RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese

Eduardo Garcia[♣*], **Nadia Silva**[♣*], **Felipe Siqueira**[◇], **Juliana Gomes**[♣], **Hidelberg O. Albuquerque**[♡♠], **Ellen Souza**[♡], **Eliomar Lima**[♣], **André de Carvalho**[◇]

[♣] Institute of Informatics, Federal University of Goiás

[◇] Institute of Mathematics and Computer Science, University of São Paulo

[♡] Mining Research Group, Federal Rural University of Pernambuco

[♠] Centro de Informática, Federal University of Pernambuco

edusantosgarcia@gmail.com, nadia.felix@ufg.br

Abstract

This work investigates the application of Natural Language Processing (NLP) in the legal context for the Portuguese language, emphasizing the importance of adapting pre-trained models, such as RoBERTa, from specialized corpora in the legal domain. We compiled and preprocessed a Portuguese Legal corpus, LegalPT corpus, addressing challenges of high document duplication in legal corpora, and measuring the impact of hyperparameters and embedding initialization. Experiments revealed that pre-training on legal and general data resulted in more effective models for legal tasks, with RoBERTaLexPT outperforming larger models trained on generic corpora, and other legal models from related works. We also aggregated a legal benchmark, PortuLex benchmark. This study contributes to improving NLP solutions in the Brazilian legal context, providing enhanced models, a specialized corpus, and a benchmark dataset. For reproducibility, we will make related code, data, and models available.

1 Introduction

Recent years have seen a significant focus on applying Natural Language Processing (NLP) techniques in the legal field. This growing interest is driven by advances in specialized NLP methods that can effectively handle the inherent complexities of legal language (Zhong et al., 2020). Legal practitioners and researchers deal with a substantial volume of legal texts on a daily basis, including legislation, jurisprudence, contracts, and petitions, all of which are characterized by highly technical and specialized language. In response to these challenges, the use of pre-trained language models[1], such as BERT (Devlin et al., 2019), adapted to meet the specific requirements of legal tasks, has emerged as a promising approach.

Pre-trained language models, like BERT, have demonstrated success in various NLP tasks (Devlin et al., 2019; Souza et al., 2020; Costa et al., 2022), and research studies have shown that their performance can be substantially improved when they are pre-trained on domain-specific corpora, such as legal (Chalkidis et al., 2020), biomedical (Lee et al., 2020), or scientific texts (Beltagy et al., 2019). This process, known as *domain adaptation*, has gained prominence and has led to improved performance in tasks within these specialized domains. It is worth noting that much of the previous work in domain adaptation for language models has been limited to the exploration into the impact of data selection on basic deduplication techniques (Lee et al., 2022; Tirumala et al., 2023), due to the universality of compute and data scaling laws which give practitioners a low-risk way to reliably improve language model performance by merely adding "more" data, not necessarily "new" data.

In the context of the Portuguese language, recent works have shown promise by training legal language models specifically tailored to Portuguese legal texts (Polo et al., 2021; Viegas et al., 2022). However, these studies have primarily focused on individual legal NLP tasks, making it challenging to assess the true benefits of domain adaptation for these models and to make meaningful comparisons among them.

In light of these, our research seeks to address these gaps, particularly within the Portuguese language legal context. Thus, our contributions are as follows: *(i)* Compiling the LegalPT Corpus[2], a Portuguese legal corpus by aggregating diverse sources of up to 125GiB data, which has shown significant performance improvement through dedupli-

*Corresponding author

[1]For the purposes of this paper, we will refer to both Causal Language Models and Masked Language Models as "language models", unless the distinction is made.

[2]The LegalPT Corpus is available at https://github.com/eduagarcia/roberta-legal-portuguese.

cation. *(ii)* Introducing the PortuLex benchmark, a Portuguese Legal benchmark composed of Named Entity Recognition (NER) and classification tasks. *(iii)* Developing RoBERTaLexPT[3] by pre-training a RoBERTa (Liu et al., 2019) base architecture on LegalPT and CrawlPT, outperforming prior Portuguese legal models, even much larger models.

This paper is structured as follows. Section 2 provides an overview of related works related to legal pre-trained models and techniques for corpus deduplication. In Section 3, we introduce the corpora employed in our pre-trained data and present the PortuLex benchmark, comparing in terms of deduplication rates. Section 4 presents the method used for pretraining and fine-tuning. Section 5 comprises the discussions and concludes the work, summarizing the findings, advantages, limitations, contributions, and research opportunities.

