Evaluation of a Hybrid Approach to Legal Entity Recognition

Fernando Hurias Lopes[1], **Luís Paulo Faina Garcia**[1]

[1]Department of Computer Science – University of Brasília (UnB)
Brasília – DF – Brazil

fernando.neto@aluno.unb.br luis.garcia@unb.br

Abstract. *The Brazilian judiciary system's extensive volume of documents and technical language necessitates efficient methods for automating legal text analysis, where Legal Entity Recognition (LER) presents a significant challenge. This study evaluated the performance of LER models within the Brazilian legal domain through a comprehensive assessment across all publicly available Portuguese legal datasets: LeNER-BR, CDJur-BR, and UlyssesNER-BR. Eleven models were evaluated, encompassing classical-based, transformer-based, and hybrid approaches. Using Precision, Recall, and F1-Score metrics, the evaluation indicated that hybrid approaches consistently outperform both classical-based and standalone transformer-based approaches in legal entity extraction tasks.*

1. Introduction

The Brazilian justice system, with its complex hierarchical structure (STF, STJ, etc.), faces critical efficiency challenges, exacerbated by a growing volume of cases and the intricate nature of contemporary legal issues [Siqueira et al. 2023]. In this context, Artificial Intelligence (AI) and Natural Language Processing (NLP) are emerging as strategic tools to optimize the analysis of legal data, which often consists of long, unstructured texts [Yadav and Bethard 2019]. A fundamental task in this process is Legal Entity Recognition (LER), which automatically identifies and classifies key elements in legal documents, such as party names, statutory articles, and precedents [Li et al. 2020].

While classical-based approaches, such as Conditional Random Fields (CRF) and Long Short-Term Memory (LSTM) and, more recently, transformer-based approaches, such as Bidirectional Encoder Representations from Transformers (BERT), have advanced the state of the art in LER, a significant gap persists in the Brazilian legal domain. There is a lack of studies that systematically compare the performance of these different classical-based, transformer-based, and, crucially, hybrid approaches to the diverse types of documents that comprise Brazil's legal ecosystem. The choice of the ideal model for a given context remains an open question, hindering the implementation of effective and reliable AI solutions.

This work directly addresses this gap by presenting the first comprehensive and systematic empirical evaluation of LER models for Brazilian legal Portuguese. Our contribution is grounded in three core pillars. First, we evaluate a broad range of models, with 11 distinct architectures covering the spectrum from classical-based and transformer-based approaches to, most importantly, hybrid configurations, allowing us to investigate whether combining architectures can overcome the limitations of standalone models. Second, we ensure complete dataset coverage by conducting experiments on the three primary public datasets for the domain LeNER-Br [Luz De Araujo et al. 2018], UlyssesNER-Br [Albuquerque et al. 2022], and CDJur-Br [Brito et al. 2023], making our findings representative of technical documents, judicial decisions, and legislative texts. Finally, we maintain statistical rigor by validating all comparisons with significance tests, ensuring that observed performance differences are consistent and not the result of chance.

Thus, our central contribution is to establish a robust benchmark for the LER task in Brazil, providing a practical guide to the advantages and disadvantages of each approach and highlighting the potential of hybrid models for practical applications in the justice system.

The article is structured as follows: Section 2 presents related work in LER; Section 3 details the proposed approach; Section 4 describes the methodology; Section 5 analyzes the results; and Section 6 presents our conclusions and directions for future research.

2. Related Work

The evolution of models for Named Entity Recognition (NER) in the Brazilian legal domain began with classical-based approaches. Luz de Araújo *et al.* (2018) [Luz De Araujo et al. 2018] implemented a BiLSTM-CRF model with GloVe and character embeddings, achieving an average F1-Score of 92.53% on the LeNER-BR corpus. However, in the more specific legislative domain of UlyssesNER-Br, Albuquerque *et al.* (2022) [Albuquerque et al. 2022] observed that a simpler CRF model outperformed the neural network architecture, reaching an F1-Score of 80.8%, while the BiLSTM-CRF obtained 76.89%, indicating that model complexity does not always guarantee better performance.

The introduction of transformer-based approaches, such as BERT, represented a leap in performance. Costa *et al.* (2022) [Costa et al. 2022] demonstrated that a fine-tuned version of BERT (BERTimbau) achieved an F1-Score of 73.90%, surpassing CRF and BiLSTM-CRF models on an NER task in a corpus that mixes formal and informal domains. The superiority of BERT was corroborated by Brito *et al.* (2023) [Brito et al. 2023] on the complex CDJUR-BR corpus, where the model obtained a macro F1-Score of 0.58, outperforming BiLSTM-CRF with 0.55 and spaCy with 0.42, especially in a scenario with 21 entity categories.

