuning using cross-lingual data augmentation enhances its diversity, but does not significantly impact NER results due to the large size of the dataset. Second, causal LLMs, such as LLaMA3 have better support for Urdu compared to the other three languages as Urdu dataset shows improvements using Generative_{Aug.}

method. Third, the Urdu-Wikiann dataset is an automatically annotated dataset that may have some inconsistencies (Mayhew et al., 2023) which can limit the effectiveness of cross-lingual augmentation.

Shahmukhi The Shahmukhi dataset demonstrates consistent performance with cluster-based data augmentation as the proposed method generates plausible augmentations that leads to improved results. The fine-tuned XLM model produced a state-of-the-art F₁ score of 88.06 in multilingual settings using the BIO annotation scheme, which outperforms the previous best score of 75.55 (Tehseen et al., 2023).

However, Generative_{Aug.} decreased NER performance for Shahmukhi. The causal model produced various augmentations that violated entity types, resulting in incorrect labeling. The low scores in-

Monolingual Settings		Glott500-base			XLM-RoBERTa-large		
Dataset	Augmentation	Precision	Recall	F ₁ Score	Precision	Recall	F ₁ Score
Urdu-Wikiann	Original dataset	95.46	95.86	95.66	95.80	96.75	96.28
	EDA-RR _{Aug.}	94.89	96.38	95.63	96.08	96.08	96.08
	Cluster-based _{Aug.}	94.51	94.61	94.56	93.97	94.34	94.16
Shahmukhi	Original dataset	79.12	73.92	76.44	80.77	73.40	76.91
	EDA-RR _{Aug.}	85.58	76.96	81.04	85.55	79.76	82.55
	Cluster-based _{Aug.}	84.04	82.23	83.13	86.52	78.71	82.43
SiNER	Original dataset	90.50	85.69	88.03	88.78	89.12	88.95
	EDA-RR _{Aug.}	88.66	87.88	88.27	88.14	90.10	89.11
	Cluster-based _{Aug.}	87.49	88.82	88.15	87.50	89.68	88.58
Pashto-Wikiann	Original dataset	51.55	32.29	39.71	49.74	38.86	43.63
	EDA-RR _{Aug.}	46.45	47.77	47.10	48.19	53.45	50.68
	Cluster-based _{Aug.}	43.93	46.96	45.40	54.23	46.54	50.09
Multilingual Settings		Glott500-base			XLM-RoBERTa-large		
Dataset	Augmentation	Precision	Recall	F ₁ Score	Precision	Recall	F ₁ Score
Urdu-Wikiann	Original dataset	95.93	96.38	96.16	96.09	96.43	96.26
	EDA-RR _{Aug.}	96.10	96.75	96.42	95.02	95.58	95.30
	Cluster-based _{Aug.}	96.07	96.18	96.12	96.23	96.28	96.25
Shahmukhi	Original dataset	83.63	80.70	82.14	83.36	81.71	82.53
	EDA-RR _{Aug.}	87.98	83.43	85.64	88.51	83.62	86.00
	Cluster-based _{Aug.}	89.03	85.50	87.22	89.29	86.85	88.06
SiNER	Original dataset	87.99	84.82	86.37	87.12	86.35	86.73
	EDA-RR _{Aug.}	88.01	86.91	87.46	90.52	86.33	88.38
	Cluster-based _{Aug.}	89.19	86.69	87.92	89.33	87.80	88.56
Pashto-Wikiann	Original dataset	87.78	84.63	86.18	86.77	86.18	86.48
	EDA-RR _{Aug.}	87.51	86.72	87.12	90.15	85.94	88.00
	Cluster-based _{Aug.}	89.00	86.29	87.62	89.14	87.45	88.29

Table 3: Micro-F₁ scores of fine-tuned multilingual Glott500-base and XLM-RoBERTa-large models for complete datasets. The results of the cluster-based augmentation are compared against the original training sets and EDA - Random Replacement (EDA-RR_{Aug.}). Improved scores are highlighted in bold.

dicates that multilingual causal LLMs have limited support for low-resource languages. The cluster-based data augmentation technique outperformed other two augmentation methods in both monolingual and multilingual experiments.

