From Courtroom to Corpora: Building a Name Entity Corpus for Urdu Legal Texts

Adeel Zafar[1], Sohail Ashraf[2], Slawomir Nowaczyk[1]
[1]CAISR, Halmstad University, Sweden
[2]Riphah International University, Islamabad, Pakistan
adeel.zafar@hh.se, cheemasohail286@gmail.com, slawomir.nowaczyk@hh.se

Abstract

This study explores the effectiveness of transformer-based models for Named Entity Recognition (NER) in Urdu legal documents, a critical task in low-resource language processing. Given the specialized terminology and complex syntax of the legal texts, accurate entity recognition in Urdu remains a challenge. We developed a legal Urdu dataset that contains 117,500 documents, generated synthetically from 47 different types of legal documents, and evaluated three BERT-based models. XLMRoBERTa, mBERT, and DistilBERT were analyzed by analyzing their performance on an annotated Urdu legal data set. mBERT demonstrated superior accuracy (0.999), and its F1 score (0.975) outperforms XLMRoBERTa and DistilBERT, highlighting its robustness in recognizing entities within low-resource languages. To ensure the privacy of personal identifiers, all documents are anonymized. The dataset for this study is publicly hosted on HuggingFace[1].

Keywords: Named Entity Recognition, Low Resource Languages, Synthetic Data, DistilBERT, XLM-RoBERTa, mBERT

1 Introduction

1.1 Motivation and Problem Statement

NER is one of the fundamental tasks of NLP that finds important uses in legal text analysis, knowledge mining, and semantic retrieval. Despite advances in NLP, low-resource languages such as Urdu still face major challenges, especially in tasks such as entity recognition within specialized domains such as law, where linguistic complexity and domain-specific terminology significantly impact accuracy and performance Naseem et al. (2020). Legal texts in Urdu pose certain challenges due to the legal vocabulary and, at the same time, due to the legal hierarchy into which the documents are classified, apart from having sectional differences. The above increases the need for models in automated legal text processing alongside custom datasets to effectively address these issues Krasadakis et al. (2024). According to the latest work, transfer learning and domain adaptation help to enhance NER performance in low-resource languages; the emphasis is placed on the requirement for fine-tuned datasets in this sphere Sasikumar and Mantri (2023).

The basic NER tasks have been resolved well by general-purpose transformer-based models such as BERT or RoBERTa. However, their application to specific NER tasks for Urdu legal texts has its limitations. General-purpose models are not as aware of the cultural and language characteristics of Urdu. The legal terminologies lead to suboptimal accuracy in identifying legal phrases and conditions. Furthermore, the absence of adequately annotated datasets, particularly those that capture the specialized vocabulary of legal documents in Urdu, significantly hampers the effectiveness of existing NER models. Although manual annotation is essential for building high-quality NER systems, it remains a labor intensive and time-consuming task, especially for low resource languages. The scarcity of large domain-specific annotated corpora in Urdu necessitates the development of a new data set tailored to legal texts, along with a systematic evaluation of methodologies capable of addressing the linguistic challenges inherent in the Urdu language.

1.2 Objectives and Contributions

This research contributes to Urdu NER for legal documents through the following:

[1]https://huggingface.co/datasets/cheemasohail/Urdu-Legal_ner_corpora

Proceedings of Recent Advances in Natural Language Processing,pages 1396–1405
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-161

1. Developed a synthetic Urdu NER dataset for legal texts by generating 117,500 annotated documents enriched with domain-specific lexical and structural features.

2. Experimental analysis of three transformer-based models for Urdu legal NER explicitly comparing the efficiency of these models in the context of the low-resource language.