Domain specialization proved to be a crucial factor for advancing results. Silveira *et al.* (2023) [Silveira et al. 2023] developed LegalBert-pt, a model pre-trained specifically on Brazilian legal documents. This model not only achieved a lower perplexity $(3,700)$ than generic models but also consistently outperformed BERTimbau-Base in NER tasks. The gains were up to 1.75% in macro F1-Score on LeNER-BR and 3.78% on the more challenging CDJUR-BR, confirming that specialized pre-training significantly improves the model's ability to understand the nuances of legal language.

Beyond model architecture and specialization, innovative training strategies have yielded promising results. Nunes *et al.* (2024) [Nunes et al. 2024] applied a self-learning and active sampling approach to the BERTimbau model. This technique raised the overall average F1-Score to $86.70 \pm 2.28\%$ on the UlyssesNER-BR corpus, a performance superior to all baselines, including BERTimbau itself without the strategy. The impact was notable in low-frequency categories, such as “Event”, where the F1-Score increased from 0% to 58.10%, demonstrating the method’s potential to overcome the scarcity of annotated data.

3. Proposal

Hybrid approaches for NER integrate multiple neural architectures within a single framework, leveraging the complementary strengths of each component. In this work, such combinations include transformer-based encoders, such as BERT or XLNet, paired with additional sequence modeling layers, for instance, BiLSTM or CRF, to capture long-range dependencies, enrich contextual representations, and enforce structural consistency in output tags [Souza et al. 2020b].

While these hybrid models have achieved state-of-the-art results on general-domain datasets, their systematic evaluation in the specific context of Brazilian legal texts is a significant research gap [Souza et al. 2020b]. The foundational papers for the main public datasets LeNER-Br, CDJur-Br, and UlyssesNER-BR focused on their creation and presented only baseline results, without a comprehensive comparison of different architectures.

This work aims precisely to fill this gap. Our main contribution is not the creation of a new dataset, but rather the systematic and in-depth evaluation of various models, including several hybrid configurations, applied rigorously to the three main publicly available Portuguese legal datasets. The experimental setup, summarized in Figure 1, is designed to explore multiple combinations of classical-based and transformer-based approaches, aiming to set a new performance benchmark for LER.

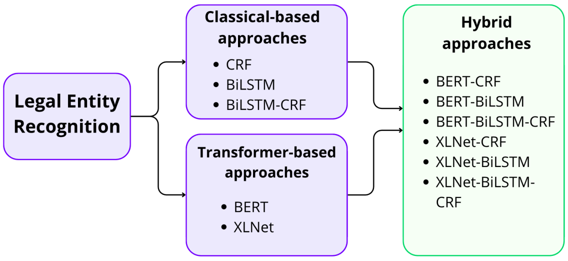

Figure 1. Explored approaches for LER

The chosen models were selected based on their effectiveness in previous LER research involving Portuguese language data. Classical-based approaches offer strong capabilities for capturing sequential patterns, while transformer-based approaches provide rich contextual representations. The selection focused on techniques already validated in related studies, ensuring the reproducibility and applicability of the results in the legal domain. Similarly, the datasets were selected due to their public availability and diverse content, encompassing different legal text types, thus enhancing the models’ generalization potential across legal subdomains [Nunes et al. 2024].

The evaluation strategy was designed to address the challenges of LER in Portuguese, notably entity imbalance and domain variability. Precision, Recall, and F1-Score were chosen for their relevance in NER tasks and their complementary insights into accuracy, coverage, and overall performance. To address class imbalance, metrics were averaged across entity types. Stratified k-fold cross-validation ensured robust and representative results, while the Friedman test, followed by Nemenyi post-hoc analysis, enabled statistically grounded model comparisons. Together, these choices aimed to ensure fair, reliable, and interpretable evaluations aligned with the study's objectives.

4. Methodology

The methodology in this study is designed to move beyond traditional systematic reviews, focusing instead on an expanded, empirical comparison of LER models. Unlike reviews that often confine themselves to cataloging existing work, our approach executes a full experimental cycle, from data preprocessing to the rigorous evaluation of 11 distinct models. As illustrated in Figure 2, the workflow begins with the selection of three key datasets from the Brazilian legal domain. Subsequently, we preprocess the data and apply classical-based, transformer-based, and hybrid approaches, including configurations not yet systematically evaluated in this context. Finally, the results are analyzed using standard metrics (Precision, Recall, and F1-Score) and statistical significance tests, enabling a direct and robust comparison of model performance.