SiNER For the Sindhi dataset, the cluster-based cross-lingual augmentation improved NER results in a multilingual setting by utilizing cross-lingual representations. This approach introduced linguistic variation and diversity that enhanced the models’ ability to generalize. For the entire dataset, EDA-RR_{Aug.} demonstrated improved results by adding cross-lingual entities that enriched the training set, making it a suitable augmentation technique for Sindhi in a monolingual training setup. However, Generative_{Aug.} had a negative impact on all low-resource training sets, highlighting limited capabilities of causal LLMs for low-resource languages. Sindhi’s use of Arabic script with additional unique letters, unlike Urdu, Shahmukhi, and Pashto, may negatively impact multilingual fine-tuning

Pashto-Wikiann The Pashto-Wikiann dataset demonstrates significant improvements with data augmentation techniques, especially in a multilin-

gual setup, except for Generative_{Aug.}. The best reported F₁ score for Pashto is 82.0 achieved from an HMM-based tagger (Momand et al., 2020). By using cluster-based augmentation, the multilingual fine-tuned Glott500 and XLM models produced F₁ scores of 87.62 and 88.29, respectively. However, these findings should be interpreted with caution due to the small size of the training and evaluation sets, which may limit the generalizability of the results.

Few-Shot Learning Table 4 presents NER results obtained from causal LLMs using few-shot learning. The performance of both LLaMA-3-8B and Mistral-7B on low-resource languages is not remarkable. LLaMa-3 performed better for Shahmukhi; however, its performance on Urdu, a relatively high-resource language, is quite low. The few-shot NER results indicate that causal LLMs are still far behind in NER for low-resource languages.

6.1 Limitations

Despite demonstrating significant advantages in the application of cross-lingual data augmentation, this study has a few limitations. The Shahmukhi, SiNER and MK-PUCIT datasets contain some an-

LLaMA-3-8B-Instruct			
Dataset	Precision	Recall	F ₁ Score
Urdu-Wikiann	20.13	24.26	22.00
Shahmukhi	74.63	72.06	73.32
SiNER	39.98	48.66	43.89
Pashto-Wikiann	48.46	56.76	52.28

Mistral-7B-Instruct-v0.3			
Dataset	Precision	Recall	F ₁ Score
Urdu-Wikiann	42.54	45.29	43.87
Shahmukhi	41.49	47.13	44.13
SiNER	27.02	38.40	31.72
Pashto-Wikiann	47.29	54.95	50.83

Table 4: Micro-F₁ scores by few-shot learning NER from LLaMA3-8B-Instruct and Mistral-7B-Instruct-v0.3. Both models have been evaluated for 1,000 sentences from each dataset except Pashto-Wikiann that has only 100 samples.

notation inconsistencies and errors that affect the overall performance of the models. Furthermore, the cluster-based data augmentation technique used entity clusters by employing an unsupervised clustering algorithm. The accuracy of the clustering process poses a limitation on the quality of the augmentation. Future work should focus on improving the annotation quality and consistency of such datasets.

7 Conclusion

This study explored various data augmentation techniques and their effect on the task of NER for low-resource languages. We used pre-trained LLMs on mono- and multilingual setups. Our findings highlight that cluster-based data augmentation improves NER performance for Shahmukhi, Sindhi and Pashto datasets by incorporating linguistically plausible text and cross-lingual diversity. Urdu-Wikiann, an automatically annotated dataset, does not take advantage of cross-lingual augmentations. Generative augmentation shows improved results on Urdu, while have a negative impact on the other three regional languages. Few-shot learning with causal models reveal their current limitations for low-resource languages when used for data augmentation and NER. Overall, the research emphasizes the potential of hybrid data augmentation techniques to enhance NER performance for low-resource languages.

References

Muhammad Tayyab Ahmad, Muhammad Kamran Malik, Khurram Shahzad, Faisal Aslam, Asif Iqbal, Zubair Nawaz, and Faisal Bukhari. 2020. Named

Entity Recognition and Classification for Punjabi Shahmukhi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TAL-LIP)*, 19(4):1–13.