2 Related Works

The acquisition of a massive amount of new data is essential to achieve optimal performance in language models. As a general rule, the more documents one can obtain, the better the models will perform in NLP tasks (Kaplan et al., 2020a).

Empirical studies have consistently demonstrated that the adaptation of Transformer encoder models, such as BERT, to domain-specific corpora (Chalkidis et al., 2020; Lee et al., 2020; Beltagy et al., 2019) can result in substantial performance improvements.

By pre-training from local legal texts, a model can learn country-specific legal capabilities (Paul et al., 2023). Works in languages such as Chinese (Xiao et al., 2021), Italian (Licari and Comandè, 2022), Romanian (Masala et al., 2021), Spanish (Gutiérrez-Fandiño et al., 2021), Arabic (AL-Qurishi et al., 2022) and French (Douka et al., 2022) revealed that legal models outperform general-domain counterparts by about 1-5%, particularly when their training data is closely aligned.

Legal language models in the Portuguese, such as BERTikal (Viegas et al., 2022) and JurisBERT (Viegas et al., 2022), have reported superior performance in a specific legal task compared to BERTimbau (Souza et al., 2020), a generic Portuguese language model. However, in another study by Niklaus et al., 2023, training was conducted on both multilingual and multiple monolingual legal models, including Portuguese, with a substantial amount of data. Despite this, the Portuguese monolingual model failed to surpass BERTimbau's performance in multiple legal tasks.

It's common practice for extensive text corpora, such as MC4 (Xue et al., 2021), CC100 (Conneau et al., 2020), and brWaC (Wagner et al., 2018), to employ techniques that remove duplicate documents. This process aims to augment data quality and prevent unintended biases during machine learning model training. However, among the sets of the Portuguese legal corpus examined in this study (Niklaus et al., 2023; Willian Sousa and Fabro, 2019; Bonifacio et al., 2020), none indicate the use of deduplication algorithms.

The work by Lee et al. (2022) demonstrates that deduplicated datasets tend to improve the performance of causal language models. Models trained on datasets with duplication tendencies may memorize the data, potentially leading to contamination between training and validation splits. We hypothesize that this performance difference can be observed in masked language models as well.

Our work is similar to Chalkidis et al. (2020); Lee et al. (2020); Beltagy et al. (2019) in pretraining BERT models for the domain. We mainly follow the model training guidelines from Liu et al. (2019), apply text deduplication as described in Lee et al. (2022), and focus on the Brazilian and European Portuguese languages. By combining contributions from each of these works, we aim to fill the gaps in state-of-the-art Portuguese models adapted to the legal domain. To the best of our knowledge, our work is also the first to propose a benchmark adapted to this domain.

3 Corpora

This work aims to acquire as much publicly available data as possible within the legal domain for the Portuguese language. We compile two main corpora for pre-training: LegalPT, a legal domain-specific corpus, and CrawlPT, a general corpus used for comparison. Additionally, we have created the PortuLex benchmark, composed of a set of legal supervised tasks designed to evaluate the language models. In table 1, we summarize the details of the corpora used in this study.

3.1 LegalPT corpus

The following legal texts are publicly available and have been aggregated to create the corpus for pre-

[3]The RoBERTaLexPT Model is available at https://github.com/eduagarcia/roberta-legal-portuguese.

Corpus	Domain	Tokens (B)	Size (GiB)
LegalPT	Legal	22.5	125.1
brWaC	General	2.7	16.3
CC100 (PT)	General	8.4	49.1
OSCAR-2301 (PT)	General	18.1	97.8

Table 1: Corpora sizes in terms of billions of tokens and file size in GiB. CrawlPT composed by brWaC and the Portuguese (PT) subsets of CC100 and OSCAR-2301.

training language models in this work, which we refer to as the "LegalPT Corpus".

MultiLegalPile (Niklaus et al., 2023) is a multilingual corpus of legal texts comprising 689 GiB of data, covering 24 languages in 17 jurisdictions. The corpus is separated by language, and the subset in Portuguese contains 92GiB of data, containing 13.76 billion words. This subset includes the jurisprudence of the Court of Justice of São Paulo (CJPG), appeals from the 5th Regional Federal Court (Menezes-Neto and Clementino, 2022) (BRCAD-5), the Portuguese subset of legal documents from the European Union, known as EUR-Lex[4], and a filter for legal documents from MC4 (Xue et al., 2021).