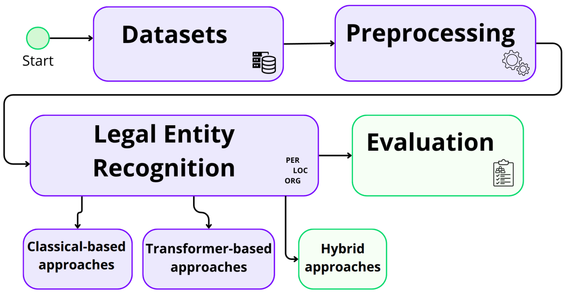

Figure 2. Methodology diagram outlining the main steps of the study.

4.1. Datasets

For this study, three annotated datasets recognized for their representativeness in the Brazilian legal domain were selected: LeNER-Br [Luz De Araujo et al. 2018], UlyssesNER-Br [Albuquerque et al. 2022], and CDJur-Br [Brito et al. 2023]. Together, these datasets provide a comprehensive foundation for LER experimentation by exploring different facets of legal language and documentation. They are, to date, the only open datasets available in the literature specifically targeting LER in the Brazilian legal context.

LeNER-Br [Luz De Araujo et al. 2018] is composed of 70 documents, including judicial decisions and legislative texts. With 318,073 tokens, its annotations cover six entity classes: "person", "organization", "location", "time", "legislation", and "jurisprudence", as detailed in Table 1. Its legal-specific focus makes it essential for applied NLP tasks.

UlyssesNER-Br [Albuquerque et al. 2022], in turn, focuses on the legislative domain, comprising 154 bills and 800 legislative consultations. Its seven entity classes: "person", "organization", "location", "event", "date", "fundament", and "legalproduct", are presented in Table 2, complementing the judicial scope of LeNER-Br.

Expanding the scope to the judicial workflow, the CDJur-Br [Brito et al. 2023] dataset was built from 1,216 legal documents from the Court of Justice of Ceará (TJCE).

It is distinguished by its 21 fine-grained entities, organized into six thematic categories, as shown in Table 3. This diversity in annotations and data sources supports a robust model evaluation.

Table 1. Entity distribution in LeNER-Br.

Entities	#Annotations	%
jurisprudence	5,370	12.06
legislation	18,317	41.15
location	1,793	4.03
organization	9,646	21.67
person	6,241	14.02
time	3,146	7.07
Total	**44,513**	**100**

Table 2. Entity distribution in UlyssesNER-BR.

Entities	#Annotations	%
date	1,000	6.89
event	10	0.07
fundament	6,700	46.18
location	1,500	10.33
organization	2,000	13.80
person	1,600	11.02
legalprod	1,700	11.71
Total	**14,510**	**100**

Table 3. Entity distribution in CDJUR-BR.

Category	#Annotations	%	Entities	#Annotations	%
Person	24,844	55.80	lawyer	735	1.65
			plaintiff	1,259	2.83
			police-identifier	2,012	4.52
			judge	576	1.29
			other	6,003	13.48
			prosecutor	363	0.82
			defendant	8,773	19.70
			witness	2,967	6.66
			victim	2,156	4.84
Evidence	3,318	7.45	evidence	3,318	7.45
Penalty	407	0.91	penalty	407	0.91
Address	2,065	4.64	plaintiff	132	0.30
			crime	466	1.05
			other	355	0.80
			defendant	693	1.56
			witness	295	0.66
			victim	124	0.28
Sentence	172	0.39	sentence	172	0.39
Norm	13,720	30.81	secondary	5,767	12.95
			jurisprudence	1,823	4.09
			main	6,130	13.77
Total	**44,526**	**100**	**Total**	**44,526**	**100**

4.2. Preprocessing

To standardize the data before training, the raw text from the datasets underwent a preprocessing pipeline. Tokenization, which is crucial for handling complex legal terms (e.g., "desembargadora" → "des", "##embar", "##gadora"), was performed using WordPiece for BERT and SentencePiece for XLNet [Schuster and Nakajima 2012, Kudo and Richardson 2018]. The process included text normalization, such as lowercasing and removing non-informative punctuation. Finally, tokens were labeled using the IOB scheme [Ramshaw and Marcus 1995] and converted into input IDs and attention masks, preparing the data for transformer-based models [Vaswani et al. 2017].