2 Literature Review

2.1 Datasets for Urdu NER and NLP

Datasets are crucial for the development and validation of NER systems. Some significant datasets of the Urdu language are:

- **MK-PUCIT Dataset:** This is the largest data set available for Urdu NER that is annotated for only person, organization and location entities. It contains 926,776 tokens and 99,718 annotated entities Kanwal et al. (2019a).
- **IJCNLP 2008 Dataset:** This dataset is one of the earliest datasets developed for the Urdu NER task that is comprehensively annotated. The dataset contains 40,408 words, 1,097 sentences, and 1,115 entities for twelve classes Hussain (2008).
- **UNER Dataset:** UNER dataset was developed from the BBC Urdu news, which includes national, international news, and sports articles. There are 48,000 words and 4,621 named entities for seven classes Khana et al. (2016).
- **Jahangir et al. Dataset:** This dataset comprises 13,860 words and 1,526 named entities for four classes Jahangir et al. (2012).
- **Kamran-PU-NE:** The dataset consists of 652852 tokens with 44480 entities having three different classes Malik (2017).

There are some other general-purpose datasets for Urdu NLP applications that include **Urdu News Dataset 1M** developed by Hamza et al. (2019), **LEGAL-UQA Dataset** developed by Faisal and Yousaf (2024), **Roman Urdu Dataset** developed by Mehmood et al. (2019), and **UQA Corpus** developed by Arif et al. (2024b).

2.2 Existing Approaches for Urdu NER

There are three main approaches used for the Urdu NER task.

2.2.1 Rule-Based Approaches:

Rule-based models work linearly and process text slowly, but they provide good accuracy. Riaz (2010) compared Conditional Random Fields, evaluated on the IJCNLP-08 dataset, with the rule-based model, applied on the Becker-Riaz corpus. The accuracy of the rule-based model was much better in comparison to the statistical model. It was also observed that each rule contributes to the accuracy of the results. Singh et al. (2012) defined ten rules to develop an Urdu NER system and applied it to the IJCNLP-08 dataset, which gave an F1 score of 88.1%. Rule-based approaches are efficient for smaller and controlled scenarios, but do not perform well for generalized contexts Arif et al. (2024a). Rule-based NER systems for the Urdu language can effectively detect named entities when rules are derived from linguistic patterns and domain-specific knowledge Tahir et al. (2024).

2.2.2 Machine-Learning Based Approaches:

Some of the models cannot be used efficiently because of an insufficient amount of datasets in the Urdu language. For example, Cotterell and Duh (2024) used character-level neural Conditional Reference Fields (CRFs) instead of log-linear Conditional Reference Fields (CRFs). A deep learning model, Artificial Neural Network (ANN) outperformed probabilistic statistical model, Hidden Markov Model (HMM) in terms of accuracy, balance between precision and recall as both models are evaluated on Kamran-PU-NE dataset Malik (2017).
Urdu NER systems can benefit from the use of diverse BERT models; however, they encounter challenges stemming from the limited availability of training data and the intrinsic linguistic differences of the language Ullah et al. (2024). Several Deep Recurrent Neural Network (DRNN) models provide more accurate results for Urdu NER as compared to the baseline linear-chain conditional random fields (CRF) and artificial neural networks (ANN) models Khan et al. (2020).

2.2.3 Hybrid Approaches:

These approaches help to overcome some difficulties related to NLP tasks. Mukund and Srihari (2009) used Conditional Random Fields (CRFs) for NER tagging that compensates for limited training data and improves performance through bootstrapping combined with grammar rules and lexicon lookups. Another combination of Conditional Ran-

dom Fields (CRFs) with language-specific heuristics was used for machine learning models to post-process the datasets. This approach also helps to generalize the models for multiple languages Gali et al. (2008). A hybrid NER system was developed by using Maximum Entropy (MaxEnt) model, and then the performance was enhanced by using language-specific rules Saha et al. (2008).

2.3 Comparative Analysis:

Table 1 presents a comprehensive analysis of existing literature on various NER tasks in the Urdu language.

2.4 Research Gaps in Existing Literature

Although there are ongoing efforts to improve NLP tasks for the Urdu language—particularly in NER several challenges still persist. These challenges stem from the unique characteristics of Urdu, such as the lack of capitalization, complex morphology, and the use of the joined-up Nastaliq script[2]. Another issue is a lack of resources in terms of large-scale annotated datasets and domain-specific vocabulary. It is also challenging for models to generalize because of the various variations in the use of words and phrases in real-world datasets.