Anil Ahmed, Degen Huang, Syed Yasser Arafat, and Imran Hameed. 2024. Enriching Urdu Ner with BERT Embedding, Data Augmentation, and Hybrid Encoder-CNN Architecture. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–38.

Wazir Ali, Junyu Lu, and Zenglin Xu. 2020. SiNER: A Large Dataset for Sindhi Named Entity Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2953–2961.

Ilaria Bartolini, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperli, and Andrea Vignali. 2022. COSINER: COntext SIMilarity data augmentation for Named Entity Recognition. In *International Conference on Similarity Search and Applications*, pages 11–24. Springer.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. *arXiv preprint arXiv:2303.00293*.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2022. Frustratingly Easy Label Projection for Cross-lingual Transfer. *arXiv preprint arXiv:2211.15613*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised Cross-lingual Representation Learning at Scale*. *CoRR*, abs/1911.02116.

Ryan Cotterell and Kevin Duh. 2024. Low-Resource Named Entity Recognition with Cross-Lingual, Character-level Neural Conditional Random Fields. *arXiv preprint arXiv:2404.09383*.

Zhuojun Ding, Wei Wei, Xiaoye Qu, and Danyang Chen. 2024. Improving Pseudo Labels with Global-Local Denoising Framework for Cross-lingual Named Entity Recognition. *arXiv preprint arXiv:2406.01213*.

Gary F. Simons Eberhard, David M. and Charles D. Fennig. 2024. *Ethnologue: Languages of the world*. *SIL International*, 27.

Toqeer Ehsan and Sarmad Hussain. 2021. Development and Evaluation of an Urdu Treebank (CLE-UTB) and a Statistical Parser. *Language Resources and Evaluation*, 55(2):287–326.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Utkarsh Tyagi, Dinesh Manocha, et al. 2024. CoDa: Constrained Generation based Data Augmentation for Low-Resource NLP. *arXiv preprint arXiv:2404.00415*.

- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. **XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation**. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Sarmad Hussain. 2008. Resources for Urdu language processing. In *Proceedings of the 6th workshop on Asian Language Resources*.
- Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. *arXiv preprint arXiv:2305.12182*.
- Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa, and Xuan Wang. 2012. N-gram and Gazetteer List based Named Entity Recognition for Urdu: A Scarce Resourced Language. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 95–104.
- Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019. Urdu Named Entity Recognition: Corpus Generation and Deep Learning Applications. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13.
- Hamza Khalid, Ghulam Murtaza, and Qaiser Abbas. 2023. Using Data Augmentation and Bidirectional Encoder Representations from Transformers for Improving Punjabi Named Entity Recognition. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–13.
- Wahab Khana, Ali Daudb, Jamal A Nasira, and Tehmina Amjada. 2016. Named Entity Dataset for Urdu Named Entity Recognition Task. *Language & Technology*, 51.
- Brayan Stiven Lancheros, Gloria Corpas Pastor, and Ruslan Mitkov. 2024. Data Augmentation and Transfer Learning for Cross-lingual Named Entity Recognition in the Biomedical Domain. *Language Resources and Evaluation*, pages 1–20.
- Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. Constrained Decoding for Cross-lingual Label Projection. *arXiv preprint arXiv:2402.03131*.
- Onkar Litake, Niraj Yagnik, and Shreyas Labhsetwar. 2024. IndiText Boost: Text Augmentation for Low Resource India Languages. *arXiv preprint arXiv:2401.13085*.
- Jian Liu, Yufeng Chen, and Jinan Xu. 2022. Low-Resource NER by Data Augmentation With Prompting. In *IJCAI*, pages 4252–4258.
- Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846.
- Wenzhong Liu and Xiaohui Cui. 2023. Improving Named Entity Recognition for Social Media with Data Augmentation. *Applied Sciences*, 13(9):5360.
- Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, and Others. 2024. **SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages**. *arXiv preprint arXiv: 2406.10118*.
- Qiuhaio Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2024. Large Language Models Struggle in Token-Level Clinical Named Entity Recognition. *arXiv preprint arXiv:2407.00731*.
- Muhammad Kamran Malik. 2017. Urdu Named Entity Recognition and Classification System using Artificial Neural Network. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–13.
- Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, et al. 2023. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. *arXiv preprint arXiv:2311.09122*.
- Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2024. MCLNER: Cross-Lingual Named Entity Recognition via Multi-View Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18789–18797.
- Rafiullah Momand, Shakirullah Waseeb, and Ahmad Masood Latif Rai. 2020. A Comparative Study of Dictionary-based and Machine Learning-based Named Entity Recognition in Pashto. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 96–101.
- Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlollah Mohaghegh, Mozdeh Rouhsedaghat, and Kai-Wei Chang. 2024. LLMs in Biomedicine: A Study on Clinical Named Entity Recognition. *arXiv preprint arXiv:2404.07376*.