Ulysses-Tesemõ[5] is a legal corpus in Brazilian Portuguese, composed of 2.2 million documents, totaling about 26GiB of text obtained from 96 different data sources. These sources encompass legal, legislative, academic papers, news, and related comments. The data was collected through web scraping of government websites.

ParlamentoPT is a corpus introduced by Rodrigues et al. (2023) for training language models in European Portuguese. The data was collected from the Portuguese government portal and consists of 2.6 million documents of transcriptions of debates in the Portuguese Parliament.

Iudicium Textum (Willian Sousa and Fabro, 2019) consists of rulings, votes, and reports from the Supreme Federal Court (STF) of Brazil, published between 2010 and 2018. The dataset contains 1GiB of data extracted from PDFs.

Acordãos TCU (Bonifacio et al., 2020)[6] is an open dataset from the Tribunal de Contas da União (Brazilian Federal Court of Accounts), containing 600,000 documents obtained by web scraping government websites. The documents span from 1992 to 2019.

DataSTF[7] is a dataset of monocratic decisions from the Superior Court of Justice (STJ) in Brazil, containing 700,000 documents (5GiB of data).

3.2 CrawlPT corpus

In order to compare the impact of deduplication and data size with other general Portuguese language models, we also applied the same process to the following Portuguese general corpora:

brWaC (Wagner et al., 2018) is a web corpus for Brazilian Portuguese from 120,000 different websites.

CC100 (Conneau et al., 2020) is a corpus created for training the multilingual Transformer XLM-R. The corpus contains two terabytes of cleaned data from the January to December of 2018 snapshots of the Common Crawl project [8] in 100 languages. We use the Portuguese subset from CC-100, which contains 49.1 GiB of text.

OSCAR-2301 (Abadji et al., 2022) is a multilingual corpus extracted from the November/December 2022 dump of Common Crawl. We use the Portuguese subset from OSCAR-2301, which contains 97.8 GiB of text.

We refer to the resulting dataset from these three corpora as "CrawlPT," a generic Portuguese corpus extracted from various web pages.

3.3 PortuLex benchmark

Our research focuses on acquiring open supervised training data meticulously annotated by legal experts. To maintain high benchmark quality, we deliberately avoided automatically generated datasets. In light of these efforts, we introduce the "PortuLex" benchmark, a four-task benchmark designed to evaluate the quality and performance of language models in the Portuguese legal domain. The composition of PortuLex is shown in Table 2.

Dataset	**Task**	**Train**	**Dev**	**Test**
RRI	CLS	8.26k	1.05k	1.47k
LeNER-Br	NER	7.83k	1.18k	1,39k
UlyssesNER-Br	NER	3.28k	489	524
FGV-STF	NER	415	60	119

Table 2: PortuLex benchmark – CLS refers to sentence classification tasks and NER to tokens sequence classification tasks.

[4] https://eur-lex.europa.eu/homepage.html

[5] https:github.com/ulysses-camara/ulysses-tesemo

[6] https://www.kaggle.com/datasets/ferraz/acordaos-tcu

[7] https://legalhackersnatal.wordpress.com/2019/05/09/mais-dados-juridicos/

[8] https://commoncrawl.org/about/

Corpus	Documents	Docs. after deduplication	Duplicates (%)
Ulysses-Tesemõ	2,216,656	1,737,720	21.61
MultiLegalPile (PT)			
CJPG	14,068,634	6,260,096	55.50
BRCAD-5	3,128,292	542,680	82.65
EUR-Lex (Caselaw)	104,312	78,893	24.37
EUR-Lex (Contracts)	11,581	8,511	26.51
EUR-Lex (Legislation)	232,556	95,024	59.14
Legal MC4	191,174	187,637	1.85
ParlamentoPT	2,670,846	2,109,931	21.00
Iudicium Textum	198,387	153,373	22.69
Acordãos TCU	634,711	462,031	27.21
DataSTF	737,769	310,119	57.97
Total (LegalPT)	**24,194,918**	**11,946,015**	**50.63**

Table 3: Duplicate rate found by the Minhash-LSH algorithm (Lee et al., 2022) for the LegalPT corpus.