4.3. Legal Entity Recognition

This study evaluated 11 LER models, covering classical-based and transformer-based approaches, including their hybrid configurations. Classical-based approaches included: a CRF model using linguistic features and L2 regularization ($\lambda = 0.1$) [Lafferty et al. 2001]; a BiLSTM with 256 hidden units and the Adam optimizer ($lr = 0.001$); and a BiLSTM-CRF model, which combines the sequence modeling of BiLSTM with the structured prediction of CRF, using a dropout rate of 0.5 [Lample et al. 2016, Huang et al. 2015].

Transformer-based approaches utilized the Portuguese variant BERTimbau [Souza et al. 2020a] and XLNet [Yang et al. 2019], both evaluated in four configurations: base, CRF, BiLSTM, and BiLSTM-CRF. All were fine-tuned for 15 epochs with a batch size of 16. For BERTimbau, learning rates ranged from 2×10^{-5} to 5×10^{-5}, while XLNet used the AdamW optimizer with rates between 2×10^{-5} and 3×10^{-5}. In the hybrid architectures, the BiLSTM was configured with 256 hidden units and a dropout of 0.3.

4.4. Evaluation

Model performance was assessed using a multifaceted approach. Standard Precision, Recall, and F1-Score metrics were computed at the entity level, both in aggregate and per class, to ensure a balanced perspective. To determine the statistical significance of performance differences, we applied non-parametric Friedman and Nemenyi tests at a significance level of $\alpha = 0.05$ [Demšar 2006]. Additionally, we employed stratified 5-fold cross-validation to mitigate bias from data splits.

The experiments were conducted in Python using the `transformers`, `seqeval`, `PyTorch`, and `pytorch-crf` libraries. The computational environment consisted of a Linux system with two NVIDIA Tesla V100 32GB GPUs, 192 GB of RAM, and an Intel Xeon Gold 5220R processor. To ensure reproducibility, fixed random seeds were applied across all processes. The complete source code is publicly available on GitHub[1].

5. Results

This section presents and discusses the results obtained with the LER models applied to the LeNER-Br, UlyssesNER, and CDJUR-BR datasets. The analysis is divided into two parts: first, the overall performance of the models is compared across datasets, along with the application of statistical tests to assess the significance of observed differences; then, a detailed analysis is conducted by named entity type, highlighting which approaches perform better for specific legal categories.

5.1. General Results and Statistical Significance Analysis

The examined models were separated into two groups: (i) classical-based approaches (CRF, BiLSTM, and BiLSTM-CRF), and (ii) transformer-based and hybrid approaches (BERT, XLNet, and their combinations with BiLSTM and CRF). Table 4 summarizes the average F1-Scores (%) across the three datasets, highlighting variations in performance and model consistency through standard deviations.

In the LeNER-Br dataset, the best performance was achieved by the hybrid approaches BERT-BiLSTM and BERT-BiLSTM-CRF, both reaching an F1-Score of 0.9440 with a low standard deviation of 0.0049, indicating strong consistency across folds. The BERT model also yielded high performance, with an F1-Score of 0.9420 ± 0.0075, though with slightly higher variation. Among classical-based approaches, the CRF achieved the highest F1-Score of 0.8900 ± 0.0063, while BiLSTM and BiLSTM-CRF showed comparatively lower performance. XLNet-based models yielded less effective results overall, with lower F1-Score and limited improvements from hybrid approaches.

[1] `https://github.com/euhurias/Master_experiments`

Table 4. Average F1-Score (%) by model and dataset

Model	LeNER-Br	UlyssesNER	CDJUR-BR
CRF	0.8900 ± 0.0063	0.6806 ± 0.0159	0.8460 ± 0.0049
BiLSTM	0.7340 ± 0.0224	0.6060 ± 0.0049	0.4780 ± 0.0147
BiLSTM-CRF	0.8300 ± 0.0126	0.5820 ± 0.0194	0.7320 ± 0.0075
BERT	0.9420 ± 0.0075	0.8000 ± 0.0261	0.8660 ± 0.0049
XLNet	0.9280 ± 0.0040	0.6520 ± 0.0271	0.8660 ± 0.0049
BERT-CRF	0.9360 ± 0.0049	**0.8020 ± 0.0204**	0.8400 ± 0.0000
BERT-BiLSTM	**0.9440 ± 0.0049**	0.4300 ± 0.2161	**0.8800 ± 0.0000**
BERT-BiLSTM-CRF	**0.9440 ± 0.0049**	0.7680 ± 0.0172	0.8520 ± 0.0040
XLNet-CRF	0.8200 ± 0.0063	0.6620 ± 0.0366	0.8680 ± 0.0040
XLNet-BiLSTM	0.9280 ± 0.0075	0.4120 ± 0.2218	0.8640 ± 0.0049
XLNet-BiLSTM-CRF	0.9280 ± 0.0075	0.5840 ± 0.0625	0.8700 ± 0.0000

