Proposta recebida em Outubro 2024 e aceite para publicação em Dezembro 2024.

RoBERTaLexPT: um modelo RoBERTa jurídico pré-treinado com deduplicação para língua Portuguesa

RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese

Eduardo Garcia
Universidade Federal de Goiás

Nádia Silva
Universidade Federal de Goiás

Juliana Gomes
Universidade Federal de Goiás

Hidelberg Albuquerque
Universidade Federal Rural de Pernambuco

Ellen Souza
Universidade Federal Rural de Pernambuco

Felipe Siqueira
Universidade de São Paulo

Eliomar Lima
Universidade Federal de Goiás

André de Carvalho
Universidade de São Paulo

Resumo

Este trabalho investiga a aplicação do Processamento de Linguagem Natural (PLN) no contexto jurídico para a língua portuguesa, enfatizando a importância de adaptar modelos pré-treinados, como o RoBERTa, a partir de *corpora* especializados no domínio jurídico. Compilamos e pré-processamos um *corpus* jurídico em português, o *corpus* "LegalPT," abordando os desafios da alta duplicação de documentos em *corpora* jurídicos e medindo o impacto dos hiperparâmetros e da inicialização de *embeddings*. Experimentos revelaram que o pré-treinamento em dados jurídicos e em dados gerais resultou em modelos mais eficazes para tarefas jurídicas, com o nosso modelo, intitulado RoBERTaLexPT, superando modelos maiores treinados em *corpora* genéricos e outros modelos jurídicos de trabalhos relacionados. Também agregamos um *benchmark* jurídico, o *benchmark* "PortuLex." Este estudo contribui para melhorar as soluções de PLN no contexto jurídico brasileiro, fornecendo modelos aprimorados, um *corpus* especializado e um conjunto de dados de referência. Para fins de reprodutibilidade, disponibilizaremos o código, os dados e os modelos relacionados.

Palavras chave

modelo de linguagem; domínio jurídico; *benchmark*

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 pre-processed 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.

Keywords

language model; legal domain; benchmark

1. Introdução

Profissionais e pesquisadores da área jurídica enfrentam, diariamente, a necessidade de lidar com um volume substancial de textos legais que abrange legislação, jurisprudência, contratos e petições, com seus subdomínios específicos (e.g., processual, penal, trabalhista, previdenciário) e suas esferas de atuação (e.g., tipos de tribunais, jurisdição) (Akbik et al., 2016; Firdaus Solihin & Makarim, 2021; Cabrera-Diego & Gheewala, 2023). A natureza desses textos adiciona desafios significativos às aplicações de Processamento de Linguagem Natural (PLN), uma vez que textos legais são tipicamente, extensos, desestruturados, apresentam ruídos e são redigidos utilizando jargões e uma linguagem técnica específica do domínio (Kapoor et al., 2022).

O número de aplicações de PLN no campo jurídico tem crescido de maneira expressiva, impulsionado, sobretudo, pelos avanços em métodos especializados capazes de lidar com a complexidade inerente à linguagem jurídica (Zhong et al., 2020). Modelos de linguagem pré-treinados, como o *Bidirectional Encoder Representations from Transformers (BERT)* (Devlin

DOI: 10.21814/lm.16.2.457

et al., 2019), têm apresentado resultados significativos em várias tarefas de PLN voltadas para a língua portuguesa (Souza et al., 2020; Costa et al., 2022). Estudos indicam que o desempenho dos modelos pode ser aprimorado quando são pré-treinados em *corpora* específicos de um determinado domínio, como textos jurídicos (Chalkidis et al., 2020). Esse processo, conhecido como adaptação de domínio, permite que os modelos se adéquem melhor às características e ao vocabulário especializado de cada área.

No que diz respeito à adaptação de domínio, estudos anteriores têm se concentrado em investigar o impacto da seleção de dados por meio de técnicas básicas de deduplicação, ou seja, remoção de duplicatas (Lee et al., 2022; Tirumala et al., 2023). A deduplicação reduz o viés nos modelos, melhora a generalização e o desempenho, e mitiga o risco de *overfitting* ao evitar a memorização de padrões repetitivos, fortalecendo sua robustez (Brown et al., 2020). Essas técnicas oferecem aos pesquisadores uma abordagem de baixo risco para aprimorar o desempenho dos modelos de linguagem, mesmo ao simplesmente adicionar mais dados, ainda que não necessariamente inéditos. No contexto da língua portuguesa, estudos recentes têm apresentado resultados promissores ao desenvolver modelos específicos para textos legais em português (Polo et al., 2021; Viegas et al., 2023). Entretanto, tais pesquisas têm focado predominantemente em tarefas isoladas de PLN aplicadas ao domínio jurídico, o que limita a avaliação dos benefícios proporcionados pela adaptação de domínio nesses modelos e dificulta a realização de comparações entre eles.

Em Garcia et al. (2024), disponibilizamos (i) o *corpus* "LegalPT", que trata dos desafios relacionados à alta duplicação de documentos em *corpora* jurídicos, (ii) o "PortuLex", um *benchmark* jurídico em português composto por tarefas de NER e classificação, e (iii) o modelo "RoBERTaLexPT", treinado com a arquitetura RoBERTa (Liu et al., 2019) sobre os *corpora* LegalPT e "CrawlPT", superando o desempenho dos modelos jurídicos disponíveis para a língua portuguesa. Nesta versão estendida, as principais contribuições deste estudo incluem: (iv) a disponibilização de uma nova versão do RoBERTaLexPT, aprimorado com um pré-treinamento adicional utilizando textos específicos do domínio; (v) a análise do impacto de realizar *further pre-training*[1] em um modelo já existente em comparação com iniciar um pré-treinamento do zero; e (vi) a avaliação de modelos jurídicos e *benchmarks* genéricos, mensurando o impacto da transferência de aprendizado.

Este artigo está estruturado da seguinte forma: a Seção 2 fornece uma visão geral dos trabalhos relacionados com foco em modelos jurídicos pré-treinados e técnicas de deduplicação de *corpus*. Na Seção 3, são apresentados os *corpora* utilizados nos dados pré-treinados e o *benchmark* PortuLex, bem como uma análise comparativa das taxas de deduplicação. A Seção 4 apresenta o método utilizado para pré-treinamento e ajuste fino dos modelos. A Seção 5 apresenta e discute os resultados. A Seção 6 apresenta as contribuições, limitações e oportunidades de pesquisa. Por fim, o Apêndice A apresenta os melhores resultados da busca de hiperparâmetros utilizados no pré-treinamento dos modelos.

2. Trabalhos relacionados

A aquisição de uma quantidade massiva de novos dados é essencial para alcançar um desempenho ótimo em modelos de linguagem. Como regra geral, quanto maior o número de documentos utilizados, maior será a chance de melhoria do desempenho dos modelos em tarefas de PLN (Kaplan et al., 2020). Estudos empíricos têm demonstrado consistentemente que a adaptação de modelos baseados no modelo Transformer como o BERT, para *corpora* específicos de domínio (Chalkidis et al., 2020; Lee et al., 2019; Beltagy et al., 2019) pode resultar em melhorias substanciais de desempenho.

Através do pré-treinamento em textos jurídicos de uma legislação, um modelo pode aprender capacidades jurídicas específicas de cada país (Paul et al., 2023). Trabalhos em idiomas como chinês (Xiao et al., 2021), italiano (Licari & Comandè, 2022), romeno (Masala et al., 2021), espanhol (Gutiérrez-Fandiño et al., 2021), árabe (Al-qurishi et al., 2022) e francês (Douka et al., 2021), revelaram que os modelos jurídicos superam seus equivalentes de domínio geral em cerca de até 5%, particularmente quando seus dados de treinamento estão estreitamente alinhados.

Modelos de linguagem jurídica em português, como o "BERTikal" (Polo et al., 2021) e o "JurisBERT" (Viegas et al., 2023), relataram desempenho superiores em uma tarefa jurídica específica em comparação com o "BERTimbau" (Souza et al., 2020), um modelo de linguagem genérico em português. Por sua vez, na pesquisa de Niklaus et al. (2024), foi realizado treinamento

[1] *Further pre-training* ou Pré-treinamento continuado é uma técnica que faz um "pré-treinamento adicional" com dados específicos do domínio.