Corpus	Documents	Docs. after deduplication	Duplicates (%)
brWaC	3,530,796	3,513,588	0.49
OSCAR-2301 (PT Subset)	18,031,400	10,888,966	39.61
CC100 (PT Subset)	38,999,388	38,059,979	2.41
Total (CrawlPT)	**60,561,584**	**52,462,533**	**13.37**

Table 4: Duplicate rate found by the Minhash-LSH algorithm for each subset composing the CrawlPT corpus.

LeNER-Br (Luz De Araujo et al., 2018) is the first Named Entity Recognition (NER) corpus for the legal domain in Brazilian Portuguese. It comprises 70 documents sourced from higher and state-level courts, annotated with six entity classes: organization, person, time, location, legislation, and jurisprudence.

Rhetorical Role Identification (RRI) (Aragy et al., 2021) is a dataset of rhetorical annotations within the legal domain, focusing on sentences extracted from judicial sentences from the Court of Justice of Mato Grosso do Sul (Brazil). It encompasses 70 initial petitions, containing approximately 10,000 manually labeled sentences. The dataset defines eight rhetorical roles in alignment with the Brazilian Civil Procedure Code, including the identification of parties, facts, arguments, legal foundation, jurisprudence, requests, case value, and "others".

FGV-STF (Correia et al., 2022) is a corpus of legal documents for entity extraction. This corpus is composed of 764 decisions from the Supreme Federal Court, manually selected by domain experts between 2009 and 2018. The data is annotated with varying levels of granularity, primarily focusing on legal foundation. These classes encompass precedents, academic citations, and legislative references, with each category containing more specific subtypes of entities. We use only the main four coarse-grained entities.

UlyssesNER-Br (Albuquerque et al., 2022) is a corpus of Brazilian legislative documents for NER. The corpus consists of bills and legislative queries from the Chamber of Deputies of Brazil. The dataset encompasses different granularity levels (Coarse/Fine), with 18 entity types manually annotated, and structured into 7 semantic classes.

4 Method

This section describes the method used in this work, including details on model architecture, the training process, datasets, and evaluation. The general training and evaluation method is summarized in Figure 1.

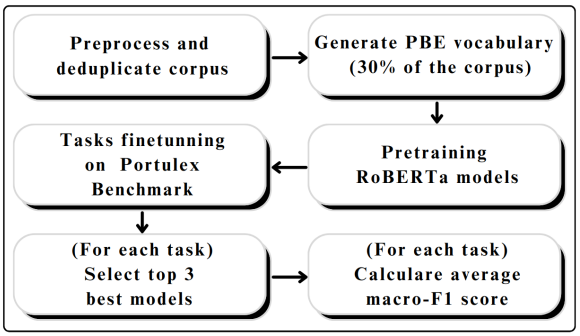

Figure 1: Method flowchart enumerating the necessary steps to pre-train a language model to evaluation on tasks of the PortuLex Benchmark.

Following the approach of Lee et al. (2022), we deduplicated all subsets of the LegalPT Corpus

using the MinHash algorithm (Broder, 2000) and Locality Sensitive Hashing (Har-Peled et al., 2012) to find clusters of duplicate documents. We used 5-grams and a signature of size 256, considering two documents to be identical if their Jaccard Similarity exceeded 0.7. The results of the deduplication process for the subsets of the LegalPT corpus can be found in Table 3 and for CrawlPT in Table 4.

To ensure that domain models are not constrained by a generic vocabulary, we utilized the HuggingFace Tokenizers[9] – BPE algorithm to train a vocabulary for each pre-training corpus used.

We employed a two-step validation methodology. First, to tune the hyperparameters of our models, we conducted a grid search by training on the training set and evaluating with the macro F1-score metric on the development set of the task data. The hyperparameters we tuned included learning rate and batch size.

After identifying the best-performing hyperparameters, we performed an evaluation using the top 3 checkpoints from the validation set and calculated the final metric as the arithmetic mean of the macro F1-Score over the dataset test splits. This method ensures that our models did not tend to overfit to the training set, thereby expecting them to perform well on unseen data.