Given these results, the two top-performing hybrid approaches, BERT-BiLSTM and BERT-BiLSTM-CRF, were selected for further analysis, as they achieved identical scores. To verify whether these combinations outperform their respective base models, we conducted a comparative analysis including BERT, BiLSTM-CRF, and BiLSTM. Since BiLSTM-CRF consistently outperformed BiLSTM, it was adopted as the representative classical-based approach for comparison. The Friedman & Nemenyi statistical test was then applied, based on average rankings, to assess whether the observed differences are statistically significant.

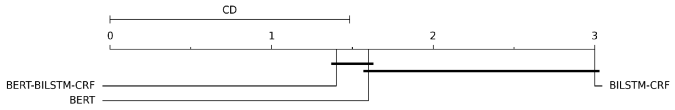

Figure 3. Critical Difference Diagram (Nemenyi test) for the LeNER-Br dataset.

To assess the statistical significance of the performance differences, the Friedman test was first applied, followed by the post-hoc Nemenyi test. As shown in Figure 3, the Friedman test revealed significant differences among BERT-BiLSTM-CRF, BERT, and BiLSTM-CRF ($\chi^2 = 8.40$, $p = 0.0147$), with a critical difference (CD) of 1.4823. BERT-BiLSTM-CRF achieved the best mean rank (1.4), followed by BERT (1.6), while BiLSTM-CRF had the lowest (3.0). The post-hoc analysis confirmed a statistically significant difference between BERT-BiLSTM-CRF and BiLSTM-CRF ($p = 0.0307$), highlighting the performance gain obtained by combining transformer-based with classical-based approaches such as BiLSTM and CRF.

As shown in Table 4, on the UlyssesNER dataset, the best performance was achieved by the hybrid approach BERT-CRF, with an F1-Score of 0.8020 ± 0.0204. BERT with 0.8000 ± 0.0261 and BERT-BiLSTM-CRF with 0.7680 ± 0.0172 followed closely, maintaining strong performance. In contrast, BERT-BiLSTM showed a notable drop with 0.4300 ± 0.2161, suggesting overfitting or instability in this dataset. XLNet-based models again underperformed. Among classical-based approaches, CRF led with 0.6806 ± 0.0159. To verify whether the top-performing model, BERT-CRF, significantly outperforms its base models, a comparative analysis including BERT and CRF was conducted using the Friedman & Nemenyi statistical test.

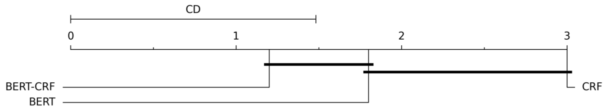

Figure 4. Critical Difference Diagram (Nemenyi test) for the UlyssesNER-BR dataset.

As shown in Figure 4, the Friedman test indicated significant differences among the top models with $\chi^2 = 8.40$, $p = 0.0150$, with a critical difference (CD) of 1.4823. BERT-CRF obtained the best mean rank (1.02), followed by BERT (1.8), while CRF ranked third (3.0). Post-hoc analysis confirmed a statistically significant difference between BERT-CRF and CRF with $p = 0.0123$, reinforcing the advantage of combining contextual embeddings with structured prediction layers.

As shown in Table 4, the best performance on the CDJUR-BR dataset was achieved by the hybrid approach BERT-BiLSTM, with an F1-score of 0.8800 ± 0.0000. BERT and XLNet also performed well. To determine whether BERT-BiLSTM significantly outperforms its base models, BERT and BiLSTM, a comparative analysis was conducted using the Friedman & Nemenyi statistical tests.

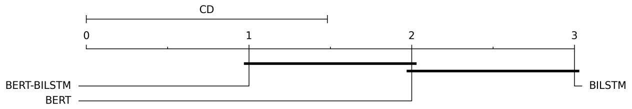

Figure 5. Critical Difference Diagram (Nemenyi test) for the CDJUR-BR dataset.