Corpus **Domínio**	**Recursos**	**Documentos (M)**	***Tokens* (B)**	**Tamanho (GiB)**
LegalPT / Jurídico	*MultiLegalPile (PT)* (Niklaus et al., 2024)	17,7	15,7	88,1
	Ulysses-Tesemô (Siqueira et al., 2024)	2,2	4,9	27
	ParlamentoPT (Rodrigues et al., 2023)	2,7	0,5	2,5
	Iudicium Textum (Willian Sousa & Fabro, 2019)	0,2	0,1	0,8
	Acórdãos TCU (Bonifacio et al., 2020)	0,6	0,6	3,2
	DataSTF (2019)[1]	0,7	0,6	3,4
	Total	**24,2**	**22,5**	**125,1**
CrawlPT / Geral	brWaC (Wagner Filho et al., 2018)	3,5	3,1	16,3
	CC100 (PT) (Conneau et al., 2020)	39	9,4	49,1
	OSCAR-2301 (PT) (Abadji et al., 2022)	18	18,1	97,8
	Total	**60,5**	**30,6**	**163,3**

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

Tabela 1: Tamanho dos *corpora* em termos de milhões de documentos, bilhões de *tokens* e tamanho de arquivo em GiB.

tanto em modelos jurídicos multilíngues quanto em modelos monolíngues, incluindo o português, com uma quantidade substancial de dados. Apesar disso, o modelo monolíngue em português não conseguiu superar o desempenho do BERTimbau em múltiplas tarefas jurídicas.

No desenvolvimento de *corpora* extensos, como "MC4" (Xue et al., 2021), "CC100" (Conneau et al., 2020) e "brWaC" (Wagner Filho et al., 2018), é comum o emprego de técnicas que removem documentos duplicados. Esse processo visa aumentar a qualidade dos dados e prevenir vieses indesejados durante o treinamento de modelos de aprendizado de máquina. No entanto, entre os *corpora* jurídicos em português examinados neste estudo (Willian Sousa & Fabro, 2019; Bonifacio et al., 2020; Niklaus et al., 2024), não foram encontradas informações sobre o uso de técnicas de deduplicação.

Lee et al. (2022) demonstram que conjuntos de dados deduplicados tendem a melhorar o desempenho de modelos de linguagem causal. Modelos treinados em conjuntos de dados com tendências de duplicação podem memorizar os dados, potencialmente levando à contaminação entre as divisões de treinamento e validação. Nossa hipótese é que essa diferença de desempenho também pode ser observada em modelos de linguagem mascarada.

Nossa pesquisa é semelhante à de Beltagy et al. (2019), Chalkidis et al. (2020), e Lee et al. (2019) quanto ao pré-treinamento de modelos BERT para o domínio. Seguimos principalmente as diretrizes de treinamento de Liu et al. (2019), aplicamos a deduplicação de texto conforme descrito em Lee et al. (2022), e nos concentramos obter *corpora* jurídicos nos idiomas português brasileiro e europeu. Ao combinar contribuições de cada um desses trabalhos, visamos preencher as lacunas nos modelos de português de última geração adaptados ao domínio jurídico. Até onde sabemos, nosso trabalho também é o primeiro a propor um *benchmark* adaptado a este domínio.

3. *Corpora*

Para esta pesquisa, buscamos adquirir o maior número possível de dados no domínio jurídico para a língua portuguesa. Desta forma, foram compilados dois *corpora* principais para o pré-treinamento: "LegalPT", um *corpus* específico para o domínio jurídico, e "CrawlPT", um *corpus* de uso geral usado para comparação. Foram utilizados unicamente recursos públicos disponíveis na internet, garantindo acessibilidade e transparência no processo de extração dos dados. Adicionalmente, criamos o *benchmark* "PortuLex", composto por um conjunto de tarefas supervisionadas jurídicas projetadas para avaliação dos modelos de linguagem desenvolvidos. A Tabela 1, resume os *corpora* utilizados neste estudo, detalhados a seguir.

3.1. *Corpus* LegalPT

Os textos jurídicos a seguir foram agregados para compor o *corpus* LegalPT, utilizado no pré-treinamento dos modelos desta pesquisa:

MultiLegalPile (Niklaus et al., 2024)[2] é um *corpus* multilíngue de textos jurídicos que compreende 689 GiB de dados, abrangendo

[2] https://huggingface.co/datasets/joelniklaus/Multi_Legal_Pile

Corpus	Documentos	Tarefa	Instâncias	Treinamento	Validação	Teste
RRI (Aragy et al., 2021)	70	CLS	10,78 mil	8,26 mil	1,05 mil	1,47 mil
LeNER-Br (Luz de Araujo et al., 2018)	70	NER	10,4 mil	7,83 mil	1,18 mil	1,39 mil
UlyssesNER-Br—PL-Corpus (Albuquerque et al., 2022)	154	NER	4,30 mil	3,28 mil	489	524
FGV-STF (Correia et al., 2022)	594	NER	594	415	60	119
Total	888	—	26,074 mil	19,785 mil	2,779 mil	3,503 mil

Tabela 2: *Benchmark* PortuLex usado nas tarefas de classificação de sentenças (CLS) e de *tokens* com NER.

24 idiomas em 17 jurisdições. O *corpus* é separado por idioma, e o subconjunto em português contém 88.1 GiB de dados e 15,7 bilhões de palavras. Este subconjunto inclui jurisprudências do Tribunal de Justiça de São Paulo, recursos do 5º Tribunal Regional Federal (Menezes-Neto & Clementino, 2022) (BRCAD-5), o subconjunto português de documentos jurídicos da União Europeia (EUR-Lex)[3], e um filtro para documentos jurídicos do MC4 (Xue et al., 2021).

Ulysses-Tesemõ (Siqueira et al., 2024)[4] é um *corpus* jurídico em português brasileiro, composto por 2,2 milhões de documentos, totalizando cerca de 27 GiB de texto obtidos de 96 fontes de dados diferentes. Essas fontes abrangem documentos jurídicos, legislativos, artigos acadêmicos, notícias e comentários relacionados. Os dados foram coletados por meio de *web scraping* de sites governamentais.

ParlamentoPT (Rodrigues et al., 2023)[5] é um *corpus* jurídico utilizado no treinamento de modelos de linguagem em português europeu. Os dados foram coletados do portal do governo português, e consiste em 2,7 milhões de documentos de transcrições de debates no Parlamento português.

Iudicium Textum (Willian Sousa & Fabro, 2019)[6] consiste em acórdãos, votos e relatórios do Supremo Tribunal Federal (STF) do Brasil, publicados entre 2010 e 2018. O conjunto de dados contém 0.8 GiB de dados extraídos de PDFs.

Acórdãos TCU (Bonifacio et al., 2020)[7] é um conjunto de dados aberto do Tribunal de Contas da União, contendo 600.000 documentos obtidos por *web scraping* de *sites* governamentais. Os documentos abrangem o período de 1992 a 2019.

DataSTF[8] é um conjunto de dados de decisões monocráticas do Superior Tribunal de Justiça do Brasil, contendo 700.000 documentos (3.4 GiB de dados).

3.2. *Corpus* CrawlPT

Para avaliar o impacto da deduplicação e do tamanho dos dados em comparação com outros modelos de linguagem gerais em português, também aplicamos o mesmo processo aos seguintes *corpora* gerais em português, para formação do *corpus* "CrawlPT":

brWaC (Wagner Filho et al., 2018)[9] é um *corpus* extenso em português brasileiro, contendo cerca de 3,5 milhões de documentos e 3,31 bilhões de tokens, extraídos de mais de 120 mil websites diferentes. Construído seguindo a metodologia *WaCky*, o *corpus* utilizou técnicas de deduplicação, garantindo alta diversidade e qualidade dos dados.