4.1 Pretraining experiments

In this section, we describe the pretraining process of our legal language model using RoBERTa$_{base}$, a Transformer-based masked language model originally introduced by Liu et al. (2019). Our model was pretrained in four different configurations: solely on the LegalPT corpus (RoBERTaLegalPT$_{base}$), solely on the CrawlPT corpus (RoBERTaCrawlPT$_{base}$), by combining both corpora (RoBERTaLexPT$_{base}$), and solely on BrWaC (RoBERTaTimbau$_{base}$).

The pretraining process involved training the model for 62,500 steps, with a batch size of 2048 sequences, each containing a maximum of 512 tokens. This computational setup is similar to the work of BERTimbau (Souza et al., 2020), exposing the model to approximately 65 billion tokens during training.

We adopted the standard RoBERTa hyperparameters (Liu et al., 2019). During pretraining, we employed the masked language modeling objective, where 15% of the input tokens were randomly masked, and the model predicted these masked words based on contextual information. The optimization was performed using the AdamW optimizer with a linear warmup and a linear decay learning rate schedule. A detailed summary of the parameters used can be found in Table 5.

Hyperparameter	RoBERTa$_{base}$
Number of layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Dropout	0.1
Attention dropout	0.1
Warmup steps	6k
Peak learning rate	4e-4
Batch size	2048
Weight decay	0.01
Maximum training steps	62.5k
Learning rate decay	Linear
AdamW ϵ	1e-6
AdamW β_1	0.9
AdamW β_2	0.98
Gradient clipping	0.0

Table 5: Hyperparameters for pre-training RoBERTa.

Our pretraining process was executed using the Fairseq library (Ott et al., 2019) on a DGX-A100 cluster, utilizing a total of 2 Nvidia A100 80 GB GPUs. The complete training of a single configuration takes approximately three days.

4.2 Fine-tuning on the PortuLex benchmark

For the evaluation of our language models on the selected datasets within the PortuLex benchmark, we implemented the fine-tuning approach proposed by Devlin et al. (2019). This method trains a bidirectional Transformer encoder for both text classification and named entity recognition tasks. Table 6 presents the search space explored during the grid search process, detailing the constants that we retained.

5 Results and Discussion

This section presents our experiments with RoBERTa-based language models, particularly RoBERTaLexPT, pre-trained on a combined legal and generic corpus. We investigate the impact of hyperparameters on model performance using PortuLex benchmark scores in Section 5.1 and explore the benefits of merging diverse datasets in Section 5.2. Additionally, in Section 5.3, we provide a comprehensive analysis of RoBERTaLexPT against established Portuguese legal language models.

[9] https://github.com/huggingface/tokenizers

Hyperparameter	Search space
Batch size	{16, 32}
Learning rate	{7.5e-6, 1e-5, 2.5e-5, 5e-5}
Dropout of task layer	0.0
Warmup steps	100
Weight decay	0.01
Maximum training epochs	50
Learning rate scheduler	Constant
Optimizer	AdamW
AdamW ϵ	1e-8
AdamW β_1	0.9
AdamW β_2	0.999
Early stopping patience	750 steps
Early stopping threshold	0.001 (F1-score)

Table 6: Hyperparameter search space for fine-tuning models trained in the PortuLex benchmark.

5.1 Replicating BERTimbau with RoBERTa

The experiments in this section aim to investigate how various hyperparameters affect the model's performance compared to RoBERTa (Liu et al., 2019) with a larger batch size.

The BERTimbau model (Souza et al., 2020) is pre-trained with a maximum input sequence length ranging from 128 to 512, a vocabulary of 29,794 tokens trained on Wikipedia PT, a batch size of 128, and runs for 1 million steps or 8 epochs on the brWaC corpus, during which the model sees a total of 65 billion tokens. It initializes the training weights from the mBERT$_{base}$ and BERT$_{large}$ models, removing the initial embedding layer to accommodate the new Portuguese vocabulary.

We evaluated variations in learning rate, number of training epochs, and initialization. The models were based on the RoBERTa$_{base}$ architecture with a fixed tokenization length of 512 tokens and a BPE vocabulary of 50,265 tokens trained on Wikipedia PT. The checkpoints were evaluated on the PortuLex benchmark proposed by this work. The results are summarized in Table 7.

To maintain computational cost comparability with Souza et al., 2020, we set a limit of 65 billion training tokens. With the new BPE vocabulary, this corresponds to approximately 17 epochs for brWaC or 62,500 training steps with a batch size of 2048 and a tokenization length of 512 tokens. We also report the results for 8 epochs (equivalent to 30,000 training steps in our setup) as per Souza et al., 2020. We used the XLM-R$_{base}$ pre-trained model (Conneau et al., 2020) as initialization, discarding its embedding layer.

