computers

MDPI

Article

Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu

Fida Ullah, Alexander Gelbukh *, Muhammad Tayyab Zamir, Edgardo Manuel Felipe Riverón and Grigori Sidorov

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Av. Juan de Dios Batiz, s/n, Mexico City 07320, Mexico; fidaullahmohmand@gmail.com or fullah-2022@cic.ipn.mx (F.U.)
* Correspondence: gelbukh@cic.ipn.mx

Abstract: Identifying and categorizing proper nouns in text, known as named entity recognition (NER), is crucial for various natural language processing tasks. However, developing effective NER techniques for low-resource languages like Urdu poses challenges due to limited training data, particularly in the nastaliq script. To address this, our study introduces a novel data augmentation method, "contextual word embeddings augmentation" (CWEA), for Urdu, aiming to enrich existing datasets. The extended dataset, comprising 160,132 tokens and 114,912 labeled entities, significantly enhances the coverage of named entities compared to previous datasets. We evaluated several transformer models on this augmented dataset, including BERT-multilingual, RoBERTa-Urdu-small, BERT-base-cased, and BERT-large-cased. Notably, the BERT-multilingual model outperformed others, achieving the highest macro F1 score of 0.982%. This surpassed the macro f1 scores of the RoBERTa-Urdu-small (0.884%), BERT-large-cased (0.916%), and BERT-base-cased (0.908%) models. Additionally, our neural network model achieved a micro F1 score of 96%, while the RNN model achieved 97% and the BiLSTM model achieved a macro F1 score of 96% on augmented data. Our findings underscore the efficacy of data augmentation techniques in enhancing NER performance for low-resource languages like Urdu.

Keywords: named entity recognition; Urdu; BERT; data augmentation; low-resource languages

check for updates

Citation: Ullah, F.; Gelbukh, A.; Zamir, M.T.; Riverón, E.M.F.; Sidorov, G. Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu. *Computers* **2024**, *13*, 258. https://doi.org/10.3390/computers13100258

Academic Editors: Phivos Mylonas, Yorghos Voutos, Cleo Sgouropoulou, Christos Troussas and Akrivi Krouska

Received: 31 July 2024
Revised: 22 September 2024
Accepted: 30 September 2024
Published: 10 October 2024

1. Introduction

Entity recognition (NER), also known as entity identification, entity chunking, or entity extraction, is a fundamental natural language processing (NLP) task. It involves the sequential labeling and classification of proper nouns into predefined categories, such as persons, locations, organizations, expressions of time, quantities, and monetary values. This task is considered an essential preliminary step in various NLP applications, including question answering, information retrieval, machine translation, and sentiment analysis. As such, NER plays a crucial role in the management and extraction of meaningful information from text [1].

Named entity recognition approaches have been explored and actively implemented for several years [2,3]. The first NER challenge was introduced during the 6th Message Understanding Conference in 1996 [4,5]. NER frameworks for English and other developed languages have been extensively established since then. However, due to the diversity and structural uniqueness of the Urdu language, Urdu NER development remains an ongoing process. In morphologically rich languages, such as Urdu, the number of words derived from a single root word is often substantial. Furthermore, compared to NER for other languages, research in this field for Urdu is significantly smaller, and the available resources are limited [6,7]. UNER researchers have primarily employed three methodologies: the rule-based approach, which is founded on constructed grammar rules; [8] the learning-based approach, which requires tagged samples and various features to perform better; [9] and a hybrid approach that combines learning-based and rule-based methods [10]. To achieve favorable outcomes, these methods leverage language-specific expertise in rule-based approaches and extensive feature engineering in learning-based strategies.

NER systems have garnered substantial attention from researchers since their introduction at the Message Understanding Conferences [5]. In the ensuing years, a wide array of NER techniques and systems have been developed [11,12], primarily catering to Western languages, particularly English, and achieving commendable accuracy [13]. Concurrently, various frameworks have been devised for other languages, including Arabic, Persian, and South Asian languages like Hindi and Bengali [14,15]. However, despite these advancements, the development of NER systems specifically tailored to the Urdu language remains in its nascent stages [16].

This research primarily aims to assess the current state of the art in NER and propose bidirectional encoder representations from transformers (BERT) multilingual. Previous efforts in Urdu named entity recognition (UNER) have heavily relied on manual feature engineering and data preprocessing, as evidenced by Zoya et al. [17].

However, recent experiments indicate that deep learning methods are gaining prominence in NLP tasks, including NER, and are often more effective than feature-based approaches, as highlighted by Çoban et al. [18]. Despite the popularity of deep neural networks, implementing them in highly morphological languages like Urdu poses certain challenges.

To address these challenges, our study was conducted to analyze the performance of various deep learning architectures against established benchmarks. The results we present surpass those of the current state-of-the-art UNER systems. The main contributions of this research are as follows.

- The existing U-NER corpus, the largest annotated dataset for Urdu named entity recognition (NER), was expanded through "contextual word embeddings augmentation" (CWEA). The original dataset of 50,692 tokens with 16,300 named entities was increased to 160,132 tokens, resulting in the creation of the UNER-II corpus.
- In the extended dataset, we annotated named entities (NEs) into specific classes: person (PER), location (LOC), and organization (ORG).
- Four distinct types of transformer models (multilingual bidirectional encoder representation from transformers (BERT), RoBERTa-Urdu-small, BERT-base-cased, and BERT-large-cased) were utilized on two datasets to classify named entities.
- Our approach, evaluated using precision, recall, and F score, significantly outperforms existing state-of-the-art DL-based NER methods for the Urdu language.
- The paper is organized as follows: The Section 2 outlines the difficulties specific to Urdu NER. The Section 3 reviews the relevant literature. The Section 4 details the research approach. The Section 5 presents and discusses the findings. Finally, the Section 7 provides concluding remarks and suggestions for future research.

2. Urdu NER Challenges

The significant presence of ambiguities concerning named entities (NEs) and the inherent linguistic complexities associated with the Urdu language collectively make the task of NER in Urdu a particularly challenging endeavor. The development of a strong and effective NER system tailored to Urdu is compounded by a series of constraints and limitations, which we describe below.

2.1. Lack of Capitalization

Various languages employ various writing systems, and some, like English, utilize capitalization as a distinctive feature. In contrast, the Urdu language lacks such an indicator, since it does not incorporate capitalization conventions. For example, (وی او اي), transcribed (VOA) in Urdu cannot be recognized as an acronym [8].

2.2. Segmentation

Segmentation, commonly known as tokenization, plays a crucial role in numerous NLP tasks such as parts of speech and named entity tagging. However, the process of segmentation is notably more intricate in Urdu compared to Western languages like English. In English, tokens (words) can be easily identified using spaces and distinct characters, whereas Urdu presents a more complex scenario for segmentation [19].

2.3. Cursive Context Sensitivity

Another challenge encountered in Urdu named entity recognition (UNER) pertains to the presence of multiple tag ambiguities and the intricate context sensitivity arising from its cursive nature. Distinguishing between common nouns and proper nouns in the Urdu language proves to be a complex task. Moreover, Urdu's cursive script involves the amalgamation of individual characters to form complete words, rendering it context-dependent. This contextual dependence leads to variations in character structures, contingent upon the preceding or succeeding characters. For instance, the token "جلال" (Jalal) refers to a person's name, whereas " جلال آباد" (Jalalabad is the capital city) pertains to a location name, resulting in potential confusion when endeavoring to accurately identify the intended entity.