- Marco Naguib, Xavier Tannier, and Aurélie Névél. 2024. Few Shot Clinical Entity Recognition in Three Languages: Masked Language Models Outperform LLM Prompting. *arXiv preprint arXiv:2402.12801*.
- Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. **Mas-**sively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.
- Caroline Sabty, Islam Omar, Fady Wasfalla, Mohamed Islam, and Slim Abdennadher. 2021. Data Augmentation Techniques on Arabic Data for Named Entity Recognition. *Procedia Computer Science*, 189:292–299.
- Sihan Song, Furoo Shen, and Jian Zhao. 2024. RoPDA: Robust Prompt-Based Data Augmentation for Low-Resource Named Entity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19017–19025.
- Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the Potential of Large Language Models (LLMs) for Low-resource Languages: A study on Named-Entity Recognition (NER) and Part-Of-Speech (POS) Tagging for Nepali Language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6974–6979.
- Amina Tehseen, Toqeer Ehsan, Hannan Bin Liaqat, Xi-angjie Kong, Amjad Ali, and Ala Al-Fuqaha. 2023. Shahmukhi Named Entity Recognition by using Contextualized Word Embeddings. *Expert Systems with Applications*, 229:120489.
- Arthur Elwing Torres, Edleno Silva de Moura, Altigran Soares da Silva, Mario A Nascimento, and Filipe Mesquita. 2024. An Experimental Study on Data Augmentation Techniques for Named Entity Recognition on Low-Resource Domains.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.
- Fabián Villena, Luis Miranda, and Claudio Aracena. 2024. llmNER:(Zero| Few)-Shot Named Entity Recognition, Exploiting the Power of Large Language Models. *arXiv preprint arXiv:2406.04528*.
- Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv preprint arXiv:1901.11196*.
- Junjie Ye, Xuanning Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhua Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. *arXiv preprint arXiv:2303.10420*.
- Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. LLM-DAA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. *arXiv preprint arXiv:2402.14568*.

A MK-PUCIT Dataset

The MK-PUCIT dataset was released with IO (Inside-Outside) annotation that has some annotation inconsistencies and errors. We converted it to the BIO (Begin-Inside-Outside) scheme automatically. For missing annotations, we extracted dictionaries with unique entities for each entity type from the training set and mapped the missing annotations throughout the dataset. After the mapping process, there was an overall increase of 19.9% in entity mentions for the train set and an increase of 13.8% for the test set. This highlights a significant number of missing annotations. Table 1 presents the updated statistics of the MK-PUCIT dataset.

We performed NER experiments by fine-tuning the Glot500 model and compared the results with different versions of the dataset in mono- and multilingual settings. Table 5 shows NER results for the MK-PUCIT. The original dataset, after conversion from IO to BIO scheme, performs with a micro F_1 score of 68.47. By performing the entity mapping for missing annotations, its performance was enhanced by 8.69 points, which is a significant improvement. Its performance remains in the same range in a multilingual setup. F_1 scores for the other three languages are lower compared to Urdu-Wikiann, therefore, we selected the Urdu-Wikiann dataset for multilingual NER experiments in this study.

Dataset	Monolingual NER		
	Precision	Recall	F_1 Score
MK-PUCIT _{Original}	74.27	63.51	68.47
MK-PUCIT _{Mapped}	81.14	73.56	77.16
MK-PUCIT _{Combined}	83.26	72.27	77.37
Shahmukhi	81.89	74.75	78.15
SiNER	81.44	79.76	80.59
Pashto-Wikiann	81.32	79.62	80.46

Table 5: NER results by fine-tuning Glot500-base on the MK-PUCIT dataset. The fine-tuned model has been trained on; 1) original dataset after conversion from IO scheme to BIO, 2) with entity mapping for missing annotations, 3) multilingual setup by combining datasets of four languages.