CC100 (Conneau et al., 2020)[10] é um *corpus* criado para treinar o modelo Transformer multilíngue XLM-R. O *corpus* contém 2 TB de dados limpos das capturas realizadas no período de janeiro a dezembro de 2018 do projeto *Common Crawl* em 100 idiomas. Para esta pesquisa, foi utilizado o subconjunto em português do CC-100, que contém

[3] https://eur-lex.europa.eu/homepage.html
[4] https:github.com/ulysses-camara/ulysses-tesemo
[5] https://huggingface.co/datasets/PORTULAN/parlamento-pt
[6] https://dadosabertos.c3sl.ufpr.br/acordaos/
[7] https://www.kaggle.com/datasets/ferraz/acordaos-tcu
[9] https://www.inf.ufrgs.br/pln/wiki/index.php?title=BrWaC
[10] https://github.com/facebookresearch/cc_net

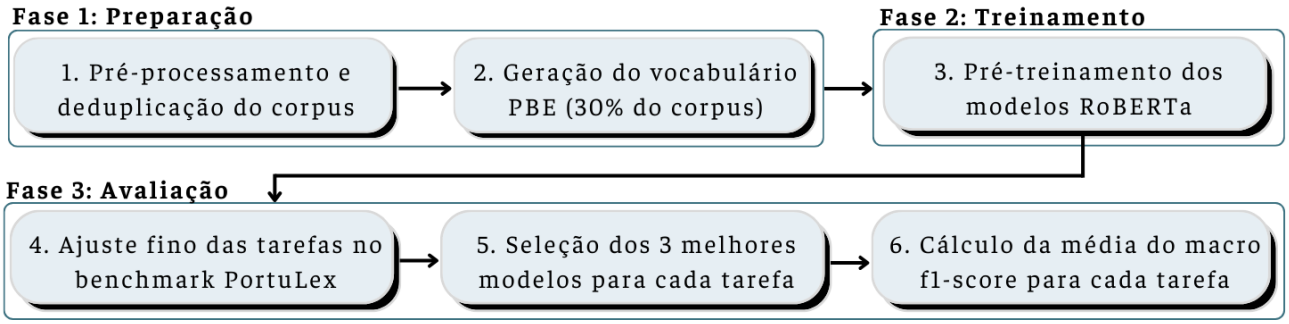

Figura 1: Fluxograma para pré-treino e avaliação no *benchmark* PortuLex.

49,1 GiB de texto e devido a sua natureza pode incluir todas as variantes do português disponíveis na *World Wide Web* como o brasileiro, europeu e africano.

OSCAR-2301 (Abadji et al., 2022)[11] é um *corpus* multilíngue extraído do *dump* de novembro a dezembro de 2022 do Common Crawl. Esta versão do OSCAR foi projetada para atender a uma ampla gama de tarefas de PLN e aprendizado de máquina, com dados organizados por idiomas e categorias temáticas. Utilizamos o subconjunto em português do OSCAR-2301, que contém 97,8 GiB de texto e também pode incluir diversas variantes do idioma português.

3.3. *Benchmark* PortuLex

Nossa pesquisa concentra-se na aquisição de dados de treinamento supervisionado abertos, anotados por especialistas jurídicos. Para manter a alta qualidade do *benchmark*, evitamos deliberadamente conjuntos de dados gerados automaticamente. À luz desses esforços, apresentamos o *benchmark* "PortuLex", composto por quatro tarefas de classificação, sendo uma para Classificação de Sentenças (CLS) e três para Classificação de Sentenças de Tokens utilizando NER, projetadas para avaliar a qualidade e o desempenho de modelos de linguagem no domínio jurídico português. A composição do PortuLex é mostrada na Tabela 2. O Portulex é composto pelos seguintes *corpora*:

RRI Identificação de Papéis Retóricos (*Rhetorical Role Identification*)(Aragy et al., 2021)[12] é um conjunto de dados de anotações retóricas no domínio jurídico, focando em sentenças extraídas de decisões judiciais do Tribunal de Justiça de Mato Grosso do Sul (Brasil). O *corpus* abrange 70 petições iniciais, contendo aproximadamente 10.000 sentenças rotuladas manualmente. São definidos oito papéis retóricos em alinhamento com o Código de Processo Civil Brasileiro, incluindo a identificação de partes, fatos, argumentos, fundamento jurídico, jurisprudência, pedidos, valor da causa e "outros".

LeNER-Br (Luz de Araujo et al., 2018)[13] é o primeiro *corpus* de NER para o domínio jurídico em português brasileiro. Ele compreende 66 documentos provenientes de tribunais superiores e estaduais, além de 04 textos de leis, totalizando 70 documentos jurídicos anotados com seis classes de entidades, sendo quatro universais (organização, pessoa, tempo e localização) e duas do domínio (legislação e jurisprudência).

UlyssesNER-Br (Albuquerque et al., 2022)[14] é um *corpus* de documentos legislativos brasileiros para NER, composto por 154 projetos de lei (PL-corpus) e 800 consultas legislativas internas (ST-corpus) da Câmara dos Deputados do Brasil. Por ter sido composto de projetos de lei públicos, somente o PL-corpus foi disponibilizado. O *corpus* abrange dois níveis de granularidade, por categorias e por tipos, com 18 tipos de entidades anotadas manualmente e estruturadas em 7 classes semânticas.

[11] https://huggingface.co/datasets/oscar-corpus/OSCAR-2301

[12] https://bit.ly/rhetoricalrole

[13] https://huggingface.co/datasets/peluz/lener_br

[14] https://github.com/ulysses-camara/ulysses-ner-br

FGV-STF (Correia et al., 2022)[15] é um *corpus* de documentos jurídicos para extração de entidades nomeadas, composto por 594 decisões do STF, selecionadas manualmente por especialistas do domínio entre 2009 e 2018. Os dados são anotados com diferentes níveis de granularidade, com foco principal em fundamento jurídico. Essas classes abrangem precedentes, citações acadêmicas e referências legislativas, com cada categoria contendo subtipos mais específicos de entidades. Usamos apenas as quatro principais entidades de granularidade mais ampla.

4. Método

Esta seção detalha o método utilizado, abrangendo a arquitetura do modelo, o treinamento, os conjuntos de dados e a avaliação. Conforme a Figura 1, o processo foi dividido em três fases: preparação dos *corpora*, treinamento e avaliação dos modelos de linguagem, que serão descritas a seguir.

4.1. Fase 1: Preparação

Nesta fase, os *corpora* foram inicialmente pré-processados para verificação de duplicatas. Seguindo a abordagem de deduplicação de Lee et al. (2022), analisamos os subconjuntos do *corpus* LegalPT usando os algoritmos *MinHash* (Broder, 2000) e *Locality Sensitive Hashing* (Har-Peled et al., 2012), identificando agrupamentos de documentos duplicados. Para isso, utilizamos duas técnicas que facilitam a comparação entre grandes volumes de dados: a técnica de *n-gramas* com sequências de cinco palavras consecutivas (5-gramas), para captura de padrões textuais, e uma assinatura de 256 elementos, para representar um resumo compacto do documento analisado, facilitando a comparação entre textos. Consideramos dois documentos idênticos quando sua Similaridade *Jaccard* (Jaccard, 1912) exceder 0.7. Quando encontradas, as duplicatas são removidas do conjunto de treinamento para melhorar a eficiência e a generalização do modelo. Os resultados do processo de deduplicação para os subconjuntos dos *corpora* LegalPT e CrawlPT podem ser encontrados na Tabela 3.

Após o passo de pré-processamento e deduplicação, buscando evitar que os modelos de domínio sejam limitados por um vocabulário genérico, aplicamos a biblioteca *HuggingFace Tokenizers*[16] e o algoritmo *Byte-Pair Encoding (BPE)* para treinar um vocabulário específico para cada *corpus* do pré-treinamento.

4.2. Fase 2: Treinamento

Nesta fase, foi executado o processo de pré-treinamento do nosso modelo de linguagem jurídica. Foi utilizado o modelo RoBERTa$_{base}$, um modelo de linguagem baseado no modelo Transformer (Liu et al., 2019). Nosso modelo foi pré-treinado com quatro configurações distintas: (i) exclusivamente no *corpus* LegalPT (RoBERTaLegalPT$_{base}$), (ii) exclusivamente no *corpus* CrawlPT (RoBERTaCrawlPT$_{base}$), (iii) combinando ambos os *corpora* (RoBERTaLexPT$_{base}$) e, (iv) exclusivamente no brWaC (RoBERTaTimbau$_{base}$).