We found that using the initialization, the model can surpass BERTimbau on the PortuLex benchmark with only 30,000 training steps, achieving an average macro F1-Score of 84.01 versus 83.78. However, training longer or adjusting the learning rate does not seem to improve the model's performance. With random initialization, our RoBERTa model shows inferior performance to BERTimbau at the 8-epoch mark but surpasses the XLM-R$_{base}$ initialization when training for a longer period. At the 17-epoch mark, the model achieved an average macro F1-Score of 84.29 on the PortuLex benchmark.

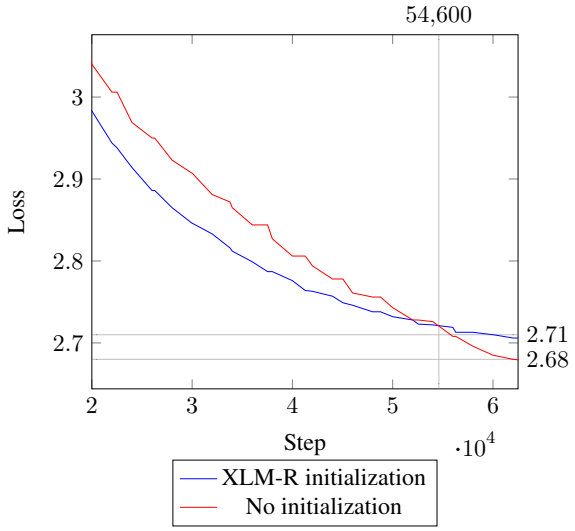

Figure 2: The MLM loss on the validation holdout set of models trained on brWaC with different initializations.

This behavior is also observed in the Masked Language Model loss graph on the validation subset in Figure 2. Between 54,000 and 55,000 training steps, the model without initialization outperforms the XLM-R initialization.

5.2 Combining generic and Legal Corpora

To our knowledge, domain adaptation techniques have not explored whether the combination of a domain-specific corpus with a generic corpus would enhance model performance due to the increased size of the pre-training corpus.

To evaluate the performance of this combination, we pre-trained language models on the CrawlPT corpus, as detailed in Section 3.2, and on the combination of CrawlPT with LegalPT. Models were trained with the hyperparameters defined in Section 5.1. The BPE vocabulary of each model was trained with 30% of the documents from their respective corpora.

Model	Batch size	Learning rate	Initialization	Steps	Epochs	PortuLex Score (%)
BERTimbau$_{base}$	128	1e-4	mBERT (no embeddings)	1,000,000	8	83.78
RoBERTaTimbau$_{base}$	2048	1e-4	XLM-R$_{base}$ (no embeddings)	30,000	8	84.01*
				62,500	17	83.96*
	2048	7e-4	XLM-R$_{base}$ (no embeddings)	30,000	8	83.40
				62,500	17	83.94*
Corpus: brWaC (16GiB)	2048	7e-4	Random	30,000	8	83.36
				62,500	17	**84.29***

Table 7: Macro F1-Score on the PortuLex benchmark for RoBERTa$_{base}$ models in Portuguese pre-trained on brWaC. Setup scores that outperformed BERTimbau$_{base}$ are marked with an asterisk, and the highest score is in bold font.

Model	LeNER	UlyNER-PL Coarse/Fine	FGV-STF Coarse	RRIP	Average (%)
BERTimbau$_{base}$ (Souza et al., 2020)	88.34	86.39/83.83	79.34	82.34	83.78
BERTimbau$_{large}$ (Souza et al., 2020)	88.64	87.77/84.74	79.71	**83.79**	84.60
Albertina-PT-BR$_{base}$ (Rodrigues et al., 2023)	89.26	86.35/84.63	79.30	81.16	83.80
Albertina-PT-BR$_{xlarge}$ (Rodrigues et al., 2023)	90.09	88.36/**86.62**	79.94	82.79	85.08
BERTikal$_{base}$ (Polo et al., 2021)	83.68	79.21/75.70	77.73	81.11	79.99
JurisBERT$_{base}$ (Viegas et al., 2022)	81.74	81.67/77.97	76.04	80.85	79.61
BERTimbauLAW$_{base}$ (Viegas et al., 2022)	84.90	87.11/84.42	79.78	82.35	83.20
Legal-XLM-R$_{base}$ (Niklaus et al., 2023)	87.48	83.49/83.16	79.79	82.35	83.24
Legal-XLM-R$_{large}$ (Niklaus et al., 2023)	88.39	84.65/84.55	79.36	81.66	83.50
Legal-RoBERTa-PT$_{large}$ (Niklaus et al., 2023)	87.96	88.32/84.83	79.57	81.98	84.02
RoBERTaTimbau$_{base}$	89.68	87.53/85.74	78.82	82.03	84.29
RoBERTaLegalPT$_{base}$	90.59	85.45/84.40	79.92	82.84	84.57
RoBERTaLexPT$_{base}$	**90.73**	**88.56**/86.03	**80.40**	83.22	**85.41**