B Dataset Analysis

To investigate the capability of pre-trained models to generalize cross-lingual entity representations, we analyzed the ratio of named entities which are common in both training and test sets. The main objective of this analysis is to determine whether

the models are only memorizing seen examples or if they are improving generalization in multilingual training setup?. Table 6 shows type-wise presence of entity mentions from the test sets in the training sets. The analysis is given for both, mono- and multilingual datasets. All four datasets demonstrate a minor increase in seen examples from monolingual to multilingual datasets. The small increase in the ratio of seen entities is evident that the models enhance their learning by generalization and produce better NER results in multilingual setups.

C Augmentation Analysis

The cluster-based data augmentation has been performed to produce enhanced datasets with multiple iterations. The X_1 iteration shows a single pass of augmentation, X_2 iteration depicts two passes, and so on. In this section, we present an experimental analysis of the cluster-based augmentation with respect to different augmentation iterations.

Table 7 presents the NER results from the fine-tuned Glot500 model with mono- and multilingual low-resource data settings. The micro F_1 scores are compared against one and two iterations. The Urdu-Wikiann dataset demonstrates some improvements for X_2 in the monolingual setup using 100 and 200 samples. However, there is a decrease in the performance in multilingual experiments for all the other training sets. Similarly, Shahmukhi shows improved performance in monolingual setup and performance degradation in multilingual training. The SiNER and Pashto-Wikiann datasets also follow the similar trend for low-resource training splits.

Table 8 further shows NER results after fine-tuning on the entire datasets. In monolingual experiments, SiNER shows a subtle increase in scores with X_2 iterations in both mono- and multilingual setups. However, all the other datasets demonstrate performance degradation with the increase of iterations of data augmentations. Based on these NER results, we presented results and comparisons against one iteration of data augmentation in the results section of the paper.

Additionally, we compared the data augmentation method by selecting all correct sentences from the top five candidates with one and two iterations. Table 9 shows the comparison for low-resource settings. In the low-resource datasets, Urdu-Wikiann and Shahmukhi datasets perform better for only 100 samples for both mono- and mul-

Monolingual Datasets				
	Urdu-Wikiann	Shahmukhi	SiNER	Pashto-Wikiann
PER	254, 82.2%	482, 48.59%	555, 32.04%	3, 10.71%
LOC	102, 30.82%	140, 52.83%	115, 28.97%	5, 12.5%
ORG	234, 77.74%	66, 42.86%	57, 22.62%	8, 23.53%
Total	590, 62.69%	688, 48.75%	727, 30.53%	16, 15.68%
Multilingual Datasets				
	Urdu-Wikiann	Shahmukhi	SiNER	Pashto-Wikiann
PER	255, 82.52%	507, 51.11%	559, 32.27%	6 21.43%
LOC	106, 32.02%	151, 56.98%	116, 29.22%	11 27.5%
ORG	234, 77.74%	69, 44.81%	57, 22.62%	9 26.47%
Total	595, 63.23%	727, 51.52%	732, 30.74%	26, 25.49%

Table 6: Analysis of presence of named entities of test sets in monolingual and multilingual training sets.

Monolingual Setup Datasets	100		200		500		1000	
	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
Urdu-Wikiann	76.62	76.48	81.00	82.32	83.78	83.13	85.31	84.73
Shahmukhi	60.78	62.24	68.03	68.79	73.17	73.03	77.11	78.15
SiNER	65.64	65.66	71.17	70.84	76.90	78.67	79.46	79.77
Pashto-Wikiann	48.54	48.51	—	—	—	—	—	—
Multilingual Setup Datasets	100		200		500		1000	
	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
Urdu-Wikiann	76.83	70.81	82.25	74.75	84.35	81.79	85.34	84.27
Shahmukhi (1k)	68.88	65.55	73.47	70.59	77.51	75.56	80.01	79.52
Sindhi	66.76	68.77	73.22	71.72	76.26	75.89	79.61	79.01
Pashto	66.63	68.67	73.12	71.60	76.05	75.68	79.35	78.81