O pré-treinamento envolveu 62.500 etapas, com um *batch size* de 2048 sequências, cada uma contendo no máximo 512 *tokens*. Esta configuração é semelhante à utilizada pelo BERTimbau, expondo o modelo a aproximadamente 65 bilhões de *tokens* durante o treinamento. Adotamos a configuração padrão dos hiperparâmetros do modelo RoBERTa. Durante o pré-treinamento, adotamos a técnica de modelagem de linguagem mascarada (*Masked Language Modeling — MLM*), onde 15% dos *tokens* de entrada foram mascarados aleatoriamente, e o modelo previu os *tokens* mascarados com base no contexto. A otimização foi conduzida com o otimizador *AdamW*, utilizando um escalonamento linear seguido de uma taxa de aprendizado com decaimento linear. A Tabela 4 apresenta os hiperparâmetros utilizados.

O processo de pré-treinamento foi executado em um cluster DGX-A100, composto de 2 GPUs Nvidia A100 com 80 GB de memória. Utilizamos a biblioteca *Fairseq* (Ott et al., 2019)[17], desenvolvida para o treinamento de modelos de linguagem usando redes neurais, entre outras tarefas. O treinamento completo de uma única configuração levou aproximadamente três dias.

4.3. Fase 3: Avaliação

A fase de Avaliação do modelo é composta por três tarefas principais: ajuste fino, seleção dos melhores modelos, e cálculo do F1-score médio para comparação dos resultados.

Inicialmente, realizou-se o ajuste fino dos modelos treinados nos conjuntos de dados do PortuLex, seguindo a abordagem proposta por

[15] O acesso ao *corpus* é disponibilizado mediante solicitação direta ao autor do artigo.

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

[17] https://github.com/facebookresearch/fairseq

Corpus	**Sub-*corpus***	**Docs.**	**Docs. depois da deduplicação**	**Duplicatas (%)**
LegalPT	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)	11.58	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	**24.194.918**	**11.946.015**	**50,63**
CrawlPT	brWaC	3.530.796	3.513.588	0,49
	OSCAR-2301 (PT)	18.031.400	10.888.966	39,61
	CC100 (PT)	38.999.388	38.059.979	2,41
	Total	**60.561.584**	**52.462.533**	**13,37**

Tabela 3: Taxa de duplicação encontrada com algoritmo Minhash-LSH (Lee et al., 2022).

Hiperparâmetros	**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	7e-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

Tabela 4: Hiperparâmetros utilizados no pré-treinamento do modelo RoBERTa.

Devlin et al. (2019). Esse método treina um codificador Transformer bidirecional para tarefas de classificação de texto e reconhecimento de entidades nomeadas. A Tabela 5 apresenta o espaço de busca explorado durante o *grid search*. Os hiperparâmetros ajustados incluíram o tamanho do lote (*batch size*) e a taxa de aprendizado (*learning rate*), enquanto os demais foram mantidos constantes.

Após a etapa de ajuste fino, realizamos a avaliação dos modelos com base nos 3 melhores *checkpoints* identificados no conjunto de validação. Por fim, a métrica final foi obtida calculando a média aritmética do macro F1-score nas divisões de teste do conjunto de dados. Essa abordagem permite avaliar a robustez dos modelos, minimizando o risco de *overfitting* nos dados de treinamento. Ao selecionar múltiplos *checkpoints*, garantimos uma avaliação mais equilibrada e robusta, com a expectativa de que os modelos mantenham um bom desempenho em dados inéditos e generalizem de forma eficiente.

Hiperparâmetros	**Espaço de busca**
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)

Tabela 5: Espaço de busca de hiperparâmetros para modelos de ajuste fino treinados no *benchmark* PortuLex.

Modelo	**Hiperparâmetros**					**PortuLex *Score* (%)**
	Tamanho do lote	**Taxa de aprendizado**	**Inicialização de pesos**	**Passos**	**Épocas**	
BERTimbau$_{base}$	128	1e-4	mBERT (sem *embeddings*)	1,000,000	8	83.78
RoBERTaTimbau$_{base}$	2048	1e-4	XLM-R$_{base}$ (sem *embeddings*)	30,000	8	84.01*
				62,500	17	83.96*
	2048	7e-4	XLM-R$_{base}$ (sem *embeddings*)	30,000	8	83.40
				62,500	17	83.94*
Corpus:				30,000	8	83.36
brWaC (16GiB)	2048	7e-4	Aleatória	62,500	17	**84.29***

Tabela 6: Macro F1-*Score* no *benchmark* PortuLex para modelos RoBERTa$_{base}$ em português pré-treinados no brWaC. As pontuações que superaram o BERTimbau$_{base}$ estão marcadas com asterisco, e a maior pontuação em negrito.

5. Resultados e Discussão

Esta seção apresenta os resultados dos experimentos com modelos baseados no RoBERTa e o RoBERTaLexPT, pré-treinado em um *corpus* combinado de textos jurídicos e genéricos. Avaliamos o impacto dos hiperparâmetros e a combinação de *corpora* (Seções 5.1 e 5.2), exploramos o pré-treinamento adicional em domínios específicos (Seção 5.3) e analisamos o aumento de parâmetros e tempo de treinamento (Seção 5.4). Por fim, comparamos nosso modelo com modelos jurídicos existentes e avaliamos sua versatilidade em *benchmarks* genéricos (Seções 5.5 e 5.6).

5.1. Replicando o BERTimbau com RoBERTa

Os experimentos nesta seção têm como objetivo investigar como diferentes hiperparâmetros afetam o desempenho do modelo em comparação com o RoBERTa, utilizando um *batch size* maior.

O modelo BERTimbau (Souza et al., 2020) foi pré-treinado com um comprimento máximo de sequência de entrada variando entre 128 e 512 *tokens*, um vocabulário de 29.794 *tokens* treinado na Wikipedia em português (Wikipedia PT), com *batch size* de 128, e executado por 1 milhão de etapas (ou 8 épocas) no *corpus* brWaC. Durante esse processo, o modelo foi exposto a um total de 65 bilhões de *tokens*. Os pesos de treinamento foram inicializados a partir dos modelos mBERT$_{base}$ e BERT$_{large}$, removendo a camada de *embeddings* inicial para acomodar o novo vocabulário em língua portuguesa.

Nesta pesquisa, avaliamos variações na taxa de aprendizado (*learning rate*) do modelo, no número de épocas de treinamento e na inicialização. Os modelos foram baseados na arquitetura RoBERTa$_{base}$, com um comprimento de *tokenização* fixo de 512 tokens e um vocabulário BPE de 50.265 *tokens*, treinado na Wikipedia PT. Os *checkpoints* foram avaliados com base no *benchmark* PortuLex proposto neste trabalho. Os resultados estão resumidos na Tabela 6.

Para manter a comparabilidade de custos computacionais com o BERTimbau, estabelecemos um limite de 65 bilhões de *tokens* de treinamento. Com o novo vocabulário BPE, corresponde a aproximadamente 17 épocas no *corpus* brWaC, ou 62.500 etapas de treinamento com um tamanho de lote de 2048 e comprimento de tokenização de 512 *tokens*. Também reportamos os resultados para 8 épocas (equivalente a 30.000 etapas de treinamento em nossa configuração). Utilizamos o modelo pré-treinado XLM-R$_{base}$ (Conneau et al., 2020) como inicialização, descartando sua camada de representações vetoriais.

Utilizando essa inicialização, foi possível perceber que o modelo conseguiu superar o BERTimbau no *benchmark* PortuLex com apenas 30.000 etapas de treinamento, alcançando um macro F1-score médio de 84,01%, em comparação com 83,78% do BERTimbau. No entanto, a execução do treinamento por mais tempo ou o ajuste da taxa de aprendizado não pareceram melhorar o desempenho do modelo. Com inicialização aleatória, nosso modelo RoBERTa apresentou desempenho inferior ao BERTimbau no ponto de 8 épocas, mas superou a inicialização do XLM-R$_{base}$ ao treinar por um período mais longo. Com 17 épocas, o modelo atingiu um macro F1-score médio de 84,29% no *benchmark* PortuLex.