3 Data Preparation

3.1 Corpus Description

The corpus includes 10 different main categories and, in total, 47 subgenres of contemporary Urdu legal text, which may pertain to different fields of the Pakistani legal system. Such sources comprise legal records of cases, contracts, real estate, affidavits, and legal announcements, among others. Every type of document differs both in terms of language and structure, and is distinguished within the legal context.

- **Judicial Documents:** Comprise judgments, orders, motions, and pleadings, each preserving different process terminology as well as coded vocabularies practiced in legal decisions and filings.
- **Contracts and Agreements:** Encompasses commercial and residential tenancy and sale contracts, and employment and nondisclosure contracts, all shared by legalistic and formulaic provisions that proscribe and prescribe relationships and behaviors.
- **Property Documents:** Includes property deeds, wills, and sales agreements, where legal ownership and transfer of property are laid out in a legal framework.
- **Affidavits and Testimonies:** Being legal documents, they are based on sworn statements, affidavits, and testimonials. These practical documents are formatted to provide accurate statements.
- **Legal Notices:** Includes claims, eviction notices, employment termination, and debt recovery notices, each a formal notice in the given legal premises.
- **Litigation Documents:** Coordinates complaints, responses, applications for an injunction, and intercepts that help clarify the case as well as pre-trial processes.
- **Financial Documents:** Such documents are financial statements, tax returns, loans, and receipts, since these are important documents documenting legal and financial responsibilities.
- **Police and Investigation Documents:** They covers FIR, charge sheets, investigation reports, and testimony records very essential in the judicial process of investigative work.
- **Family Documents:** Includes divorce deeds, marriage certificates, wills, alimony agreements, and guardianship, all of which document family and personal transactions.
- **Business Documents:** These consist of business incorporation, agreements and articles, licenses, shareholding, and shareholders' agreements that are vital to corporate control and legal requirements.

[2]Nastaliq is a calligraphic style that serves Persian script document creation for Urdu together with Persian and Pashto language writing purposes. The calligraphic style presents an elegant appeal to the eye.

Reference	Approach	Dataset	Models	No. of NEs	F1 Score
Singh et al. (2012)	Rule-Based	IJCNLP-08	10 rules	13	88.1%
Riaz (2010)	Rule-Based	Becker-Riaz corpus	11 Rules	6	91.5%
Khan et al. (2022)	ML-Based	UNER-I	Condition Random Field (CRF)	7	74.67%
Malik (2017)	ML-Based	KPU-NE	Hidden Morkov Model (HMM)& Artifitial Neural Network (ANN)	3	84.17%
Riaz et al. (2020)	ML-Based	IJCNLP-08	Maximum Entropy Model	12	92%
Kanwal et al. (2019b)	ML-Based	UNER	NN & RNN	3	49%
Khan et al. (2020)	DL-Based	IJCNLP-08	Deep RNN & LSTM	3	81%
Ullah et al. (2021)	DL-Based	UNER	Bi LSTM with self attention & CRF	3	93%
Ullah et al. (2022)	DL-Based	MK-PUICT corpus	Attention-Bi-LSTM-CRF	3	92%
Saha et al. (2008)	Hybrid	IJCNLP NERSSEAL	MaxEnt Model & Defined Rules	12	48%
Gali et al. (2008)	Hybrid	IJCNLP-08	CRF & Heuristic Rules	12	45%
Mukund and Srihari (2009)	Hybrid	Urdu Newswires	CRF & Bigram HMM	3	69%
Khan et al. (2024)	DL-Based	UrduDic	FMM and RMM	-	97%
Muskaan et al. (2023)	DL-Based	Wikiann	RoBERTa	-	98%
Patil et al. (2020)	ML-Based	UNER-I	CRF Model	12	74.81%
Ullah et al. (2024)	DL-Based	CWEA	BERT-multilingual	-	98.20%
Anam et al. (2024)	DL-Based	UNER	BiLSTM	-	98%

Table 1: Detailed table of model specifications and performances

The corpus enables the NER model to capture legal Urdu languages and document structure, allowing it to learn about variations within distinct documentation. The list of all 47 documents is mentioned in Appendix table 5.