2.4. Agglutination

The presence of agglutination is a distinctive attribute in the Urdu language, where words are composed of multiple components, including prefixes, roots, and suffixes, resulting in a complex morphology. For instance, consider the word "نازمند" (needy), which is deconstructed into "ناز" (need), denoting the NE category of a person and "مند" (having the quality of) representing the NE type of "other". In contrast, English typically employs single-word formations, simplifying the recognition of named entities [8].

2.5. Diacritics

Diacritics in Urdu are not obligatory, but serve a purpose in differentiating named entities. Urdu, following the Arabic script, employs diacritics to clarify and represent specific abbreviated vowel sounds. The Urdu language encompasses a range of diacritical marks, including "zaber", "zair", "paish", "khari zaber", "juzm", and several others.

2.6. Variation in Spelling

In news articles, it is not uncommon to encounter instances where various authors or reporters may spell names differently, even when referring to native Urdu names. For instance, consider the case of "مسعود" and "مسود", both of which denote the same person, "Masood". "مسعود" (Masood) represents the Arabic-style rendering of the name, including an additional vowel, whereas "مسود" (Masood) adheres to the native Urdu form of writing the name.

2.7. Loanwords from Other Languages

Urdu incorporates numerous loanwords from different languages. For instance, words like "استری", meaning "world", have been borrowed from Arabic. Additionally, terms such as "tOlyA" for "towel" and "almArI" denoting "cupboard" have their origins in Portuguese. Furthermore, words like "jindRI" for "life" and "acAr", referring to "pickle", have been borrowed from the Persian language [20]. Given the current challenges we face, creating NER systems for Urdu is a complex task. In our effort to make a valuable contribution to the development of Urdu NER, we have employed a straightforward yet innovative data augmentation technique in combination with a state-of-the-art transformer model. This approach has been instrumental in building an effective Urdu NER system.

3. Related Work

The research on named entity recognition for Western languages has a long-standing history dating back to the early 1990 [4]. Since then, numerous studies have been conducted to address the NER problem, employing a range of techniques, from rule-based approaches to purely supervised methods. Additionally, researchers have explored the use of hybrid approaches for NER in various texts [11–21]. However, it has been observed that NER techniques developed for specific domains may not be equally effective across other domains [22]. Similarly, the techniques developed for NER in one language may not be as efficient when applied to other languages. For instance, an NER system designed for Spanish may not be readily usable for Turkish or Chinese. The majority of NER research has predominantly focused on Western languages. However, some work has been done on UNER, mainly based on rule-based, machine learning, and hybrid-based approaches.

Riaz et al. [8] pioneered the development of a rule-based algorithm for named entity recognition in Urdu, with a core focus on six entity categories. Their system was evaluated using a benchmark dataset created by Becker–Riaz, which yielded impressive performance metrics, including precision, recall, and F-measure values of 91.5%, 90.7%, and 91.1%, respectively. Additionally, Singh et al. [15] developed a rule-based approach for Urdu named entity recognition, concentrating on 12 distinct entity categories. This system was evaluated on the IJCNLP-2008 dataset and achieved an accuracy score of 74%.

Jahangir et al. [23] developed n-gram-based models, specifically Unigram and Bigram, and applied smoothing algorithms to recognize five distinct named entity classes. The evaluation of their system yielded the following performance metrics: for the Unigram model, the precision, recall, and F-measure values were 65.21%, 88.63%, and 75.14%, respectively; while for the Bigram model, the corresponding values were 66.20%, 88.18%, and 75.83%.

Mukund et al. [16] developed a conditional random field-based model for Urdu named entity recognition, which considered three entity classes. Their approach achieved an F measure of 68.90%. Additionally, the authors [10] proposed a hybrid system for Urdu named entity recognition, primarily incorporating CRF, hidden Markov model, and manual heuristics. This hybrid system yielded the highest reported precision, recall, and F-measure values of 56.21%, 37.15%, and 44.73%, respectively.

Malik et al. [9] pioneered the use of neural networks for Urdu named entity recognition. Their system was evaluated on the KPU-NE corpus and achieved remarkable performance metrics, with the highest reported precision, recall, and F-measure values of 81.05%, 87.54%, and 84.17%, respectively. In another study, Kanwal et al. [24] introduced the MK-PUCIT dataset and explored two machine learning techniques, namely, artificial neural networks and recurrent neural networks, for the task of Urdu named entity recognition. The ANN-based approach achieved the highest reported precision, recall, and F-measure values of 76.5%, 73.2%, and 73.8%, respectively. The RNN-based model yielded the best performance metrics of 76.3% precision, 78.9% recall, and 77.5% F measure.

In a study by Kumar Saha et al. [10], a hybrid named entity recognition system is developed for five languages: Hindi, Bengali, Telugu, Oriya, and Urdu. Their approach combines linguistic rules with a maximum entropy model and utilizes gazetteer lists to enhance performance. The reported F-measure values for the five languages are 65.13%, 65.96%, 44.65%, 18.74%, and 35.47%, respectively. Additionally, Gali et al. [25] introduce a hybrid system that employs tailor-made rules in conjunction with conditional random fields. However, due to insufficient training data, their system achieves an F measure of only 43.46%.

Khan et al. [26] employed a deep recurrent neural network for Urdu NER, incorporating context windows and parts-of-speech features. Their approach was meticulously validated across diverse datasets, both for independent and dependent feature extractions. The results demonstrated the superiority of their proposed method over previous approaches, including conditional random fields and artificial neural networks. F-measure values of 81.1%, 79.94%, and 63.21% were achieved on three benchmark datasets, further reinforcing the efficacy of their approach.

Similarly, the UNER-1 dataset was also utilized by Wahab et al. [5], the main purpose of that study being to recognize Urdu named entities employing the conditional random field (CRF) methodology. Their experimental outcomes revealed that their novel approach surpassed the baseline technique for the dataset, leading to a noteworthy enhancement in F1 scores ranging from 1.5% to 3%. Furthermore, the results provided evidence that the improved dataset proved to be highly beneficial for facilitating learning and prediction within the context of a supervised learning framework.

Ullah et al. [27] proposed a novel Bi-LSTM and CRF-based approach for Urdu named entity recognition, which incorporated enhancements to the attention layer. Their innovative model achieved a remarkable F1 score of 92%, outperforming previously reported results and demonstrating the effectiveness of their novel architectural modifications. Haq et al. [19] introduced a sophisticated NER system for the Urdu language, incorporating deep learning techniques to address intricate feature extraction challenges. Their study also introduced a manually annotated tweets dataset in Urdu, encompassing five named entity classes. The deep learning approaches demonstrated substantial advancements over existing state-of-the-art NER techniques, culminating in a 6.26% enhancement in the F1 score.