Table 7: Micro-F₁ scores by fine-tuning Glot500-base on low-resource multilingual datasets by using data augmentation with one (X₁) and two (X₂) iterations.

lingual experiments. The other data splits start performance degradation. SiNER demonstrates some improvements for 1,000 sentences in monolingual experiment and for 100 train samples for multilingual setup. The performance degradation is observed for all the other training sets. Pashto-Wikiann is a smaller dataset that contains only 100 sentences and it shows improvements by learning cross-lingual representations in multilingual setup.

We further compared the results by selecting all correct sentences for entire datasets as shown in Table 10. The F₁ score for Urdu-Wikiann remains in the same range for monolingual training but decreases significantly in the multilingual training setup. However, F₁ scores for Shahmukhi and Sindhi are quite low compared to X₁ and X₂ iterations. Pashto-Wikiann shows the similar behaviour.

The Shahmukhi and SiNER datasets were further analyzed for one, two and three augmentation iterations for low-resource monolingual settings as shown in Table 11. Shahmukhi shows improvements by training with three iterations. However, in the multilingual setup, it shows performance degradation when adding more augmented sentences (Table 10). On the other hand, SiNER performs with mixed results but it also demonstrates decreased

performance in multilingual training setup with increased data augmentation iterations. Based on these analysis, augmentation with one iteration produces optimal performance for Urdu-Wikiann, Shahmukhi, SiNER and Pashto-Wikiann datasets. Therefore, in the main paper, we presented the results achieved by using one iteration of the cluster- and EDA-based data augmentation methods for all the selected datasets.

Table 12 presents the F₁ scores for Shahmukhi and SiNER Few-Shot experiments with five different randomly selected training sets to analyze the variation in scores across datasets. Pashto-Wikiann is a small dataset with only 100 instances, and our data augmentation technique does not perform well on Urdu-Wikiann; therefore, we experimented only on the Shahmukhi and SiNER datasets. Shahmukhi exhibits a consistent trend across all Few-Shot settings, with a mean score closely aligning with the actual scores. However, SiNER, on the other hand, demonstrates higher variance for the smaller number of examples.

D Hyperparameters

In the fine-tuning process, the learning rate of 2e-5 was used along with the AdamW optimizer. The batch size was set to 8, which helped to maintain

Monolingual Setup		X_1			X_2		
Datasets	Precision	Recall	F_1	Precision	Recall	F_1	
Urdu-Wikiann	94.51	94.61	94.56	93.75	94.14	93.94	
Shahmukhi	84.04	82.23	83.13	82.33	82.20	82.27	
SiNER	87.49	88.82	88.15	88.91	88.01	88.48	
Pashto-Wikiann	52.08	45.45	48.54	58.46	41.45	48.51	
Multilingual Setup		X_1			X_2		
Datasets	Precision	Recall	F_1	Precision	Recall	F_1	
Urdu-Wikiann	96.07	96.18	96.12	94.76	95.70	95.23	
Shahmukhi	89.03	85.50	87.22	86.83	86.08	86.45	
SiNER	89.19	86.69	87.92	88.16	88.01	88.08	
Pashto-Wikiann	89.00	86.29	87.62	87.85	87.54	87.69	

Table 8: Micro-F₁ scores by fine-tuning Glot500-base on multilingual setting for the entire datasets by using data augmentation with one (X₁) and two (X₂) iterations.