Esse comportamento também foi observado no gráfico de perda do MLM no subconjunto de validação, como mostrado na Figura 2. Entre 54.000 e 55.000 etapas de treinamento, o modelo sem inicialização superou o modelo inicializado com o XLM-R. Esse ponto de inflexão

sugere que, em determinado estágio, o modelo sem inicialização pode alcançar maior capacidade de generalização, apesar do início mais lento em comparação com o modelo pré-treinado. Isso reforça a importância de explorar diferentes abordagens de inicialização ao otimizar modelos de linguagem.

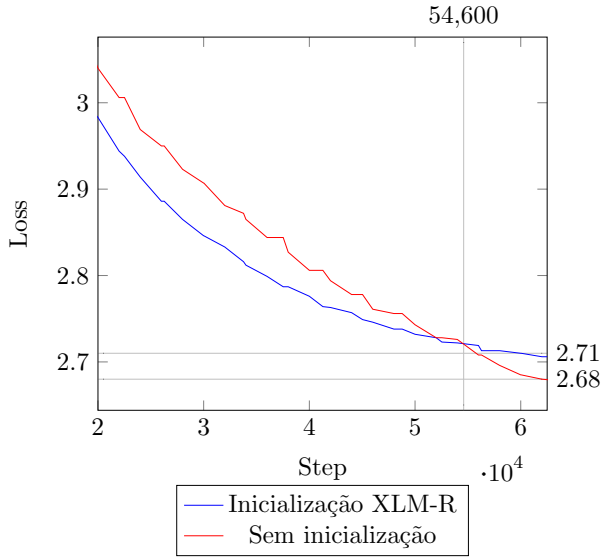

Figura 2: Perda do MLM no conjunto de validação dos modelos treinados no brWaC com diferentes inicializações.

Modelo	***Corpus***	**PortuLex *Score* (%)**
RoBERTaTimbau$_{base}$	brWaC	84,29
RoBERTaCrawlPT$_{base}$	CrawlPT	84,83
RoBERTaLegalPT$_{base}$	LegalPT	84,57
RoBERTaLexPT$_{base}$	LegalPT+CrawlPT	**85,41**

Tabela 7: Macro F1-*score* médio nos modelos pré-treinados no *benchmark* PortuLex. O RoBERTaLexPT$_{base}$, pré-treinado no *corpus* específico de domínio LegalPT e no *corpus* geral CrawlPT, alcança a maior pontuação.

5.2. Combinando *corpora* genéricos e jurídicos

Até o nosso conhecimento, as técnicas de adaptação de domínio ainda não exploraram se a combinação de um *corpus* específico de domínio com um *corpus* genérico pode melhorar o desempenho do modelo de aprendizagem, devido ao aumento do tamanho do *corpus* de pré-treinamento.

Para avaliar essa combinação, pré-treinamos modelos de linguagem no *corpus* CrawlPT (Seção 3.2), e na combinação de CrawlPT com LegalPT. Os modelos foram treinados com os hiperparâmetros definidos na Seção 5.1. O vocabulário BPE de cada modelo foi treinado com 30% dos documentos de seus respectivos *corpora*. A Tabela 7 apresenta o resultado da avaliação dos novos modelos no *benchmark* PortuLex, comparando-os com o modelo RoBERTaTimbau$_{base}$.

Curiosamente, quando usados individualmente para o pré-treinamento, o modelo CrawlPT apresentou desempenho superior ao LegalPT, apesar de seu domínio genérico. Mesmo com tamanhos semelhantes, o *corpus* CrawlPT possui mais dados únicos, com uma taxa de duplicação de 13,37%, em comparação com 50,63% no LegalPT. Este resultado pode indicar que um *corpus* genérico de alta qualidade pode ser tão eficiente quanto um *corpus* específico de domínio para o pré-treinamento de modelos de linguagem.

No entanto, ao combinar os dois *corpora*, o modelo RoBERTaLexPT resultante apresentou desempenho superior em comparação com os modelos pré-treinados em conjuntos de dados individuais. Esse resultado está de acordo com as conclusões de Kaplan et al. (2020) que afirmam que o tamanho do *corpus* é um fator chave para aumentar o desempenho do modelo, embora o estudo deles tenha examinado modelos de linguagem causais, diferindo dos modelos de linguagem mascarada utilizados em nossa pesquisa.

5.3. Pré-treinamento Adicional em Dados Específicos do Domínio

O pré-treinamento adicional (*further pre-training*) adapta modelos de linguagem previamente treinados a domínios específicos, aproveitando o conhecimento adquirido em *corpora* amplos e diversificados (Gururangan et al., 2020). Essa técnica combina a compreensão geral da linguagem com conhecimento específico do domínio, melhorando o desempenho em tarefas específicas, oferecendo vantagens computacionais ao reduzir o tempo e os recursos necessários, evitando o treinamento completo do zero.

Para explorar o impacto do pré-treinamento adicional, realizamos experimentos usando nosso modelo RoBERTaCrawlPT$_{base}$, que foi inicialmente pré-treinado no *corpus* CrawlPT. Continuamos seu pré-treinamento no *corpus* LegalPT por mais 30.000 passos, testando diferentes taxas de aprendizado. Os resultados desses experimentos são apresentados na Tabela 8.

Observamos que o pré-treinamento adicional do RoBERTaCrawlPT no *corpus* LegalPT por mais 30 mil passos melhorou seu desempenho no *benchmark* PortuLex. O melhor resultado foi alcançado com uma taxa de aprendizagem de 3e-4,

Modelo	**Passos**	***Corpus***	**Taxa de aprendizado**	**PortuLex Score (%)**
RoBERTaCrawlPT$_{base}$	62,5 mil	CrawlPT	7e-4	84,83
+ pré-treino continuado	+ 30 mil	+ LegalPT	7e-4	84,95
	+ 30 mil	+ LegalPT	3e-4	85,37
	+ 30 mil	+ LegalPT	1e-4	85,29
RoBERTaLexPT$_{base}$	62,5 mil	LegalPT + CrawlPT	7e-4	**85,41**

Tabela 8: Comparação dos resultados e um pré-treino adicional de domínio a partir de um modelo genérico e o RoBERTaLexPT$_{base}$, F1-*score* médio avaliados nas divisões de teste do benchmark PortuLex.

produzindo um F1-score médio de 85,37%, o que representa uma melhoria significativa em relação aos 84,83% do modelo inicial.

Curiosamente, essa abordagem de pré-treinamento adicional chega perto de igualar o desempenho do RoBERTaLexPT (85,41%), que foi pré-treinado no *corpus* combinado LegalPT+CrawlPT desde o início. A pequena diferença de desempenho (0,04%) sugere que ambas as estratégias (pré-treinamento combinado e pré-treinamento adicional) podem ser eficazes para o desenvolvimento de modelos de linguagem específicos para o domínio.

Os resultados destacam a importância do ajuste da taxa de aprendizagem em cenários de pré-treinamento adicional. Observamos que uma taxa de aprendizagem mais baixas (3e-4 e 1e-4) que a taxa de aprendizagem inicial de pré-treino do modelo (7e-4) tem um resultado melhor.

Embora o RoBERTaLexPT ainda mantenha uma leve vantagem no desempenho, a abordagem de pré-treinamento adicional oferece uma alternativa viável, especialmente quando os recursos computacionais são limitados ou quando se adaptam modelos existentes a novos domínios.

5.4. Expandindo Parâmetros e Tempo de Treinamento

Para investigar o impacto do tamanho do modelo e do tempo de treinamento no desempenho, conduzimos experimentos adicionais com variações do RoBERTaLexPT. Especificamente, treinamos uma versão do modelo base por um período mais longo (RoBERTaLexPT+$_{base}$) e implementamos uma versão maior do modelo (RoBERTaLexPT$_{large}$), também treinando-a por períodos curto e longo. A Tabela 9 apresenta os resultados desses experimentos.

Os resultados revelam que o aumento do tamanho do modelo de *base* para *large* não proporcionou benefícios substanciais para as tarefas avaliadas no *benchmark* PortuLex. O RoBERTaLexPT$_{large}$ com 62.5k passos de treinamento apresentou um desempenho ligeiramente inferior ao RoBERTaLexPT$_{base}$ (85.30% vs 85.41%). Quando treinados por 125k passos, ambos os modelos base e large atingiram o mesmo F1-score médio de 85.46%, indicando uma melhoria marginal em relação ao treinamento mais curto.