4. Methodology and Material

Our proposed approach begins with the acquisition of data, followed by the preprocessing of raw data. Subsequently, we employ a data augmentation technique to expand the existing corpus. The dataset is then divided into training and testing sets, and we apply various transformer models, including BERT-multilingual, RoBERTa-Urdu-small, BERT-base-cased, and BERT-large-cased, to perform named entity recognition (NER) tasks on Urdu data. For further explanation, see Figure 1.

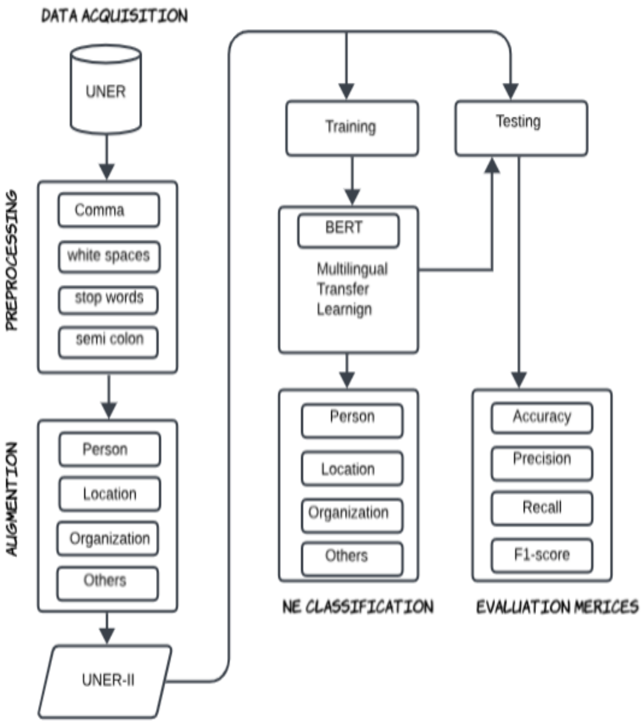

Figure 1. Main steps of the proposed methodology.

4.1. Corpus

In this study, we leveraged the publicly available UNER dataset (https://github.com/javaidiqbal11/Named-Entity-Recognition-for-Urdu/blob/master/ner%20(1).txt) (accessed on 5 November 2023), which is available for the Urdu named entity recognition task. The UNER dataset, a subset derived from an Urdu language corpus, consists of 50,692 texts. These texts were meticulously extracted from the preceding UNER dataset, presenting a rich source for NER model training and evaluation. The NER labels in this dataset encompass four distinct entities—person, location, organization, and others—as outlined in Table 1. For the creation of our annotated dataset, we drew upon the recent research efforts of Javed Iqbal et al. This original dataset encapsulates a total of 50,692 tokens, wherein 2380 instances pertain to person entities, 1547 to locations, and 1545 to organizations. Additionally, 45,220 instances were classified as other named entities (NEs). The diversity and size of this dataset provide a robust foundation for addressing the complexities of Urdu NER, enabling the exploration and evaluation of models designed for entity recognition in the Urdu language. This annotated dataset, sourced from reputable research, serves as a valuable resource for advancing the field of NER in the context of Urdu language processing.

Table 1. Urdu named entity recognition dataset.

Label	PER	LOC	ORG	O
Total Entities	2380	1547	1545	45,220

4.2. Data Processing and Augmentation

The preprocessing of text stands as a pivotal stage in the NLP pipeline, holding significance in the proficient analysis and representation of Urdu text data. The primary objective of text preprocessing is to convert raw text data into a standardized numerical format, making it apt for subsequent analysis and computational algorithm processing. This multifaceted process encompasses tasks such as eliminating extraneous information from the text, encoding the text uniformly, and converting the textual content into a numerical representation.

The dataset was cleaned using various techniques, such as removing stop words, commas, and semicolons. Additionally, the elimination of white spaces was carried out to enhance data cleanliness, and these refined data were subsequently utilized for data augmentation. The entire preprocessing operation was executed using Python libraries, with a particular emphasis on leveraging the Natural Language Toolkit (NLTK), a widely utilized tool by NLP engineers and researchers globally. Moving on to the subsequent phase, we employed a data augmentation technique centered on "contextual word embedding augmentation" (CWEA). A detailed methodology of the data augmentation process is provided in Supplementary Materials Table S4. In this step, a publicly available dataset containing named entities (NEs) related to persons, locations, organizations, and others was utilized. Extracted data from this source were incorporated into a pool, from which NEs were randomly selected to extend both sentences and datasets.

It was ensured that the repetition rate of NEs remained low by selecting words only if their occurrence in the text was below a specified threshold of 0.1. This strategic selection practice aimed to maintain diversity within our extended dataset, a crucial factor contributing to the subsequent development of an enhanced named entity recognition (NER) system for Urdu. An illustrative example of a generated new sentence is presented in Figure 2. We used a data augmentation method to increase the amount of text from 50,000 to 154000. The augmentation was implemented for three types of named entities: person, location, and organization. By means of inner augmentation, 47,000 additional annotated texts for a person, 14,000 annotated texts for location, and 30,945 texts for the organization were created. Our augmented dataset now comprises 160,132 tokens with 114,912 annotated NEs. The focus during the creation of our UNER-II corpus, which is the augmented version of the original UNER dataset, centered on four primary NE groups: person, location, organization, and others. This selection was based on the widespread implementation areas of these NE groups, considering that most existing datasets primarily encompass these types [24]. A comparative analysis between the extended dataset and the existing UNER corpus data is detailed in Table 2.