3.2 Synthetic Data Generation

A major problem for training legal tools is the shortage of large annotated datasets for Urdu legal documents. To solve this problem, we have used synthetic data by merging base CoNLL-formatted documents and custom dictionaries. The key steps are described below.

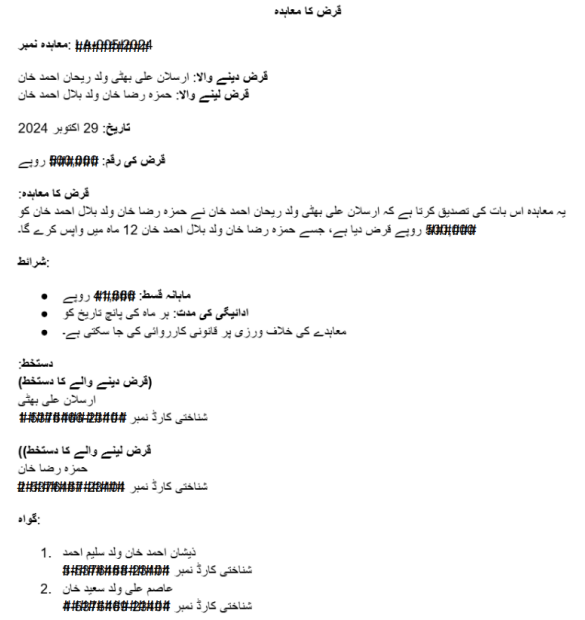
- **Step 1: Base CoNLL Format Creation**

  Representative samples from each of the 47 document types (a sample document in Urdu and translated version are shown in Figures 1 & 2) were converted into a widely used structure called CoNLL, which works well for NER tasks. The CoNLL format, line by line, divides tokenized words and links them with the respective entity tags. We merged these samples in the CoNLL format to build a primitive dataset that describes the overall structural and valuable reproduction of Urdu legal papers in terms of vocabulary. The CoNLL format annotation scheme was employed, and each word is tagged with an entity tag showing its position in the document. The 12 entity types were tagged in the B-I-O (Beginning-Inside-Outside) method because they include multi-token entities such as legal terms and personal names.

- **Step 2: Identifying Field and Dictionary Types**

  There are 12 main entities related to the Urdu legal frameworks when analyzing our dataset, as shown in Table 2 and the distribution of named entities in Figure 3. To maintain dataset diversity, we created a main dictionary with 25 varied entries per entity, reflecting usage in Urdu legal contexts. These examples—names, locations, dates, legal terms, numerals—were chosen for lexical and contextual variability.

- **Step 3: Creation of Synthetic Dataset** We utilized the original documents as templates and compiled dictionaries containing 25 unique entries for each named entity. A



قرض کا معاہدہ

معاہدہ نمبر: LA-005/2024

قرض دینے والا: ارسلان علی بھٹی ولد ریحان احمد خان
قرض لینے والا: حمزہ رضا خان ولد بلال احمد خان

تاریخ: 29 اکتوبر 2024

قرض کی رقم: 500,000 روپے

قرض کا معاہدہ:
یہ معاہدہ اس بات کی تصدیق کرتا ہے کہ ارسلان علی بھٹی ولد ریحان احمد خان نے حمزہ رضا خان ولد بلال احمد خان کو 500,000 روپے قرض دیا ہے، جسے حمزہ رضا خان ولد بلال احمد خان 12 ماہ میں واپس کرے گا۔

شرائط:

- ماہانہ قسط: 41,666 روپے
- ادائیگی کی مدت: ہر ماہ کی پانچ تاریخ کو
- معاہدے کی خلاف ورزی پر قانونی کارروائی کی جا سکتی ہے۔