Esta falta de ganho significativo com o modelo maior pode ser atribuída principalmente ao tamanho limitado do *corpus* de treinamento. Com 125k passos, os modelos completam aproximadamente 2,5 épocas no dataset combinado LegalPT+CrawlPT. O fenômeno observado está alinhado com as descobertas de Kaplan et al. (2020) que demonstraram que o desempenho dos modelos de linguagem está intrinsecamente ligado à quantidade de dados de treinamento. No nosso caso, o modelo *large*, com sua maior capacidade, provavelmente requer um *corpus* substancialmente maior para demonstrar vantagens significativas sobre o modelo base.

Além disso, o RoBERTaLexPT$_{base}$ demonstrou notável eficiência, alcançando um desempenho competitivo com significativamente menos parâmetros (125M vs 355M) e em menos tempo de treinamento. Isso sugere que, para as tarefas específicas do PortuLex, o modelo base já possui capacidade suficiente para capturar as nuances necessárias.

Estes resultados destacam a importância de considerar cuidadosamente o equilíbrio entre o tamanho do modelo, o volume de dados de treinamento e a complexidade das tarefas alvo. Para o *benchmark* PortuLex e o *corpus* utilizado, o RoBERTaLexPT$_{base}$ parece oferecer o melhor equilíbrio entre eficiência computacional e desempenho.

Modelo	**Parâmetros (M)**	**Passos (mil)**	**Taxa de aprendizado**	**PortuLex) *Score* (%)**
RoBERTaLexPT$_{base}$	125	62,5	7e-4	85,41
RoBERTaLexPT+$_{base}$	125	125	7e-4	**85,46**
RoBERTaLexPT$_{large}$	355	62,5	2e-4	85,30
RoBERTaLexPT+$_{large}$	355	125	2e-4	**85,46**

Tabela 9: Comparação de desempenho entre modelos RoBERTaLexPT de diferentes tamanhos e tempos de treinamento, com F1-*score* médio nas divisões de teste do *benchmark* PortuLex. O símbolo **"+"** indica treinamento estendido.

Modelo	**LeNER**	**UlyNER-PL** Categorias/Tipos	**FGV-STF** Amplo	**RRIP**	**Média (%)**
BERTimbau$_{base}$	88,34	86,39/83,83	79,34	82,34	83,78
BERTimbau$_{large}$	88,64	87,77/84,74	79,71	**83,79**	84,60
Albertina-PT-BR$_{base}$	89,26	86,35/84,63	79,30	81,16	83,80
Albertina-PT-BR$_{xlarge}$	90,09	88,36/**86,62**	79,94	82,79	85,08
BERTikal$_{base}$	83,68	79,21/75,70	77,73	81,11	79,99
JurisBERT$_{base}$	81,74	81,67/77,97	76,04	80,85	79,61
BERTimbauLAW$_{base}$ (Viegas et al., 2023)	84,90	87,11/84,42	79,78	82,35	83,20
Legal-XLM-R$_{base}$	87,48	83,49/83,16	79,79	82,35	83,24
Legal-XLM-R$_{large}$	88,39	84,65/84,55	79,36	81,66	83,50
Legal-RoBERTa-PT$_{large}$	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
RoBERTaLexPT+$_{base}$	**90,97**	88,49/86,16	80,12	83,45	**85,46**
RoBERTaLexPT+$_{large}$	90,61	88,30/86,16	**80,60**	83,41	**85,46**

Tabela 10: Média de Macro F1-*score* (%) para vários modelos avaliados nas divisões de teste do *benchmark* PortuLex.

5.5. Comparação com outros modelos jurídicos

A Tabela 10 apresenta o desempenho do modelo RoBERTaLexPT em comparação a modelos jurídicos abertos em português no conjunto de dados do *benchmark* PortuLex.

Destacamos inicialmente que, apesar de usar apenas a configuração *base*, o RoBERTaLexPT+$_{base}$ superou modelos significativamente maiores, como BERTimbau$_{large}$, Albertina-PT-BR$_{xlarge}$ (Rodrigues et al., 2023), e Legal-XLM-R$_{large}$ (Niklaus et al., 2024). Este resultado destaca a eficácia do RoBERTaLexPT, resultante do pré-treinamento em dados jurídicos e genéricos combinados.

O RoBERTaLexPT alcançou o melhor desempenho nos conjuntos LeNER e FGV-STF, mesmo quando comparado a modelos muito maiores. Comparando com o UlyssesNER-Br/Pl-corpus, o RoBERTaLexPT+$_{base}$ obteve resultados competitivos. O único conjunto em que o RoBERTaLexPT+$_{base}$ é superado é o RRI, onde o BERTimbau$_{large}$ tem uma pequena vantagem de 0,34% de F1-score.

Em contraste, alguns trabalhos anteriores afirmaram ter obtido desempenho superior ao BERTimbau em certas tarefas jurídicas (Polo et al., 2021; Viegas et al., 2023). Esses modelos, na verdade, apresentaram um desempenho inferior ao BERTimbau em nossos experimentos com o *benchmark* PortuLex. Por exemplo, o JurisBERT alcançou apenas um F1-score médio de 79,61%, comparado aos 83,78% do BERTimbau. Uma possível explicação para essa discrepância é que as avaliações originais foram limitadas a um único conjunto de dados, provavelmente favorecendo os dados de treinamento específicos do modelo.

Em resumo, o RoBERTaLexPT e suas variações demonstrou consistentemente uma alta eficácia em PLN jurídico, mesmo em seu tamanho base. Com dados suficientes de pré-treinamento, acreditamos que nosso modelo pode superar modelos sobreparametrizados. Os resultados destacam a importância de dados de treinamento diversos em vez do simples aumento da escala do modelo.

5.6. Avaliações em *benchmarks* genéricos

Para avaliar a versatilidade dos modelos desenvolvidos neste trabalho, realizamos experimentos adicionais em tarefas genéricas de PLN em português. Utilizamos os *benchmarks* ASSIN2 Real et al. (2020) e HAREM (Santos et al., 2006), seguindo a metodologia proposta por Souza et al. (2020). Estes *benchmarks* incluem tarefas de Similaridade Textual de Sentenças (STS), Reconhecimento de Inferência Textual (RTE) e Reconhecimento de Entidades Nomeadas (NER). A Tabela 11 apresenta os resultados comparativos.

Os resultados mostram que nossos modelos, especialmente o RoBERTaLexPT+$_{large}$, apresentam desempenho competitivo em tarefas genéricas de PLN, apesar de terem sido treinados com foco em dados jurídicos. O RoBERTaLexPT+$_{large}$ supera o BERTimbau$_{large}$ em todas as tarefas, exceto no ASSIN2 RTE, onde a diferença é marginal (89.71% vs 89.77%). Notavelmente, o RoBERTaLexPT+$_{large}$ alcança o melhor desempenho nas tarefas de NER do HAREM, tanto no cenário padrão quanto no seletivo.

É interessante observar que o RoBERTaLegalPT$_{base}$, treinado exclusivamente em dados jurídicos, apresenta um desempenho inferior nas tarefas genéricas, como esperado. No entanto, o RoBERTaLexPT+$_{base}$, que combina dados jurídicos e genéricos, consegue manter um desempenho competitivo, superando até mesmo o BERTimbau$_{base}$ e o Albertina-PT-BR$_{base}$ em algumas tarefas.

O Albertina-PT-BR$_{xlarge}$ ainda mantém a liderança na média geral, principalmente devido ao seu desempenho superior nas tarefas do ASSIN2. Isso pode ser atribuído ao seu tamanho significativamente maior (900 milhões de parâmetros) e ao seu treinamento com maiores recursos.

Estes resultados demonstram que a abordagem de combinar dados jurídicos e genéricos no pré-treinamento, como feito no RoBERTaLexPT, não apenas melhora o desempenho em tarefas jurídicas específicas, mas também mantém uma forte capacidade de generalização para tarefas de PLN mais amplas. Isso sugere que o modelo adquiriu uma compreensão robusta da língua portuguesa em geral, além de conhecimentos específicos do domínio jurídico.