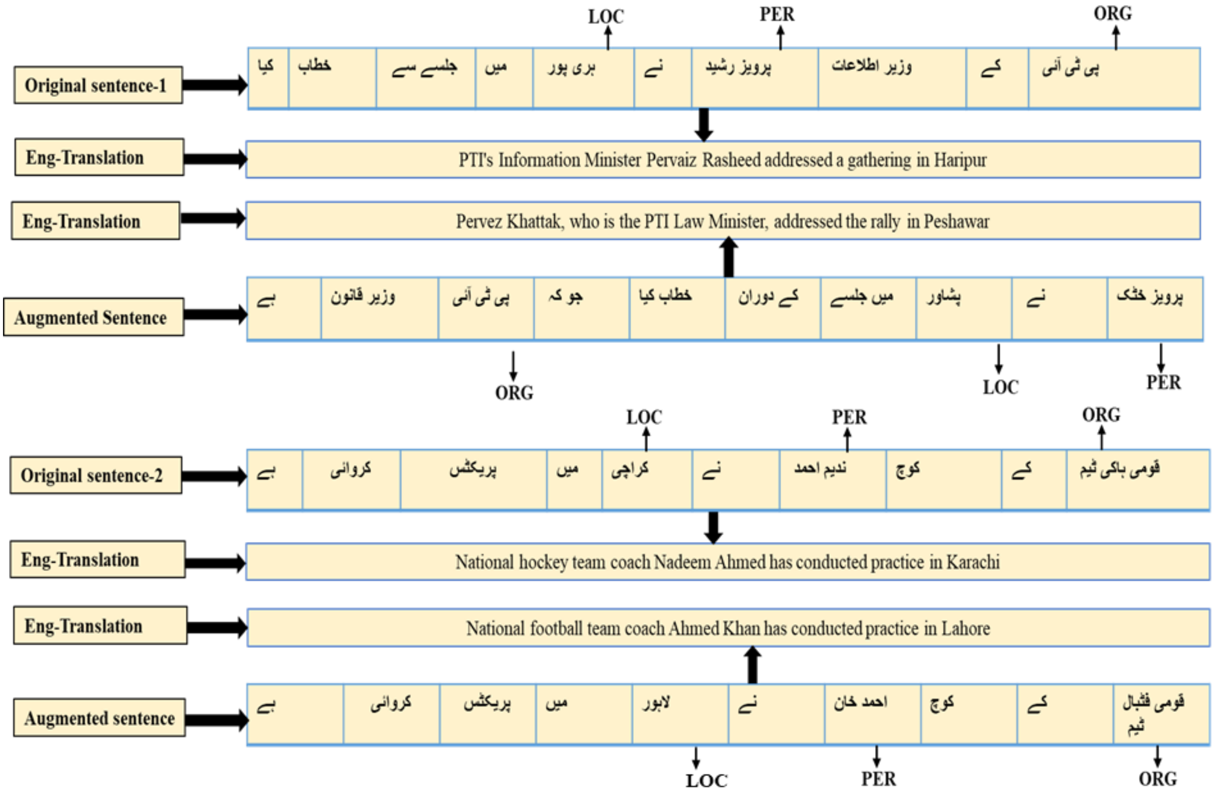
LOC
PER
ORG
Original sentence-1
پی ٹی آئی کے وزیر اطلاعات پرویز رشید نے ہری پور میں جلسے سے خطاب کیا
Eng-Translation
PTI's Information Minister Pervaiz Rasheed addressed a gathering in Haripur
Eng-Translation
Pervez Khattak, who is the PTI Law Minister, addressed the rally in Peshawar
Augmented Sentence
پرویز خٹک نے پشاور میں جلسے کے دوران خطاب کیا جو کہ پی ٹی آئی وزیر قانون ہے
ORG
LOC
PER
LOC
PER
ORG
Original sentence-2
قومی ہاکی ٹیم کے کوچ ندیم احمد نے کراچی میں پریکٹس کروائی ہے
Eng-Translation
National hockey team coach Nadeem Ahmed has conducted practice in Karachi
Eng-Translation
National football team coach Ahmed Khan has conducted practice in Lahore
Augmented sentence
قومی فٹبال ٹیم کے کوچ احمد خان نے لاہور میں پریکٹس کروائی ہے
LOC
PER
ORG

Figure 2. Example of original and augmented sentences with English translation.

Table 2. Comparison of UNER-II with existing UNER dataset.

Characteristics	Existing Dataset	Extension	UNER-II
Total tokens	50,692	109,440	160,132
Person	2380	47,600	49,980
Location	1547	30,940	32,487
Organization	1545	30,900	32,445
Other	45,220	--	45,220

4.3. Pre-Trained BERT Model

Transfer learning methods [28], such as pre-trained deep neural networks have demonstrated significant advancements in various natural language processing (NLP) tasks, including text classification [29], machine translation [30], and text summarization. Transformer models like BERT [28], ALBERT [31], and RoBERTa [32] have particularly excelled in NLP. These models undergo pre-training on extensive datasets, often sourced from Wikipedia articles, through unsupervised learning. They can then be fine-tuned for specific tasks like named entity recognition (NER). Unlike recurrent neural networks (RNNs), transformers do not process input sequentially. They rely on attention mechanisms, making them suitable for NLP tasks due to their intricate architecture, involving embedding layers, self-attention layers, and feed-forward layers. BERT [33] for instance, was trained with objectives involving masked language modeling, where it predicts masked words in a randomly masked input sequence, and next-word prediction, where it determines if two sentences follow each other. BERT achieved remarkable results across various language understanding tasks. For Urdu named entity recognition, our proposed BERT model architecture is illustrated in Figure 3.

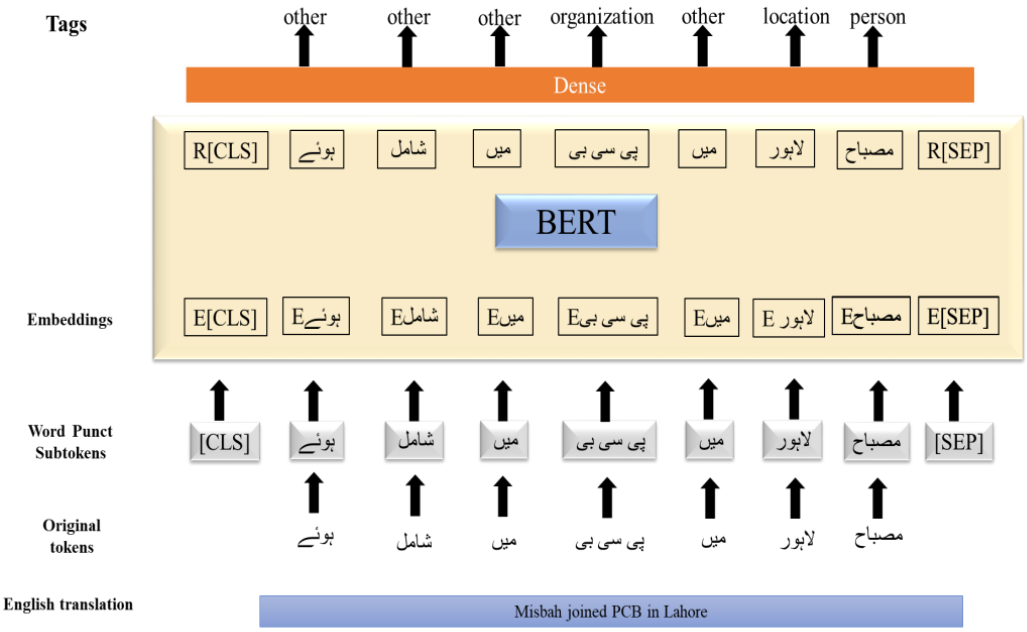

Figure 3. The architecture of the proposed BERT model.

4.4. Experimental Setup

We opted for a pre-trained language model, specifically a multilingual BERT base model, for two primary reasons. Firstly, it comes pre-trained on Urdu language data, which is advantageous for transfer learning tasks, particularly in the context of NER classification. The Urdu language shares the same writing script as the Arabic script, facilitating knowledge transfer. Secondly, this model has been previously employed by researchers in Urdu NLP tasks, as evidenced by Daud et al. [5]. Our results align with these findings affirming the model's effectiveness for tasks in low-resource languages. Moreover, alternatives within the BERT family, such as RoBERTa-Urdu-small, BERT-base-cased, and BERT-large-cased [34], exist. We selected a model pre-trained in 104 languages, including Urdu, using self-supervised methods. Notably, the model boasts 12 attention layers, totaling 110 million parameters. The base model features a hidden layer size of 768 with 12 self-attention heads. WordPiece embedding, with a vocabulary of 30,000 tokens, was employed. During fine-tuning, we utilized a cross-entropy loss employing the Adam optimizer, with a learning rate of 2×10^{-5}. The model underwent training for 5 epochs, utilizing a batch size of 32 and a sequence length of 256. To address overfitting, a dropout value of 0.2 was implemented. All experiments were conducted on Google Colab, leveraging a Tesla K80 12 GB GPU and 32 GB of RAM. The development of the software involved Python (version 3.10.12) and TensorFlow (version 2.17.0). The dataset was partitioned with a 70:10:20 ratio, with 70% allocated for training the BERT model, 10% for validation, and the remaining 20% reserved for testing the model. The subsequent section provides a detailed discussion of the experimental results. More details about hyperparameters, training time, and computational resources are shown in Table 3 for our experiments.