Table 8: Macro F1-Score (%) for multiple models evaluated on PortuLex benchmark test splits.

Model	Corpus	Avg. F1
RoBERTaTimbau$_{base}$	brWaC	84.29
RoBERTaCrawlPT$_{base}$	CrawlPT	84.83
RoBERTaLegalPT$_{base}$	LegalPT	84.57
RoBERTaLexPT$_{base}$	LegalPT+CrawlPT	**85.41**

Table 9: Average macro F1-score on pretrained models on PortuLex benchmark. RoBERTaLexPT$_{base}$, pretrained on both domain-specific LegalPT corpus and general CrawlPT corpus, achieves the highest score.

We evaluated the new models on the PortuLex benchmark and compared them with RoBERTaTimbau$_{base}$. The results can be found in Table 9.

Interestingly, when used individually for pretraining, the CrawlPT model exhibits superior performance to LegalPT, despite CrawlPT's generic domain. Even with a similar size, the CrawlPT corpus has more unique data, with a 13.37% duplication rate compared to 50.63% for the LegalPT corpus. This indicates that a high-quality generic corpus can be comparable to a domain-specific corpus for pre-training language models.

However, upon combining the two corpora, the resulting model, RoBERTaLexPT, shows superior performance compared to that of models pretrained on individual datasets. This outcome aligns with the conclusions drawn by Kaplan et al. (2020b) that corpus size is a key factor in increasing model performance, although their study examined causal language models, which differs from the masked language models in our research.

5.3 Comparing with other Legal models

Table 8 presents the performance of RoBERTaLexPT compared to prior open Portuguese legal language models in the PortuLex benchmark datasets.

The primary finding is that despite using only a base configuration, RoBERTaLexPT outperforms even much larger models such as Albertina-PT-BR$_{xlarge}$, BERTimbau$_{large}$, and Legal-XLM-R$_{large}$. This highlights RoBERTaLexPT's effectiveness resulting from pre-training on combined legal and generic data.

Specifically, RoBERTaLexPT achieves the highest performance on the LeNER and FGV-STF datasets, even when compared to significantly larger models. For UlyssesNER-Br, RoBERTaLexPT attains competitive results with the top models. The only dataset where RoBERTaLexPT is surpassed is RRI, where BERTimbau$_{large}$ has a slight edge of 0.57% in F1-score.

In contrast, some prior works claimed superior performance over BERTimbau for certain legal tasks (Polo et al., 2021; Viegas et al., 2022). However, these models actually underperform BERTimbau in our PortuLex benchmark experiments. For instance, JurisBERT only reaches an average F1-score of 79.61% compared to BERTimbau's 83.78%. One possible explanation for this discrepancy is that the original evaluations were limited to a single selected dataset, likely favoring the model's specific training data.

In summary, RoBERTaLexPT consistently achieves top legal NLP effectiveness despite its base size. With sufficient pre-training data, it can surpass overparameterized models. The results highlight the importance of domain-diverse training data over sheer model scale.

6 Conclusion

This work introduces RoBERTaLexPT, a Portuguese legal language model pre-trained on a combined legal and generic corpus. Throughout this process, we created the largest Portuguese legal corpus (LegalPT) by aggregating diverse sources, resulting in significant performance improvements through deduplication and introducing the PortuLex benchmark for rigorous model evaluation.

We also demonstrated that using other models as weight initialization for pre-training language models can boost performance in a limited resource setting, but it has a trade-off if trained for longer training settings.