6. Conclusão

Este trabalho apresenta o RoBERTaLexPT, um modelo de linguagem jurídica em português pré-treinado em um *corpus* combinado de textos jurídicos e gerais. Até o nosso conhecimento, ao longo do processo, criamos o maior *corpus* jurídico em português (LegalPT) agregando diversas fontes, obtendo melhorias significativas de desempenho por meio da deduplicação, além de introduzir o *benchmark* PortuLex para uma avaliação rigorosa dos modelos.

Demonstramos que utilizar outros modelos como inicialização de pesos para o pré-treinamento de modelos de linguagem pode melhorar o desempenho em cenários com recursos limitados, embora isso tenha um *trade-off* em configurações de treinamento mais longas.

Nossos resultados indicam que combinar um *corpus* específico de domínio (LegalPT) e um *corpus* genérico (CrawlPT) para o pré-treinamento oferece benefícios complementares. Apesar de seu tamanho compacto em comparação com modelos anteriores, o RoBERTaLexPT na configuração base demonstra eficácia de ponta em PLN jurídico em português, destacando a importância dos dados de pré-treinamento em relação à escala do modelo.

Investigações adicionais sobre o impacto do tamanho do modelo e do tempo de treinamento revelaram que, para as tarefas do PortuLex, o aumento do tamanho do modelo de *base* para *large* não proporcionou benefícios substanciais. Isso sugere que o tamanho do *corpus* de treinamento pode ser um fator limitante mais significativo do que o tamanho do modelo para melhorias de desempenho em nosso contexto específico.

Além disso, avaliamos a versatilidade do RoBERTaLexPT em tarefas genéricas de PLN utilizando os *benchmarks* ASSIN2 e HAREM. Os resultados demonstraram que nosso modelo mantém um desempenho competitivo em tarefas não jurídicas, superando o BERTimbau em várias métricas. Isso indica que a abordagem de combinar dados jurídicos e genéricos no pré-treinamento não apenas melhora o desempenho em tarefas jurídicas específicas, mas também mantém uma forte capacidade de generalização.

O RoBERTaLexPT, juntamente com o LegalPT e o PortuLex, representam um avanço sig-

Modelo	**ASSIN2 RTE** (F1-*Score*)	**ASSIN2 STS** (Pearson)	**HAREM Padrão** (F1-*Score*)	**HAREM Seletivo** (F1-*Score*)	**Média (%)**
BERTimbau$_{base}$	88,76	84,15	74,02	80,53	81,87
BERTimbau$_{large}$	89,77	85,53	75,04	80,97	82,83
Albertina-PT-BR$_{base}$	86,91	82,25	73,72	78,82	80,43
Albertina-PT-BR$_{xlarge}$	**91,60**	**85,78**	76,62	82,75	**84,19**
RoBERTaTimbau$_{base}$	87,73	83,11	73,12	78,99	80,74
RoBERTaCrawlPT$_{base}$	88,28	83,20	74,51	80,24	81,56
RoBERTaLegalPT$_{base}$	85,34	78,89	67,73	75,25	76,80
RoBERTaLexPT+$_{base}$	88,68	83,19	73,97	81,36	81,80
RoBERTaLexPT+$_{large}$	89,71	85,50	**76,72**	**82,83**	83,69

Tabela 11: Comparação dos diversos modelos treinados neste trabalhos com modelos de linguagem genéricos em português nos *benchmarks* do ASSIN2 e HAREM.

nificativo para o processamento de linguagem natural do domínio jurídico em português, enfrentando limitações de recursos e modelos. Estes recursos oferecem uma base sólida para futuras pesquisas e aplicações no campo do PLN jurídico em português.

Trabalhos futuros podem explorar o pré-treinamento de modelos RoBERTa ainda maiores, investigar técnicas para melhor aproveitamento de modelos grandes com *corpora* limitados, expandir o *corpus* LegalPT, e aprimorar o *benchmark* PortuLex. Além disso, a aplicação do RoBERTaLexPT em tarefas jurídicas mais complexas e sua adaptação para outros domínios especializados são direções promissoras para pesquisas futuras.

Agradecimento

Este trabalho foi apoiado pelo Centro de Excelência em Inteligência Artificial (CEIA) do Instituto de Informática da Universidade Federal de Goiás (INF-UFG). Ellen Souza e Nadia Félix são financiadas pela FAPESP, por meio de um acordo entre a USP e a Câmara dos Deputados do Brasil. André C.P.L.F. de Carvalho é apoiado pelo CNPq. Agradecemos ao CEIA-UFG, ao Instituto de Inteligência Artificial (IAIA) e às agências de fomento à pesquisa, às quais expressamos nossa gratidão pelo suporte concedido ao desenvolvimento desta pesquisa.

Referências

Abadji, Julien, Pedro Ortiz Suarez, Laurent Romary & Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. arXiv [cs.CL]. 10.48550/arXiv.2201.06642

Akbik, Alan, Laura Chiticariu, Marina Danilevsky, Yonas Kbrom, Yunyao Li & Huaiyu Zhu. 2016. Multilingual information extraction with PolyglotIE. Em *26th International Conference on Computational Linguistics (COLING)*, 268–272.

Al-qurishi, Muhammad, Sarah Alqaseemi & Riad Souissi. 2022. AraLegal-BERT: A pretrained language model for Arabic legal text. Em *Natural Legal Language Processing Workshop (NLLP)*, 338–344. 10.18653/v1/2022.nllp-1.31

Albuquerque, Hidelberg O., 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 & Adriano L. I. Oliveira. 2022. UlyssesNER-Br: A corpus of Brazilian legislative documents for named entity recognition. Em *Computational Processing of the Portuguese Language (PROPOR)*, 3–14. 10.1007/978-3-030-98305-5_1

Aragy, Roberto, Eraldo Rezende Fernandes & Edson Norberto Caceres. 2021. Rhetorical role identification for Portuguese legal documents. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 557–571. 10.1007/978-3-030-91699-2_38

Beltagy, Iz, Kyle Lo & Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. Em *Conference on Empirical Methods in Natural Language Processing and*

the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3615–3620. 10.18653/v1/D19-1371

Bonifacio, Luiz Henrique, Paulo Arantes Vilela, Gustavo Rocha Lobato & Eraldo Rezende Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in Portuguese. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 648–662. 10.1007/978-3-030-61377-8_46

Broder, Andrei Z. 2000. Identifying and filtering near-duplicate documents. Em *Combinatorial Pattern Matching (CPM)*, 1–10. 10.1007/3-540-45123-4_1

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. arXiv [cs.CL]. 10.48550/arXiv.2005.14165

Cabrera-Diego, Luis Adrián & Akshita Gheewala. 2023. Jus Mundi at SemEval-2023 Task 6: Using a frustratingly easy domain adaption for a legal named entity recognition system. Em *17th International Workshop on Semantic Evaluation (SemEval)*, 1783–1790. 10.18653/v1/2023.semeval-1.247

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras & Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of law school. Em *Findings of the Association for Computational Linguistics*, 2898–2904. 10.18653/v1/2020.findings-emnlp.261

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. Em *58th Meeting of the Association for Computational Linguistics*, 8440–8451. 10.18653/v1/2020.acl-main.747

Correia, Fernando A., Alexandre A. A. Almeida, José Luiz Nunes, Kaline G. Santos, Ivar A. Hartmann, Felipe A. Silva & Hélio Lopes. 2022. Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. *Information Processing & Management* 59(1). 102794. 10.1016/j.ipm.2021.102794

Costa, Rosimeire, 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 & Adriano L. I. Oliveira. 2022. Expanding UlyssesNER-Br named entity recognition corpus with informal user-generated text. Em *EPIA Conference on Artificial Intelligence*, 767–779. 10.1007/978-3-031-16474-3_62

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Douka, Stella, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani & David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. Em *Natural Legal Language Processing Workshop (NLLP)*, 95–101. 10.18653/v1/2021.nllp-1.9

Firdaus Solihin, Rizal Fathoni Aji, Indra Budi & Edmon Makarim. 2021. Advancement of information extraction use in legal documents. *International Review of Law, Computers & Technology* 35(3). 322–351. 10.1080/13600869.2021.1964225

Garcia, Eduardo A. S., Nadia F. F. Silva, Felipe Siqueira, Hidelberg O. Albuquerque, Juliana R. S. Gomes, Ellen Souza & Eliomar A. Lima. 2024. RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 374–383.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey & Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. Em *58th Meeting of the Association for Computational Linguistics (ACL)*, 8342–8360. 10.18653/v1/2020.acl-main.740