Table 3. Model hyperparameters, time, and computational resources.

Model	Epochs	Batch Size	Learning Rate	Optimizer	Dropout Rate	Training Duration (µs)	Computational Resources
Neural Network	10	32	0.001	Adam	0.2	1.0041×10^8	Tesla K80 12 GB GPU and 32 GB RAM
Recurrent Neural Network	10	32	0.001	Adam	0.2	4.0526×10^8	Tesla K80 12 GB GPU and 32 GB RAM
BiLSTM	10	32	0.001	Adam	0.2	6.606×10^8	Tesla K80 12 GB GPU and 32 GB RAM
RoBERTa-urdu-small	5	16	2×10^{-5}	AdamW	N/A	1.77×10^8	Tesla K80 12 GB GPU and 32 GB RAM
BERT-large-cased	5	16	2×10^{-5}	AdamW	N/A	4.29×10^8	Tesla K80 12 GB GPU and 32 GB RAM
BERT-base-cased	5	16	2×10^{-5}	AdamW	N/A	1.39×10^8	Tesla K80 12 GB GPU and 32 GB RAM
BERT-multilingual	5	16	2×10^{-5}	AdamW	N/A	2.2×10^8	Tesla K80 12 GB GPU and 32 GB RAM

5. Results and Analysis

We leveraged the Urdu script to enhance the performance of Urdu NER. Moreover, we devised a data augmentation technique to expand the initial dataset. In this section, we present the results obtained by applying the BERT-multilingual model to both the original dataset and the augmented dataset. To assess the overall effectiveness of the final system, we employed Fida et al. [35] for the evaluation of metrics such as precision, recall, and F1 score.

Table 4 displays the comprehensive outcomes of the proposed Urdu named entity recognition (NER) approach. The results show a significant improvement in the NER system's performance with the extended dataset. Notably, the existing literature is scarce on the same dataset, prompting a comparison with results from different datasets. Initial training of BERT on the original dataset yielded an F1 score of 0.846, surpassing the F1 score achieved by the RNN-based method presented by Kanwal et al. [24]. It is noteworthy that the RNN approach employed Word2Vec, Glove, and FastText embedding, while our pre-trained BERT model utilized WordPiece embedding.

Our outcomes exhibited superiority over the RNN-based approach. Leveraging BERT's pre-training on Urdu corpora and fine-tuning for similar writing scripts significantly enhanced the NER system's performance. Furthermore, the extension of the dataset through CWEA augmentation and subsequent fine-tuning of the BERT model resulted in a remarkable improvement, yielding an enhanced F1 score of 0.98228%. This underscores the effectiveness of our data augmentation methodology in expanding the existing UNER dataset. In our results, Figures 4 and 5 shows the confusion matrix for the best-performing model for the UNER-II dataset and visually represents the outcomes of our BERT-multilingual model on the extended dataset, comparing them with recent research, particularly focusing on the best-performing model, i.e., RNN. Additionally, our method's comparative performance with a machine learning approach is detailed in Table 4. The analysis underscores that our BERT-multilingual model, coupled with augmentation, outperformed the top-performing approach in recent studies, namely, RNN. While BERT initially exhibited commendable performance against RNN, the introduction of CWEA augmentations significantly elevated the overall performance, a trend also noted in analogous studies like Dai et al. [36], where data augmentation contributed to the development of more accurate systems.

Table 4. Comparison of the BERT-multilingual model with RNN using MkPUCIT dataset.

Study	Dataset	Methods	Precision	Recall	F1-Score
Kanwal et al. [24]	MkPUCIT	MEMM	0.73	0.53	0.61
		CRF	0.77	0.61	0.68
		NN	0.76	0.75	0.75
		RNN	0.76	0.79	0.77
Proposed	UNER (Original Data)	NN	0.87	0.78	0.82
		RNN	0.86	0.83	0.84
		BiLSTM	0.85	0.83	0.84
		Ruberta-Urdu-small	0.79	0.77	0.78
		BERT-large-cased	0.73	0.67	0.70
		BERT-base-cased	0.64	0.59	0.62
		BERT-multilingual	0.82	0.80	0.85
	UNER-II (Augmented Data)	NN	0.96	0.96	0.96
		RNN	0.96	0.96	0.97
		BiLSTM	0.96	0.96	0.96
		RoBERTa-Urdu-small (CWEA)	0.89	0.88	0.88
		BERT-large-cased (CWEA)	0.87	0.86	0.92
		BERT-base-cased (CWEA)	0.88	0.89	0.91
		BERT-multilingual (CWEA)	**0.979**	**0.984**	**0.982**

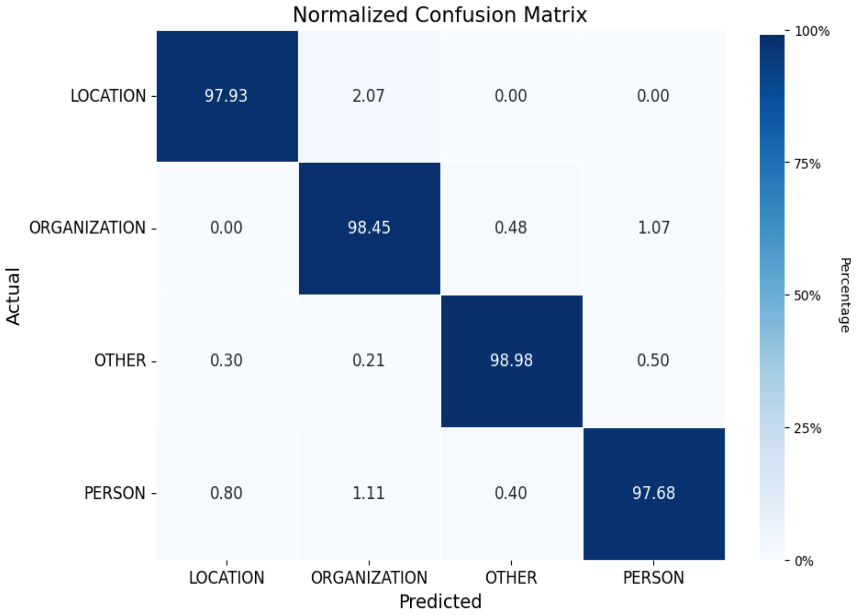

Figure 4. Confusion matrix for the best model.