As shown in Figure 5, the Friedman test indicated significant differences among the evaluated models with $\chi^2 = 10.00$, $p = 0.0067$. BERT-BILSTM achieved the best mean rank (1.0), followed by BERT (2.0), while BILSTM ranked third (3.0). Post-hoc analysis confirmed a statistically significant difference between BERT-BILSTM and BILSTM with $p = 0.0045$, while no statistically significant difference was found between BERT-BILSTM and BERT ($p = 0.2538$), despite the higher average score of the hybrid.

5.2. Performance by Entity Type

For a detailed evaluation, Tables 5, 7, and 8 present the mean F1-score and standard deviation per entity across the datasets.

Table 5. Mean F1 Score (± standard deviation) per Entity for the LeNER-Br Dataset

Entity	BiLSTM-CRF	BERT	BERT-BiLSTM-CRF
jurisprudence	0.7580 ± 0.0504	**0.9100 ± 0.0110**	0.8880 ± 0.0204
legislation	0.8600 ± 0.0089	**0.9540 ± 0.0120**	0.9420 ± 0.0075
location	0.7080 ± 0.0791	0.9240 ± 0.0326	**0.9280 ± 0.0147**
organization	0.8520 ± 0.0147	0.9240 ± 0.0049	**0.9320 ± 0.0040**
person	0.8620 ± 0.0232	**0.9740 ± 0.0102**	0.9720 ± 0.0075
time	0.8140 ± 0.0258	0.9740 ± 0.0049	**0.9780 ± 0.0117**

The results show that hybrid and transformer-based approaches clearly outperform classical-based approaches across most entity types. BERT achieved the best scores for “jurisprudence” with 0.9100, “legislation” with 0.9540, and “person” with 0.9740, highlighting its strength in capturing semantic and syntactic patterns. BERT-BiLSTM-CRF excelled in "location" with an F1-Score of 0.9280, "organization" with 0.9320, and "time" with 0.9780, indicating that combining contextual embeddings with sequential modeling and label dependencies benefits entities with structural variability or temporal references. Additionally, hybrid models showed lower standard deviations, such as "person" with ± 0.0102 and "time" with ± 0.0049 with BERT, reflecting more stable predictions. This consistency is especially valuable in legal contexts. Overall, the per-entity evaluation confirms that hybrid approaches are not only superior on average but also more robust and reliable for legal entity recognition in the LeNER-Br dataset. To further investigate model behavior across specific entity types in the UlyssesNER dataset, Table 7 presents the average F1-Scores and standard deviations per entity, comparing CRF, BERT, and BERT-CRF.

Table 6. Mean F1 Score (± standard deviation) per Entity for the UlyssesNER Dataset

Entity	CRF	BERT	BERT-CRF
date	0.2320 ± 0.1993	0.6340 ± 0.0948	**0.6520 ± 0.1530**
event	0.6540 ± 0.1437	**0.6580 ± 0.1485**	0.6540 ± 0.1048
fundament	0.6540 ± 0.0524	**0.7500 ± 0.0434**	0.7400 ± 0.0385
location	0.6480 ± 0.1026	0.7840 ± 0.0641	**0.8380 ± 0.0371**
organization	0.6980 ± 0.0471	**0.8180 ± 0.0676**	0.8100 ± 0.0566
person	0.6620 ± 0.0299	0.8260 ± 0.0224	**0.8400 ± 0.0400**
legalprod	0.7380 ± 0.0312	**0.7840 ± 0.0162**	0.7740 ± 0.0206

Table 7. Mean F1 Score (± standard deviation) per Entity for the UlyssesNER Dataset

Entity	CRF	BERT	BERT-CRF
date	0.2320 ± 0.1993	0.6340 ± 0.0948	**0.6520 ± 0.1530**
event	0.6540 ± 0.1437	**0.6580 ± 0.1485**	0.6540 ± 0.1048
fundament	0.6540 ± 0.0524	**0.7500 ± 0.0434**	0.7400 ± 0.0385
location	0.6480 ± 0.1026	0.7840 ± 0.0641	**0.8380 ± 0.0371**
organization	0.6980 ± 0.0471	**0.8180 ± 0.0676**	0.8100 ± 0.0566
person	0.6620 ± 0.0299	0.8260 ± 0.0224	**0.8400 ± 0.0400**
legalprod	0.7380 ± 0.0312	**0.7840 ± 0.0162**	0.7740 ± 0.0206

The per-entity analysis confirms the superiority of BERT-based models over CRF in most categories. BERT-CRF achieved the highest F1-scores for "date", "location", and "person", highlighting its effectiveness in capturing temporal and sequential context. Meanwhile, BERT excelled in recognizing semantically rich entities such as "fundament", "organization", and "legalprod". The "event" category showed similar results across models, suggesting lower complexity for its identification. Notably, hybrid approaches like BERT-CRF also demonstrated lower standard deviations, indicating more stable and reliable predictions, an important factor for legal and literary tasks.