Our findings indicate that combining a domain-specific corpus (LegalPT) and a generic corpus (CrawlPT) for pre-training yields complementary benefits. Despite its compact size compared to prior models, the RoBERTaLexPT base model demonstrates state-of-the-art effectiveness in Portuguese legal NLP. This underscores the significance of pre-training data over model scale.

RoBERTaLexPT, LegalPT, and PortuLex significantly advance Portuguese legal NLP, addressing resource and model limitations. Future work can explore pre-training larger RoBERTa models, expanding the LegalPT corpus, and enhancing the PortuLex benchmark.

There remain opportunities for future work to build upon these contributions. Potential research directions include pre-training larger RoBERTa models, expanding the LegalPT corpus, and enhancing the PortuLex benchmark.

Acknowledgements

This work has been supported by the AI Center of Excellence (Centro de Excelência em Inteligência Artificial – CEIA) of the Institute of Informatics at the Federal University of Goiás (INF-UFG). Ellen Souza and Nadia Félix are supported by FAPESP, agreement between USP and the Brazilian Chamber of Deputies. To the CEIA, to the Institute of Artificial Intelligence (IAIA), and to research funding agencies, to which we express our gratitude for supporting the research.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. ArXiv:2201.06642 [cs].

Muhammad AL-Qurishi, Sarah AlQaseemi, and Riad Soussi. 2022. AraLegal-BERT: A pre-trained language model for Arabic Legal text. ArXiv:2210.08284 [cs].

Hidelberg O. Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 3–14, Cham. Springer International Publishing.

Roberto Aragy, Eraldo Rezende Fernandes, and Edson Norberto Caceres. 2021. Rhetorical Role Identification for Portuguese Legal Documents. In *Intelligent Systems*, Lecture Notes in Computer Science, pages 557–571, Cham. Springer International Publishing.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv:1903.10676 [cs].

Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems*, pages 648–662, Cham. Springer International Publishing.

Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. In *Combinatorial Pattern Matching*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. ArXiv:2010.02559 [cs].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Fernando A. Correia, Alexandre A. A. Almeida, José Luiz Nunes, Kaline G. Santos, Ivar A. Hartmann, Felipe A. Silva, and Hélio Lopes. 2022. Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. *Information Processing & Management*, 59(1):102794.

Rosimeire Costa, Hidelberg Oliveira Albuquerque, Gabriel Silvestre, Nádia Félix F. Silva, Ellen Souza, Douglas Vitório, Augusto Nunes, Felipe Siqueira, João Pedro Tarrega, João Vitor Beinotti, Márcio de Souza Dias, Fabíola S. F. Pereira, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text. In *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 767–779, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2022. JuriBERT: A Masked-Language Model Adaptation for French Legal Text. ArXiv:2110.01485 [cs].

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish Legalese Language Model and Corpora. ArXiv:2110.12201 [cs].

Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. 2012. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. *Theory of Computing*, 8(1):321–350.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020a. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs, stat].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020b. Scaling laws for neural language models.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. ArXiv:1901.08746 [cs].

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. ArXiv:2107.06499 [cs].

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, volume 3256 of *CEUR Workshop Proceedings*, Bozen-Bolzano, Italy. CEUR. ISSN: 1613-0073.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Pedro Henrique Luz De Araujo, Teófilo E. De Campos, Renato R. R. De Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, *Computational Processing of the Portuguese Language*, volume 11122, pages 313–323. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT Model for Legal Judgement Prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PLOS ONE*, 17(7):e0272287. Publisher: Public Library of Science.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023. MultiLegalPile: A 689GB Multilingual Legal Corpus.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: a case study on indian law. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 187–196.

Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J. Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. 2021. LegalNLP – Natural Language Processing methods for the Brazilian Legal Language. ArXiv:2110.15709 [cs].

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. ArXiv:2305.06721 [cs].

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Lecture Notes in Computer Science, pages 403–417, Cham. Springer International Publishing.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification.

Charles F O Viegas, Bruno Catais Costa, and Renato Porfirio Ishii. 2022. JurisBERT: Transformer-based model for embedding legal texts.

Jorge Wagner, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese.

Antonio Willian Sousa and Marcos Fabro. 2019. *Iudicium Textum Dataset Uma Base de Textos Jurídicos para NLP*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents. ArXiv:2105.03887 [cs].

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. ArXiv:2010.11934 [cs].

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.