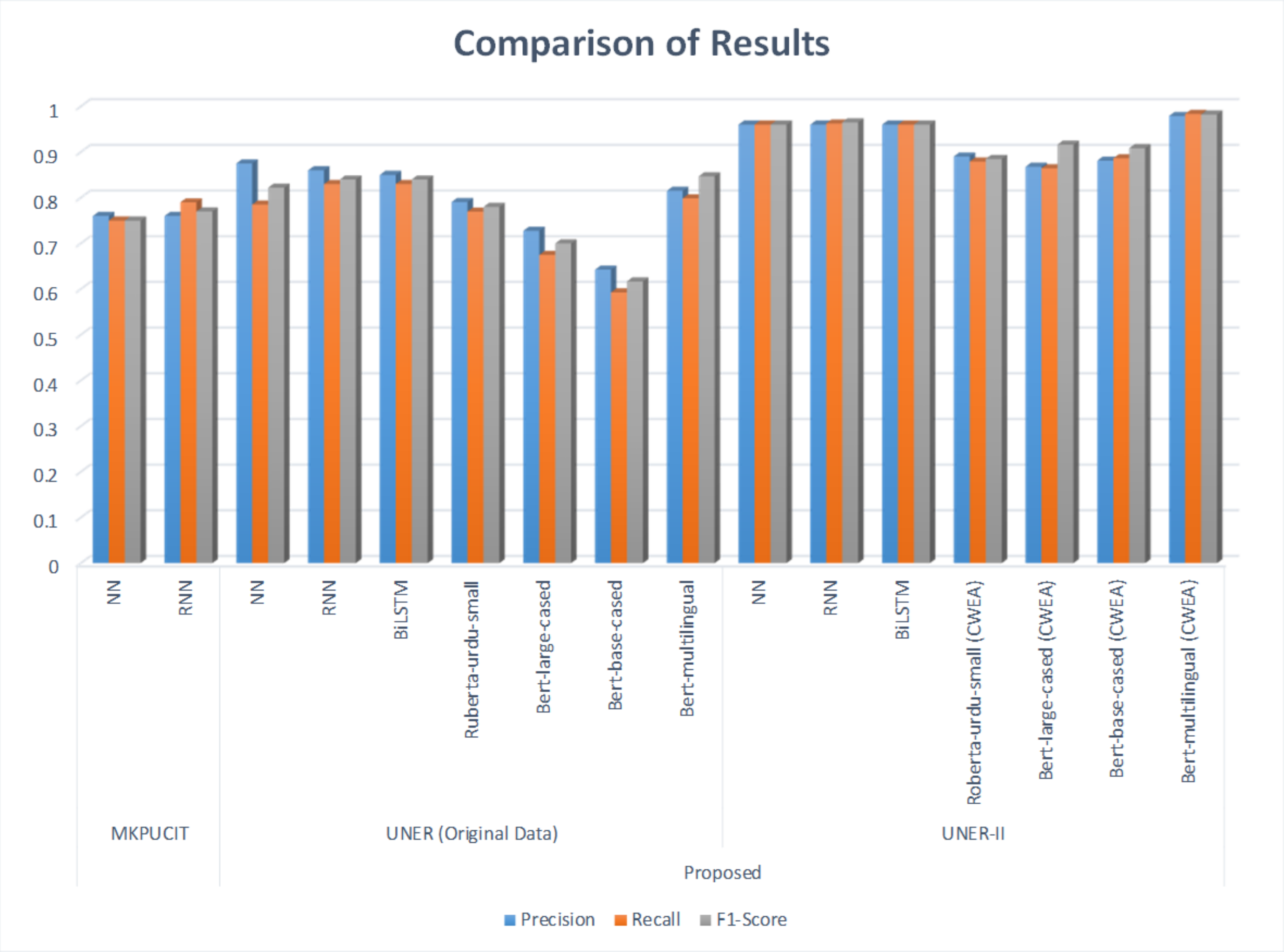

Figure 5. Comparison of the results with and without augmentation [24].

6. Discussion and Error Analysis

Named entity recognition (NER) boasts a broad spectrum of applications across various natural language processing (NLP) tasks, including streamlined search algorithms, content recommendation systems, and customer support applications, as evidenced by studies conducted by [37]. Therefore, in the current research, we introduce a novel data augmentation method for Urdu, aiming to enrich existing datasets, comprising 160,132 tokens and 114,912 labeled entities, significantly enhancing the coverage of named entities compared to previous datasets. We also evaluated several transformer models, including BERT-multilingual, RoBERTa-Urdu-small, BERT-base-cased, and BERT-large-cased, on this augmented dataset. For instance, the application of NER in news content classification is noteworthy, particularly for entities like news and publishing houses dealing with vast amounts of data daily. NER proves invaluable in automatically scanning extensive datasets to identify key information such as major locations, time, and individuals. However, the effectiveness of machine learning algorithms in training NER models relies heavily on the availability of abundant data, which is abundant for high-resource languages like English, but poses a significant challenge for low-resource languages. In the context of low-resource languages, such as Urdu, the limitations in available resources hinder the reliability of NER systems. In our study, the BERT-multilingual model given the highest macro F1 score of 0.982% on augmented data.

Unlike high-resource languages, where fine-tuning existing models on new datasets suffices, low-resource languages lack the requisite massive datasets. Addressing NER for low-resource languages has become a significant challenge in the research field. In response, researchers have begun improving low-resource word representations through knowledge transfer from high-resource languages [38–40]. The methods and findings from the current study have significant implications for named entity recognition (NER) systems in other low-resource languages beyond Urdu. Many low-resource languages, such as Pashto, Somali, Korean, Indonesian, and Amharic, face challenges similar to those of Urdu, including limited availability of labeled datasets, insufficient research focus, and unique linguistic complexities such as morphological richness and diverse script systems. Most importantly, our novel contextual word embeddings augmentation (CWEA) method, introduced to enhance Urdu NER datasets, demonstrates potential for application in other low-resource languages with comparable characteristics. Languages that utilize the Arabic script or exhibit agglutinative structures (e.g., Kurdish and Persian) could benefit from our data augmentation techniques. Furthermore, our evaluation of transformer models, such as BERT-multilingual, indicates that multilingual models can effectively leverage shared linguistic features across languages, making them suitable for use in other low-resource contexts. Additionally, the potential of cross-lingual transfer learning is recognized as a promising approach for improving NER systems in low-resource languages by utilizing datasets from high-resource languages. Languages that share vocabulary or structural similarities with extensively researched languages (e.g., Arabic for Urdu) may experience improved model performance by applying these techniques, despite the challenges of limited data availability. This generalization is crucial for extending the applicability of our approach to a wider range of languages.

In this paper, we proposed a BERT-multilingual approach for Urdu named entity recognition (NER) that tremendously improved the accuracy of the F1 score. We can suggest that the current approach could be adapted for other low-resource, ethnic minority languages, potentially extending our architecture to additional NLP tasks. We hope that our results inspire further advancements in NLP applications for low-resource agglutinative languages.