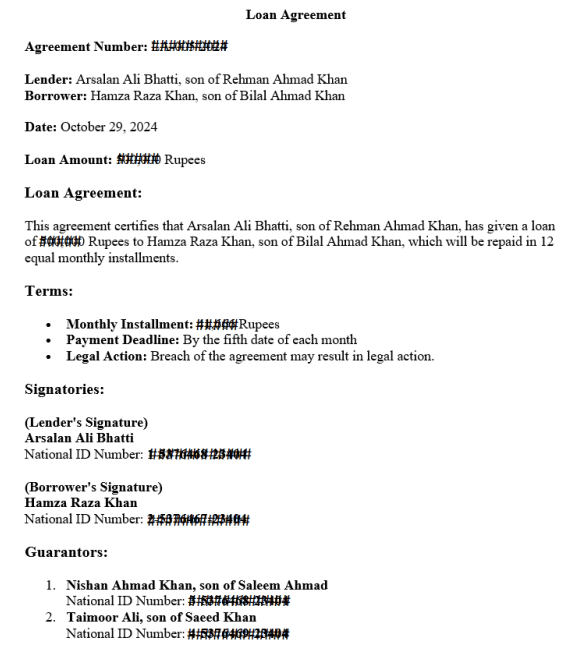
دستخط:
(قرض دینے والے کا دستخط)
ارسلان علی بھٹی
شناختی کارڈ نمبر 35202-1234567-1

(قرض لینے والے کا دستخط)
حمزہ رضا خان
شناختی کارڈ نمبر 35202-7654321-3

گواہ:

1. نیشان احمد خان ولد سلیم احمد
   شناختی کارڈ نمبر 35202-1122334-5
2. عاصم علی ولد سعید خان
   شناختی کارڈ نمبر 35202-9988776-1

Figure 1: A sample of an Urdu document before converting into CoNLL format.



Loan Agreement

Agreement Number: LA-005/2024

Lender: Arsalan Ali Bhatti, son of Rehman Ahmad Khan
Borrower: Hamza Raza Khan, son of Bilal Ahmad Khan

Date: October 29, 2024

Loan Amount: 500,000 Rupees

Loan Agreement:

This agreement certifies that Arsalan Ali Bhatti, son of Rehman Ahmad Khan, has given a loan of 500,000 Rupees to Hamza Raza Khan, son of Bilal Ahmad Khan, which will be repaid in 12 equal monthly installments.

Terms:

- **Monthly Installment:** 41,666 Rupees
- **Payment Deadline:** By the fifth date of each month
- **Legal Action:** Breach of the agreement may result in legal action.

Signatories:

(Lender's Signature)
Arsalan Ali Bhatti
National ID Number: 35202-1234567-1

(Borrower's Signature)
Hamza Raza Khan
National ID Number: 35202-7654321-3

Guarantors:

1. **Nishan Ahmad Khan, son of Saleem Ahmad**
   National ID Number: 35202-1122334-5
2. **Taimoor Ali, son of Saeed Khan**
   National ID Number: 35202-9988776-1

Figure 2: English translated version before converting into CoNLL format.

Sr. No.	ENTITY	EXPLANATION
1	LEGAL_ACTION	Title of the document
2	ID	CNIC number
3	PERSON	Name of any person
4	CASE_NUMBER	Reference ID of the document
5	DATE	Date
6	AMOUNT	Money
7	DESIGNATION	Official post or designation
8	ORG	Organization name
9	LOCATION	City name
10	PERCENTAGE	Percentage
11	ITEM	Thing or piece of land
12	DURATION	Amount of time

Table 2: List of entities and their description

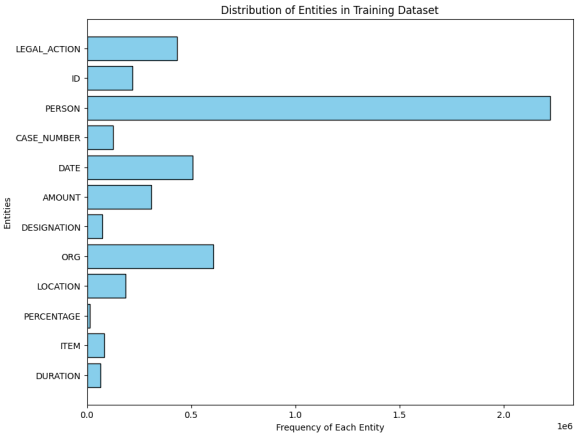


Figure 3: Entity distribution on the entire dataset

Python script was developed to process all 47 templates—already formatted in CoNLL—by randomly selecting and inserting entity values from the corresponding dictionaries to generate synthetic documents. The formation of the actual structures in the documents and the placement of entities are realistic and possible for Urdu legal documents. This process was done in parallel, and we generated 2500 synthetic documents for each type. The purpose of using random sampling served well to guarantee high variation in terms of getting full representation of diversity, reduction of bias, and generalization of findings by generating documents. This methodology reflects real-life legal language structures and variations within each type of document. This makes a synthetic collection of documents in CONLL format of 117,500 documents, which serves as a strong dataset for the NER models.