Table 8 presents the average F1 scores and standard deviations per entity in the CDJUR-BR dataset, comparing BILSTM, BERT, and BERT-BILSTM. The results obtained on the CDJUR-BR dataset clearly demonstrate the superiority of transformer-based and hybrid approaches, especially when combined with sequential modeling, as in BERT-BiLSTM. The BiLSTM model showed significantly lower performance in critical categories such as "address-victim" with 0.0000 ± 0.0000, "sentence" 0.0100 ± 0.0126, and "address-plaintiff" 0.0160 ± 0.0185, highlighting its limitations in capturing complex legal domain patterns. In contrast, BERT achieved substantial improvements, exceeding 0.9000 F1-Score in categories such as "person-police-identifier", "person-defendant",

Table 8. Mean F1 Score (± standard deviation) per Entity for the CDJUR-BR Dataset

Category	Entity	BiLSTM	BERT	BERT-BiLSTM
Person	lawyer	0.0600 ± 0.0167	0.5880 ± 0.0286	**0.6280 ± 0.0255**
	plaintiff	0.3740 ± 0.0422	0.7760 ± 0.0398	**0.7880 ± 0.0237**
	police-identifier	0.6960 ± 0.0361	0.9200 ± 0.0063	**0.9300 ± 0.0075**
	judge	0.1820 ± 0.0445	0.8020 ± 0.0279	**0.8260 ± 0.0248**
	other	0.4100 ± 0.0253	0.8220 ± 0.0075	**0.8320 ± 0.0089**
	prosecutor	0.1460 ± 0.0484	0.6880 ± 0.0279	**0.7200 ± 0.0193**
	defendant	0.6380 ± 0.0564	0.9200 ± 0.0126	**0.9240 ± 0.0089**
	witness	0.2260 ± 0.0689	0.8080 ± 0.0117	**0.8200 ± 0.0137**
	victim	0.5520 ± 0.0421	0.9000 ± 0.0219	**0.9060 ± 0.0150**
Evidence	evidence	0.1740 ± 0.0287	0.6460 ± 0.0273	**0.6720 ± 0.0237**
Penalty	penalty	0.4100 ± 0.0938	0.8540 ± 0.0969	**0.8680 ± 0.0545**
Address	plaintiff	0.0160 ± 0.0185	0.4880 ± 0.0880	**0.5560 ± 0.0699**
	crime	0.1900 ± 0.0482	0.7900 ± 0.0460	**0.8120 ± 0.0117**
	other	0.1620 ± 0.0492	0.7300 ± 0.0518	**0.7360 ± 0.0253**
	defendant	0.0900 ± 0.0690	0.7260 ± 0.0215	**0.7520 ± 0.0237**
	witness	0.0760 ± 0.0162	0.6680 ± 0.0643	**0.6840 ± 0.0591**
	victim	0.0000 ± 0.0000	0.5020 ± 0.0778	**0.5700 ± 0.0643**
Sentence	sentence	0.0100 ± 0.0126	0.5780 ± 0.1089	**0.6200 ± 0.0703**
Norm	secondary	0.6700 ± 0.0167	**0.8980 ± 0.0098**	0.8940 ± 0.0117
	jurisprudence	0.8040 ± 0.0441	**0.9820 ± 0.0040**	0.9800 ± 0.0036
	main	0.6820 ± 0.0440	**0.9020 ± 0.0040**	0.9000 ± 0.0024

"norm-jurisprudence", and "norm-main", due to its strong contextualization capabilities. BERT-BiLSTM achieved the best results across most entities, with notable gains in challenging categories such as "evidence" with 0.6720 ± 0.023, "penalty" with 0.8680 ± 0.0545, "address-crime" with 0.8120 ± 0.0117, "address-witness" with 0.6840 ± 0.0591, and "address-victim" with 0.5700 ± 0.0643.