In the realm of NLP, various data augmentation techniques have been proposed for text classification, including translation, back-translation, and synonym word replacement [41]. However, these methods pose challenges in the context of Urdu due to the lack of reliable Urdu-to-English translation for the Arabic script, rendering methods like translation and back-translation impractical. Manual translation efforts are also deemed impractical due to the extensive time and effort required from individual data annotators and developers. Although some tech giants offer Urdu-to-English translation, they predominantly use the Arabic script, complicating the use of such augmentation methods. Additionally, the scarcity of synonym choices in pure Urdu makes synonym word replacement an unsuitable solution. Consequently, traditional augmentation methods are not directly applicable to address our research problem. In response to the challenges posed by data availability, our research proposes CWEA as an innovative approach. This augmentation method significantly improves Urdu NER results compared to recent studies by extending the existing Urdu dataset through simple yet effective CWEA augmentations.

Error Analysis

Misclassification is a frequent challenge in NER tasks involving the Urdu language. To investigate the misclassified tokens, we conduct a class-wise comparison between the actual and predicted labels using the test data for the best-performing model configurations. Table 5 presents the confusion matrix for the top-performing model across the entire dataset. Each entry in the matrix indicates the number of instances where the model's prediction corresponds to the actual class labels.

Table 5. Misclassifications of named entities.

Classes	Location	Organization	Other	Person
Location	6400	135	0	0
Organization	0	5916	29	64
Person	80	111	40	9730

Our manual investigation identified five main factors contributing to these errors: improper tokenization, insufficient representation of uncommon named entities in the training data, the presence of abbreviations or nicknames in the testing dataset, non-Urdu text, and incorrect disambiguation of tokens that can map to multiple named entity types. In the following sections, each named-entity type and examples of the misclassified instances are mentioned.

The examination revealed that the proposed technique occasionally misclassifies the named entities that are labeled in the dataset. In the following sections, we provided examples to clearly illustrate these issues.

1. LOCATION: The model accurately identifies 6400 instances as location (true positives for location). However, it misclassifies 135 instances as organization and makes no errors in categorizing instances as other or person. For example, for Mardan University (مردان یونورسٹی), the first part of the entity is location, and then یونورسٹی (university) [O].
2. ORGANIZATION: For the organization class, the model correctly identifies 5916 instances. There are 29 instances incorrectly labeled as OTHER and 64 as person, indicating some confusion between an organization and these classes. For instance, (باچا خان یونورسٹی) Bacha khan University was tokenized into two separate tokens, باچا خان and یونورسٹی, which were marked as person and other, respectively.
3. PERSON: For the person class, the model correctly classified 9730 instances. However, it incorrectly labeled 80 instances as location, 111 as organization, and 40 as other, indicating some level of confusion between person and these other categories. Examples of misclassified persons due to incorrect tokenization and scarcity of availability were اقبال لاہوری (Iqbal Lahorey) and شرپاؤ ہسپتال (Sherpao Hospital).

Overall, the model demonstrates strong performance in most categories, with higher accuracy for location, organization, and person. However, there is some confusion between these categories, especially between organization and other and between person and the other classes. This error analysis highlights areas where the model could be improved, particularly in distinguishing between organization and other, as well as person and other classes. To strengthen this section, future research should delve deeper into specific types of errors, particularly those arising from tokenization and category misclassification. Advanced tokenization methods such as byte-pair encoding (BPE) or WordPiece could address issues with compound words and entity splitting. Furthermore, leveraging contextualized embeddings (e.g., BERT or mBERT) can help better differentiate between categories by capturing richer context around ambiguous tokens. Addressing the issue of insufficient representation could be approached by augmenting the training data with rare entities or using external resources like gazetteers for named entities. Additionally, focusing on multilingual and code-mixed data handling could mitigate errors caused by non-Urdu text in the test dataset. Expanding the analysis to discuss these strategies would provide valuable insights for improving NER models and reducing misclassifications in future research.

7. Conclusions and Future Work

The availability of labeled data for NER, especially in the domain of computer science (CS), is notably scarce and poses a significant challenge for Urdu language text. In this research, we present an enhanced NER system for the Urdu script by leveraging multilingual

BERT and introducing a novel data augmentation technique known as CWEA. We extended the dataset to address the scarcity of existing Urdu data, resulting in 160,132 tokens and 114,912 named entities (NEs). Our experimental findings indicate that the extended dataset contributes to an overall improvement in the performance of Urdu NER, achieving an F1 score of 0.982%. This suggests that employing a practical augmentation approach can positively impact the NER task, particularly for languages with limited resources, by mitigating the need to gather and annotate new data while enhancing performance concurrently.

In the future, we aim to develop and curate a large, comprehensive dataset specifically tailored to the education domain. This dataset will feature a broader variety of entity types, allowing for more diverse and nuanced natural language processing applications within the education sector. Additionally, we will focus on designing and implementing advanced deep learning models and large language models to enhance named entity recognition capabilities within this specialized educational context.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/computers13100258/s1, Table S1: Data Collection, Table S2: Calculation of Mean and standard derivation differences, Table S3: Conduct paired *t*-test, Table S4: Detail approach for the Data augmentation Process. The steps of algorithms and pseudo-code used for the augmentation were also present in Supplementary Materials.

Author Contributions: Conceptualization, A.G. and F.U.; methodology, F.U. and M.T.Z.; software, G.S., F.U. and M.T.Z.; validation, A.G. and E.M.F.R.; formal analysis, A.G.; investigation, F.U.; resources, A.G., G.S. and E.M.F.R.; data curation, F.U. and M.T.Z.; writing—original draft preparation, F.U.; writing—review and editing, F.U. and E.M.F.R.; visualization, A.G.; supervision, A.G. and E.M.F.R. project administration, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was done with partial support from the Mexican government through grant A1-S-47854 of CONACYT, Mexico, and grants 20241816, 20,241,819, and 20,240,951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Acknowledgments: The authors thank CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alshammari, N.; Alanazi, S. The impact of using different annotation schemes on named entity recognition. *Egypt. Inform. J.* **2021**, *22*, 295–302. [CrossRef]
2. Yadav, V.; Bethard, S. A survey on recent advances in named entity recognition from deep learning models. *arXiv* **2019**, arXiv:1910.11470.
3. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.R.; Abdelmajeed, M.; Sadiq, M.T. Automatic detection of offensive language for Urdu and Roman Urdu. *IEEE Access* **2020**, *8*, 91213–91226. [CrossRef]
4. Sundheim, B.M. Overview of results of the MUC-6 evaluation. In Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, MA, USA, 6–8 November 1995.
5. Khan, W.; Daud, A.; Shahzad, K.; Amjad, T.; Banjar, A.; Fasihuddin, H. Named entity recognition using conditional random fields. *Appl. Sci.* **2022**, *12*, 6391. [CrossRef]
6. Khattak, A.; Asghar, M.Z.; Saeed, A.; Hameed, I.A.; Hassan, S.A.; Ahmad, S. A survey on sentiment analysis in Urdu: A resource-poor language. *Egypt. Inform. J.* **2021**, *22*, 53–74. [CrossRef]
7. Khan, I.U.; Khan, A.; Khan, W.; Su'ud, M.M.; Alam, M.M.; Subhan, F.; Asghar, M.Z. A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and roman Urdu language. *Computers* **2021**, *11*, 3. [CrossRef]
8. Riaz, K. Rule-based named entity recognition in Urdu. In Proceedings of the 2010 Named Entities Workshop, Uppsala, Sweden, 16 July 2010.
9. Malik, M.K.; Sarwar, S.M. Urdu named entity recognition and classification system using conditional random field. *Sci. Int.* **2015**, *5*, 4473–4477.