- **Step 4: Quality Control** After the generation of a large Urdu legal dataset, the quality and accuracy of the language in synthetic documents must be up to the mark, so we also introduced a quality check mechanism. We randomly selected a subset of 10% of the dataset for manual analysis. To examine this subset, we took help from domain experts who were also fluent Urdu speakers. The documents were investigated for grammatical acceptability, general linguistic features for Urdu, and context for legal purposes. After the results of this manual review, we identified that about 2% of the sampled documents need some corrections. These corrections are made on the selected subset of the dataset but not on the full synthesized dataset, to maintain the realistic nature of the data. The process of confirming the accuracy of the data is important because that will also reflect in the performance of the NER system.

3.3 Preprocessing and Standardization

As shown in Fig. 3, the entity distribution requires the real documents to be processed according to the CoNLL format—this involves breaking sentences so that each token appears on a separate line for proper tagging.

3.4 Data Privacy and Ethics

The original legal documents that are publicly available are used to create this dataset. To ensure the privacy of the individuals, we removed all the identifiers, including names, case numbers, national identity numbers, and addresses. These personal identifiers were replaced by randomly selecting placeholders from the dictionary created to keep the context of the documents preserved. All the sensitive and identifiable content is removed. This dataset is strictly for academic and research purposes.

4 Methodology

We applied three transformer-based models namely, **DistilBERT**, **XLM-RoBERTa** and **Multilingual BERT (mBERT)** for entity identification. These models are especially useful for low-resource languages with much better computational efficiency. The architecture details of each model are given in Table 3.

4.1 Implementation Constraints and Model Execution

4.1.1 Constraints:

Particular operational limitations governed the entire research process from methodology development until the project was finished. Computational

Parameters	DistilBERT	XLM-RoBERTa	mBERT
Weight decay	0.01	0.01	0.01
Linear Layer Size	768	768	768
Sequence Length	512	512	512
Attention Heads	12	12	12
Encoder Blocks	6	12	12
Optimizer	AdamW	AdamW	AdamW
Dropout	0.1	0.1	0.1
Batch Size	32	32	32
Learning Rate	2e-5	2e-5	2e-5
Epochs	1	1	1

Table 3: Hyperparameter configuration for DistilBERT, mBERT, and XLMRoBERTa

boundaries represented the key challenge because they reduced the potential to execute full-scale training on the dataset. The research was carried out using a quarter of the complete dataset of synthetic Urdu legal documents created for analysis. We had to reduce the data set by 75% to make it feasible to train and evaluate advanced analytics models with current hardware and run each model for one epoch.

4.1.2 Model Execution:

Due to operational restrictions, the NER model implementation demanded specific training configuration choices to achieve efficient learning results during limited training cycles. To maintain the diversity and unique features of the dataset, we randomly selected a subset of entire dataset that contains 25% of the original corpus. Multiple epochs help models to get more exposure to the dataset but for the limited computational resources, we trained all of three models only for one epoch and with 32 batch size that also helped us to reduce training time significantly. Most of the hyperparameters are kept same for every model applied, so a fair comparison between models could be possible.

4.1.3 Evaluation Metrics

Special attention was given to selecting evaluation metrics that provide meaningful insights, even with shorter training durations. The model's performance was assessed using macro-averaged Accuracy, which calculates the average accuracy across all entity classes; macro F1-score, which balances precision and recall to evaluate the model's ability to handle misclassified named entities; macro Recall, which measures the proportion of correctly predicted positive instances; and macro Precision,

Model	Accuracy Macro	F1 Macro	Recall	Precision Macro	Training Loss	Validation Loss
DistilBERT	0.965	0.265	0.261	0.320	0.206	0.168
XLM-RoBERTa	0.996	0.845	0.858	0.840	0.052	0.034
mBERT	0.999	0.975	0.975	0.979	0.028	0.014

Table 4: Performance comparison of language models

which reflects the proportion of correct positive predictions among all predicted positives. Additionally, training and validation loss were monitored to ensure a reliable evaluation of the reduced dataset.