Entities related to people, such as "person-plaintiff", "person-judge", "person-other", "person-witness", and "person-prosecutor", also showed consistent improvements with the use of BERT-BiLSTM, reflecting better modeling of sequential and contextual relationships. Additionally, the lower variance observed in models like BERT and BERT-BiLSTM, for example, "norm-jurisprudence" with ± 0.0040 and ± 0.0036, respectively, indicates greater stability and reliability, which are essential in the legal domain. These findings confirm that hybrid approaches such as BERT-BiLSTM not only outperform classical-based approaches but also refine performance compared to pure transformer-based approaches, particularly in entities with structural variability or unstable semantic roles.

Overall, the results reinforce the effectiveness of transformer-based approaches for LER tasks, with hybrid approaches often enhancing performance. While classical-based approaches like CRF showed solid results in certain datasets, they were generally outperformed by more advanced approaches. XLNet-based models demonstrated variable performance, suggesting that further tuning may be required for consistent gains across diverse legal datasets. Nonetheless, it is worth noting that hybrid approaches, despite their performance advantages, tend to demand greater computational resources, including longer training times and the need for specialized hardware such as high-end GPUs. These factors can limit their applicability in environments with constrained infrastructure and may increase the complexity of deploying such models.

6. Conclusion

We presented a novel and in-depth benchmark for LER in Portuguese by jointly evaluating a diverse set of models, ranging from classical-based to transformer-based and hybrid approaches, on three heterogeneous legal datasets. Unlike previous studies, which typically lack comprehensive comparisons across both models and datasets, our evaluation provides a more detailed view of model effectiveness across varied legal contexts.

Our experiments offer the first in-depth evaluation of LER models applied to the Brazilian legal domain, analyzing eleven classical-based, transformer-based, and hybrid approaches across the LeNER-Br, CDJUR-BR, and UlyssesNER datasets. The results consistently demonstrate the superiority of hybrid approaches, particularly those combining BERT with BiLSTM and CRF layers, over standalone approaches. These models not only achieved the highest F1-Scores across multiple datasets but also exhibited greater robustness and stability across entity types, especially those with structural variability or complex semantic patterns. The statistical analyses reinforce the significance of these improvements, confirming that hybrid approaches can effectively bridge the limitations of classical-based and transformer-based approaches in legal contexts.

In future work, we intend to explore alternative transformer-based approaches for LER, which enables a generative and multitask-oriented approach to sequence labeling. Additionally, we plan to refine the performance of XLNet by pretraining it on a large-scale Portuguese legal dataset, aiming to better align the model with the linguistic and structural nuances of the domain. These efforts will contribute to expanding the repertoire of effective models for LER in Portuguese, while also enabling a broader understanding of how different transformer-based paradigms influence LER outcomes.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., Da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., De Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Computational Processing of the Portuguese Language*. Springer.

Brito, M., Pinheiro, V., Furtado, V., Neto, J. A. M., Bomfim, F. d. C. J., da Costa, A. C. F., Silveira, R., and Aragão, N. (2023). CDJUR-BR – A Golden Collection of Legal Document from Brazilian Justice with Fine-Grained Named Entities. *arXiv*.

Costa, R., Albuquerque, H. O., Silvestre, G., Silva, N. F. F., Souza, E., Vitório, D., Nunes, A., Siqueira, F., Pedro Tarrega, J., Vitor Beinotti, J., de Souza Dias, M., Pereira, F. S. F., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text. In *Progress in Artificial Intelligence*. Springer.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Machine Learning Research*.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Human Language Technologies*. Association for Computational Lingustics.

Li, J., Sun, A., Han, J., and Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *arXiv*.

Luz De Araujo, P. H., De Campos, T. E., De Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*. Springer International Publishing.

Nunes, R. O., Balreira, D. G., Spritzer, A. S., and Freitas, C. M. D. S. (2024). A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Association for Computational Lingustics.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. *arXiv*.

Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain. In *Intelligent Systems*. Springer Nature Switzerland.

Siqueira, D. P., Mendes Junior, F., and Santos, M. F. D. (2023). Poder judiciário na era digital: o impacto das novas tecnologias de informação e de comunicação no exercício da jurisdição. *Consinter de Direito*.

Souza, F., Nogueira, R., and Lotufo, R. (2020a). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*. Springer International Publishing.

Souza, F., Nogueira, R., and Lotufo, R. (2020b). Portuguese named entity recognition using bert-crf. *arXiv*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.

Yadav, V. and Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv*.