Train Size	Iteration	Urdu-Wikiann	Shahmukhi	SiNER	Pashto-Wikiann
Monolingual Setup					
100	X_1	76.62	60.78	65.64	48.54
	X_2	76.48	62.25	65.66	48.51
	All correct	72.31	64.39	65.27	49.78
200	X_1	81.00	68.03	71.17	—
	X_2	82.32	68.79	70.84	—
	All correct	81.18	67.85	71.46	—
500	X_1	83.78	73.17	76.88	—
	X_2	83.13	73.03	78.67	—
	All correct	84.57	73.87	76.00	—
1000	X_1	85.31	77.11	79.46	—
	X_2	84.73	78.15	79.77	—
	All correct	81.76	77.16	80.98	—
Multilingual Setup					
100	X_1	76.83	66.88	66.76	66.63
	X_2	70.81	65.55	68.77	68.67
	All correct	79.10	67.85	64.89	64.84
200	X_1	82.25	73.47	73.22	73.12
	X_2	74.75	70.59	71.72	71.60
	All correct	79.58	71.55	72.84	72.70
500	X_1	84.35	77.51	76.26	76.05
	X_2	81.79	75.56	75.89	75.68
	All correct	81.49	76.00	77.03	76.86
1000	X_1	85.34	80.01	79.61	79.35
	X_2	84.27	79.52	79.01	78.81
	All correct	85.03	79.13	79.18	79.53

Table 9: Micro-F₁ scores by fine-tuning Glot500-base on monolingual and multilingual low-resource datasets by using data augmentation with one (X₁) and two (X₂) iterations and all correct from top five augmentations.

Monolingual Setup			
Dataset	X_1	X_2	All correct
Urdu-Wikiann	94.56	93.94	94.58
Shahmukhi	83.13	82.27	81.79
SiNER	88.15	88.48	86.84
Pashto-Wikiann	48.54	48.51	49.78
Multilingual Setup			
Urdu-Wikiann	96.12	95.23	91.82
Shahmukhi	87.22	86.45	83.42
SiNER	87.92	88.08	84.82
Pashto-Wikiann	87.62	87.69	84.52

Table 10: Micro-F₁ scores by fine-tuning Glot500-base on monolingual low-resource datasets by using data augmentation with one (X₁) and two (X₂) iterations and all correct from top five augmentations.

Train Size	Iteration	Shahmukhi	SiNER
100	X_1	60.78	65.64
	X_2	62.25	65.66
	X_3	61.85	65.03
200	X_1	68.03	71.17
	X_2	68.79	70.84
	X_3	70.35	70.38
500	X_1	73.17	76.88
	X_2	73.03	78.67
	X_3	73.89	75.98
1000	X_1	77.11	79.46
	X_2	78.15	79.77
	X_3	77.68	80.53

Table 11: Micro-F₁ scores by fine-tuning Glot500-base on monolingual low-resource datasets by using data augmentation with one (X₁), two (X₂) and three (X₃) iterations.

memory and training efficiency. The models were fine-tuned by setting various number of epochs for low-resource datasets depending on the training samples. Early stopping was further implemented based on the micro F₁ score on the validation set. The maximum sequence length was set to 100 tokens. These hyperparameters ensured optimal performance of the models.

E Few-Shot NER - Prompt

You are an expert in identifying named entities for language. The INPUT contains text followed by an OUTPUT sequence of BIO labels. Perform named entity recognition and return the labels. Three examples are provided for your reference:

EXAMPLE 1:

INPUT: Foreign advisor Sartaj Aziz will visit Afghanistan today.

OUTPUT: O O B-PER I-PER O O B-LOC O.

Shahmukhi				
RUNs	100	200	500	1000
Run 1	62.12	66.04	72.05	76.69
Run 2	60.77	67.00	71.47	75.45
Run 3	60.49	65.86	72.62	77.15
Run 4	61.81	64.85	73.36	77.23
Run 5	62.58	67.33	72.05	74.98
Mean	61.55	66.22	72.31	76.30
Variance	0.7963	0.9698	0.5098	1.0511
Standard Deviation	0.8924	0.9848	0.7140	1.0252
SiNER				
RUNs	100	200	500	1000
Run 1	64.81	70.40	75.83	77.78
Run 2	60.54	66.62	75.21	79.32
Run 3	63.40	65.89	74.94	76.81
Run 4	63.67	69.44	74.57	78.74
Run 5	64.65	69.86	73.55	79.45
Mean	63.41	68.44	74.82	78.42
Variance	2.9505	4.1682	0.7155	1.2437
Standard Deviation	1.7177	2.0416	0.8458	1.1152

Table 12: Mean, variance, and standard deviation by fine-tuning Glot500-base for Shamukhi and SiNER Few-Shot settings on five randomly selected train sets.