5 Results and Discussion

All performance indicators shown in Table 4 demonstrate that DistilBERT produces inferior results in comparison to alternative models during this experiment. The reduced model complexity, along with smaller size, explains why this model shows limited ability to understand complex legal language in Urdu. XLM-RoBERTa demonstrates superior performance across all metrics compared to DistilBERT because it possesses solid resistance to diverse linguistic inputs while leveraging its massive multilingual training process. The best performance emerges from mBERT, which demonstrates nearly perfect execution across all metrics, indicating its training on a wide multi-language dataset has produced effective Urdu legal text processing.

The observed difference in system performance stems from multiple elements that span model depth and training dataset breadth, and language structure capabilities of both systems. Models using stronger contextual embedding capacities, such as mBERT and XLM-RoBERTa, demonstrate better performance in operating with legal documents that contain special vocabulary elements and challenging sentence complexity.

6 Conclusion

This research contributes to the field of NER systems for low-resource languages by introducing a large annotated Urdu legal NER dataset. The dataset is specifically tailored to the domain-specific challenges and complexities in the legal context. This dataset can play a crucial role as a foundational resource for the advancement of the NER task in the Urdu language. By implementing deep multilingual models, we demonstrated that mBERT and XLM-RoBERTa provide high performance for NER tasks for Urdu, while DistilBERT is computationally efficient, but it provides less ac-

curacy, which can be the trade-off between speed and accuracy in practical applications.

We found that the selection of the model plays a crucial role in domain-specific NER tasks for low-resource languages. For future work, the diversity and volume of the dataset can be expanded by adding more types of documents or by including more and more real documents. Different strategies can be used to train NER systems or models to improve speed-performance balance and accuracy as well.

References

Rimsha Anam, Muhammad Waqas Anwar, Muhammad Hasan Jamal, Usama Ijaz Bajwa, Isabel de la Torre Diez, Eduardo Silva Alvarado, Emmanuel Soriano Flores, and Imran Ashraf. 2024. A deep learning approach for named entity recognition in urdu language. *Plos one*, 19(3):e0300725.

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024a. Generalists vs. specialists: Evaluating large language models for urdu. *arXiv preprint arXiv:2407.04459*.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024b. Uqa: Corpus for urdu question answering. *arXiv preprint arXiv:2405.01458*.

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*.

Faizan Faisal and Umair Yousaf. 2024. Legal-uqa: A low-resource urdu-english dataset for legal question answering. *arXiv preprint arXiv:2410.13013*.

Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth M Shishtla, and Dipti Misra Sharma. 2008. 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*.

Syed Ali Hamza, Bilal Tahir, and Muhammad Amir Mehmood. 2019. Domain identification of urdu news text. In *2019 22nd International Multitopic Conference (INMIC)*, pages 1–7. IEEE.

Sarmad Hussain. 2008. Resources for urdu language processing. In *Proceedings of the 6th workshop on Asian Language Resources*.

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. 2019a. 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.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019b. 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.

Asif Khan, Khairullah Khan, Wahab Khan, Sadiq Nawaz Khan, and Rafiul Haq. 2024. Knowledge-based word tokenization system for urdu. *Journal of Informatics and Web Engineering*, 3(2):86–97.

Wahab Khan, Ali Daud, Fahd Alotaibi, Naif Aljohani, and Sachi Arafat. 2020. Deep recurrent neural networks with word embeddings for urdu named entity recognition. *ETRI Journal*, 42(1):90–100.

Wahab Khan, Ali Daud, Khurram Shahzad, Tehmina Amjad, Ameen Banjar, and Heba Fasihuddin. 2022. Named entity recognition using conditional random fields. *Applied Sciences*, 12(13):6391.