10. Saha, S.K.; Chatterji, S.; Dandapat, S.; Sarkar, S.; Mitra, P. A hybrid named entity recognition system for south and south east asian languages. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, Hyderabad, India, 12 January 2008.
11. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
12. Roberts, A.; Gaizauskas, R.J.; Hepple, M.; Guo, Y. Combining Terminology Resources and Statistical Methods for Entity Recognition: An Evaluation. In Proceedings of the LREC, Miyazaki, Japan, 7–12 May 2008.
13. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the CoNLL-2003, Edmonton, AB, Canada, 31 May–1 June 2003.
14. Shaalan, K.; Raza, H. NERA: Named entity recognition for Arabic. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1652–1663. [CrossRef]
15. Singh, U.; Goyal, V.; Lehal, G.S. Named entity recognition system for Urdu. In Proceedings of the COLING 2012, Mumbai, India, 8–15 December 2012.
16. Mukund, S.; Srihari, R.; Peterson, E. An information-extraction system for Urdu—A resource-poor language. *ACM Trans. Asian Lang. Inf. Process. (TALIP)* **2010**, *9*, 1–43. [CrossRef]
17. Zoya Latif, S.; Latif, R.; Majeed, H.; Jamail, N.S.M. Assessing Urdu Language Processing Tools via Statistical and Outlier Detection Methods on Urdu Tweets. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–31. [CrossRef]
18. Çoban, Ö.; Özel, S.A.; İnan, A. Deep learning-based sentiment analysis of Facebook data: The case of Turkish users. *Comput. J.* **2021**, *64*, 473–499. [CrossRef]
19. Haq, R.; Zhang, X.; Khan, W.; Feng, Z. Urdu named entity recognition system using deep learning approaches. *Comput. J.* **2023**, *66*, 1856–1869. [CrossRef]
20. Naz, S.; Umar, A.I.; Razzak, M.I. A hybrid approach for NER system for scarce resourced language-URDU: Integrating n-gram with rules and gazetteers. *Mehran Univ. Res. J. Eng. Technol.* **2015**, *34*, 349–358.
21. Collins, M.; Singer, Y. Unsupervised models for named entity classification. In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, USA, 21–22 June 1999.
22. Capstick, J.; Diagne, A.K.; Erbach, G.; Uszkoreit, H.; Leisenberg, A.; Leisenberg, M. A system for supporting cross-lingual information retrieval. *Inf. Process. Manag.* **2000**, *36*, 275–289. [CrossRef]
23. Jahangir, F.; Anwar, W.; Bajwa, U.I.; Wang, X. 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, Mumbai, India, 9 December 2012; pp. 95–104.
24. Kanwal, S.; Malik, K.; Shahzad, K.; Aslam, F.; Nawaz, Z. Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2019**, *19*, 1–13. [CrossRef]
25. Gali, K.; Surana, H.; Vaidya, A.; Shishtla, P.M.; Sharma, D.M. Aggregating machine learning and rule based heuristics for named entity recognition. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, Hyderabad, India, 12 January 2008.
26. Khan, W.; Daud, A.; Alotaibi, F.; Aljohani, N.; Arafat, S. Deep recurrent neural networks with word embeddings for Urdu named entity recognition. *ETRI J.* **2020**, *42*, 90–100. [CrossRef]
27. Ullah, F.; Zeeshan, M.; Ullah, I.; Alam, M.N.; Al-Absi, A.A. Towards Urdu Name Entity Recognition Using Bi-LSTM-CRF with Self-attention. In Proceedings of the International Conference on Smart Computing and Cyber Security: Strategic Foresight, Security Challenges, and Innovation, Gosung, Republic Korea, 28–29 October 2021; Springer: Singapore, 2021.
28. Balouchzahi, F.; Sidorov, G.; Shashirekha, H.L. ADOP FERT-Automatic Detection of Occupations and Profession in Medical Texts using Flair and BERT. In *IberLEF@SEPLN*; 2021. Spain. Available online: https://www.researchgate.net/publication/354795026_ADOP_FERT-Automatic_Detection_of_Occupations_and_Profession_in_Medical_Texts_using_Flair_and_BERT (accessed on 31 July 2024).
29. Sathyanarayanan, D.; Ashok, A.; Mishra, D.; Chimalamarri, S.; Sitaram, D. Kannada Named Entity Recognition and Classification using Bidirectional Long Short-Term Memory Networks. In Proceedings of the 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEEECCOT), Msyuru, India, 14–15 December 2018; pp. 65–71.
30. Dedes, K.; Utama AB, P.; Wibawa, A.P.; Afandi, A.N.; Handayani, A.N.; Hernandez, L. Neural Machine Translation of Spanish-English Food Recipes Using LSTM. *JOIV Int. J. Inform. Vis.* **2022**, *6*, 290–297. [CrossRef]
31. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
32. Suleman, M.; Asif, M.; Zamir, T.; Mehmood, A.; Khan, J.; Ahmad, N.; Ahmad, K. Floods Relevancy and Identification of Location from Twitter Posts using NLP Techniques. *arXiv* **2023**, arXiv:2301.00321.
33. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
34. Agrawal, A.; Tripathi, S.; Vardhan, M.; Sihag, V.; Choudhary, G.; Dragoni, N. BERT-based transfer-learning approach for nested named-entity recognition using joint labeling. *Appl. Sci.* **2022**, *12*, 976. [CrossRef]
35. Ullah, F.; Ullah, I.; Kolesnikova, O. Urdu named entity recognition with attention bi-lstm-crf model. In *Mexican International Conference on Artificial Intelligence*; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 3–17.
36. Dai, X.; Adel, H. An analysis of simple data augmentation for named entity recognition. *arXiv* **2020**, arXiv:2010.11683.

37. Daud, A.; Khan, W.; Che, D. Urdu language processing: A survey. *Artif. Intell. Rev.* **2017**, *47*, 279–311. [CrossRef]
38. Feng, X.; Feng, X.; Qin, B.; Feng, Z.; Liu, T. Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. *IJCAI* **2018**, *1*, 4071–4077.
39. Jin, G.; Yu, Z. A Korean named entity recognition method using Bi-LSTM-CRF and masked self-attention. *Comput. Speech Lang.* **2021**, *65*, 101134. [CrossRef]
40. Gunawan, W.; Suhartono, D.; Purnomo, F.; Ongko, A. Named-entity recognition for indonesian language using bidirectional lstm-cnns. *Procedia Comput. Sci.* **2018**, *135*, 425–432. [CrossRef]
41. Bayer, M.; Kaufhold, M.-A.; Reuter, C. A survey on data augmentation for text classification. *ACM Comput. Surv.* **2021**, *55*, 146. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.