Wahab Khana, Ali Daudb, Jamal A Nasira, and Tehmina Amjada. 2016. Named entity dataset for urdu named entity recognition task. *Language & Technology*, 51.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3).

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.

Khawar Mehmood, Daryl Essam, and Kamran Shafi. 2019. Sentiment analysis system for roman urdu. In *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 1*, pages 29–42. Springer.

Smruthi Mukund and Rohini K Srihari. 2009. Ne tagging for urdu based on bootstrap pos learning. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies (CLIAWS3)*, pages 61–69.

Maurya Muskaan, Mandal Anupam, Maurya Manoj, Gupta Naval, and Nayak Somya. 2023. Neural language model embeddings for named entity recognition: A study from language perspective. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 44–51.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.

Nita Patil, Ajay Patil, and BV Pawar. 2020. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188.

Fatima Riaz, Muhammad Waqas Anwar, and Humaira Muqades. 2020. Maximum entropy based urdu named entity recognition. In *2020 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–5. IEEE.

Kashif Riaz. 2010. Rule-based named entity recognition in urdu. In *Proceedings of the 2010 named entities workshop*, pages 126–135.

Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, and Pabitra Mitra. 2008. 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*.

Nevasini Sasikumar and Krishna Sri Ipsit Mantri. 2023. Transfer learning for low-resource clinical named entity recognition. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 514–518, Toronto, Canada. Association for Computational Linguistics.

UmrinderPal Singh, Vishal Goyal, and Gurpreet Singh Lehal. 2012. Named entity recognition system for urdu. In *Proceedings of COLING 2012*, pages 2507–2518.

Muhammad Shoaib Tahir, Mahnoor Amjad, Minnaa Ahmad, Mahnoor Ikram, and Namra Fazal. 2024. Named entity recognition for urdu language. *Remittances Review*, 9(1):2724–2732.

Fida Ullah, Alexander Gelbukh, Muhammad Tayyab Zamir, Edgardo Manuel Felipe Rivern, and Grigori Sidorov. 2024. Enhancement of named entity recognition in low-resource languages with data augmentation and bert models: A case study on urdu. *Computers*, 13(10):258.

Fida Ullah, Ihsan Ullah, and Olga Kolesnikova. 2022. Urdu named entity recognition with attention bi-lstm-crf model. In *Mexican International Conference on Artificial Intelligence*, pages 3–17. Springer.

Fida Ullah, Muhammad Zeeshan, Ihsan Ullah, Md Nur Alam, and Ahmed Abdulhakim Al-Absi. 2021. Towards urdu name entity recognition using bi-lstm-crf with self-attention. In *International conference on smart computing and cyber security: strategic foresight, security challenges and innovation*, pages 403–407. Springer.

A Appendix

A.1 Types of Legal Documents in the Corpus

These are the titles of the legal documents used to generate the synthetic dataset.

Sr. No	Document Name	Sr. No	Document Name
1	Judicial Documents	25	Legal Notices
2	Judgments	26	Notice of Claim
3	Case Orders	27	Eviction Notice
4	Petitions	28	Employment Termination Notices
5	Claims	29	Debt Recovery Notice
6	Prosecution Documents	30	Litigation Documents
7	Status Reports	31	Complaint
8	Contracts and Agreements	32	Answer/Response
9	Sale Agreements	33	Injunction Applications
10	Lease/Rental Agreements	34	Interrogatories
11	Partnership Agreements	35	Voluntary Statements
12	Employment Contracts	36	Financial Documents
13	Loan Agreements	37	Financial Statements
14	Non-Disclosure Agreements (NDA)	38	Tax Documents
15	Property Documents	39	Loan Documents
16	Land Records	40	Receipts for Payments
17	Inheritance Documents	41	Police and Investigation Documents
18	Sale Deeds	42	FIR (First Information Report)
19	Possession Certificates	43	Charge Sheet
20	Registration Documents	44	Investigation Reports
21	Affidavits and Testimonies	45	Testimony Records
22	Affidavits	46	Family Documents
23	Witness Statements	47	Divorce Deeds
24	Attestations		

Table 5: Types of legal documents in the corpus