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre & Marta

Villegas. 2021. Spanish legalese language model and corpora. arXiv [cs.CL/cs.AI]. 10.48550/arXiv.2110.12201

Har-Peled, Sariel, Piotr Indyk & Rajeev Motwani. 2012. Approximate nearest neighbors: Towards removing the curse of dimensionality. *Theory of Computing* 8(1). 321–350. 10.4086/toc.2012.v008a014

Jaccard, Paul. 1912. The distribution of the flora in the alpine zone. *New phytologist* 11(2). 37–50. 10.1111/j.1469-8137.1912.tb05611.x

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu & Dario Amodei. 2020. Scaling laws for neural language models. arXiv [cs.LG/stat.ML]. 10.48550/arXiv.2001.08361

Kapoor, Arnav, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru & Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. Em *Findings of the Association for Computational Linguistics*, 3521–3536. 10.18653/v1/2022.findings-acl.278

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So & Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4). 1234–1240. 10.1093/bioinformatics/btz682

Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch & Nicholas Carlini. 2022. Deduplicating training data makes language models better. Em *60th Meeting of the Association for Computational Linguistics (ACL)*, 8424–8445. 10.18653/v1/2022.acl-long.577

Licari, Daniele & Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A pre-trained transformer language model for italian law. Em *23rd International Conference on Knowledge Engineering and Knowledge Management*,

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv [cs.CL]. 10.48550/arXiv.1907.11692

Luz de Araujo, Pedro Henrique, Teófilo E. De Campos, Renato R. R. De Oliveira, Matheus Stauffer, Samuel Couto & Paulo Bermejo. 2018. LeNER-Br: A dataset for named entity recognition in Brazilian legal text. Em *Computational Processing of the Portuguese Language (PROPOR)*, 313–323. 10.1007/978-3-319-99722-3_32

Masala, Mihai, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea & Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. Em *Natural Legal Language Processing Workshop (NLLP)*, 86–94. 10.18653/v1/2021.nllp-1.8

Menezes-Neto, Elias Jacob de & 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. 10.1371/journal.pone.0272287

Niklaus, Joel, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis & Daniel Ho. 2024. MultiLegalPile: A 689GB multilingual legal corpus. Em *62nd Meeting of the Association for Computational Linguistics (ACL)*, 15077–15094. 10.18653/v1/2024.acl-long.805

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier & Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 48–53. 10.18653/v1/N19-4009

Paul, Shounak, Arpan Mandal, Pawan Goyal & Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: a case study on indian law. Em *19th International Conference on Artificial Intelligence and Law (ICAIL)*, 187–196. 10.1145/3594536.3595165

Polo, Felipe, Gabriel Mendonça, Kauê Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Ferreira, Leticia Lima, Antônio Maia & Renato Vicente. 2021. LegalNLP - natural language processing methods for the Brazilian legal language. Em *18th Encontro Nacional de Inteligência Artificial e Computacional*, 763–774. 10.5753/eniac.2021.18301

Real, Livy, Erick Fonseca & Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. Em *Computational Processing of the Portuguese Language (PROPOR)*, 406–412. 10.1007/978-3-030-41505-1_39

Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso & Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer Albertina PT-*. Em *EPIA Conference on Artificial Intelligence*, 441–453. 10.1007/978-3-031-49008-8_35

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*, 1986–1991.

Siqueira, Felipe A., Douglas Vitório, Ellen Souza, José A. P. Santos, Hidelberg O. Albuquerque, Márcio S. Dias, Nádia F. F. Silva, André C. P. L. F. de Carvalho, Adriano L. I. Oliveira & Carmelo Bastos-Filho. 2024. Ulysses Tesemõ: a new large corpus for brazilian legal and governmental domain. *Language Resources and Evaluation* 10.1007/s10579-024-09762-8

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em Ricardo Cerri & Ronaldo C. Prati (eds.), *Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. 10.1007/978-3-030-61377-8_28

Tirumala, Kushal, Daniel Simig, Armen Aghajanyan & Ari Morcos. 2023. D4: Improving LLM pretraining via document de-duplication and diversification. Em *Advances in Neural Information Processing Systems*, vol. 36, 53983–53995.

Viegas, Charles F. O., Bruno Catais Costa & Renato Porfirio Ishii. 2023. JurisBERT: Transformer-based model for embedding legal texts. Em *International Conference on Computational Science and Its Applications (ICCSA)*, 349–365. 10.1007/978-3-031-36805-9_24

Wagner Filho, Jorge A., Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4339–4344.

Willian Sousa, Antonio & Marcos Fabro. 2019. Iudicium Textum Dataset: Uma base de textos jurídicos para NLP. Em *14th Simpósio Brasileiro de Banco de Dados*, 1–11.

Xiao, Chaojun, Xueyu Hu, Zhiyuan Liu, Cunchao Tu & Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open* 2. 79–84. 10.1016/j.aiopen.2021.06.003

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua & Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. Em *North American Chapter of the Association for Computational Linguistics (NAACL)*, 483–498. 10.18653/v1/2021.naacl-main.41

Zhong, Haoxi, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu & Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. Em *58th Meeting of the Association for Computational Linguistics (ACL)*, 5218–5230. 10.18653/v1/2020.acl-main.466

Modelo	***Dataset***	***Learning Rate***	**Batch Size**	**Passos**	**Épocas**	**F1-*Score* (Validação)**
RoBERTaTimbau$_{base}$	LeNER-Br	2.5e-5	16	3000	6.1	89.68
RoBERTaTimbau$_{base}$	UlyssesNER-Br T.	5.e-5	16	1750	12.2	89.73
RoBERTaTimbau$_{base}$	UlyssesNER-Br C.	5.e-5	16	3750	26.2	88.79
RoBERTaTimbau$_{base}$	FGV-STF Amplo	5.e-5	32	3000	15.0	77.93
RoBERTaTimbau$_{base}$	RRI	5.e-5	16	1500	2.9	83.17
RoBERTaCrawlPT$_{base}$	LeNER-Br	7.5e-6	32	3500	14.2	89.76
RoBERTaCrawlPT$_{base}$	UlyssesNER-Br T.	5.e-5	16	750	5.2	88.74
RoBERTaCrawlPT$_{base}$	UlyssesNER-Br C.	2.5e-5	16	2000	14.0	88.27
RoBERTaCrawlPT$_{base}$	FGV-STF Amplo	2.5e-5	16	5000	13.8	76.55
RoBERTaCrawlPT$_{base}$	RRI	5.e-5	32	3000	11.6	82.82
RoBERTaLegalPT$_{base}$	LeNER-Br	5.e-5	32	2000	8.1	90.62
RoBERTaLegalPT$_{base}$	UlyssesNER-Br T.	5.e-5	16	750	5.2	88.62
RoBERTaLegalPT$_{base}$	UlyssesNER-Br C.	5.e-5	16	2000	14.0	88.36
RoBERTaLegalPT$_{base}$	FGV-STF Amplo	5.e-5	16	2500	7.6	75.98
RoBERTaLegalPT$_{base}$	RRI	5.e-5	16	1000	1.9	82.94
RoBERTaLexPT$_{base}$	LeNER-Br	1.e-5	32	3750	15.2	89.83
RoBERTaLexPT$_{base}$	UlyssesNER-Br T.	5.e-5	32	2250	31.2	90.71
RoBERTaLexPT$_{base}$	UlyssesNER-Br C.	5.e-5	32	1750	24.3	89.12
RoBERTaLexPT$_{base}$	FGV-STF Amplo	2.5e-5	16	4000	11.9	75.28
RoBERTaLexPT$_{base}$	RRI	5.e-5	16	4000	7.7	84.29
RoBERTaLexPT+$_{base}$	LeNER-Br	5.e-5	16	2750	5.6	90.39
RoBERTaLexPT+$_{base}$	UlyssesNER-Br T.	5.e-5	16	1750	12.2	90.07
RoBERTaLexPT+$_{base}$	UlyssesNER-Br C.	5.e-5	16	2250	15.7	89.36
RoBERTaLexPT+$_{base}$	FGV-STF Amplo	5.e-5	16	2000	6.0	75.11
RoBERTaLexPT+$_{base}$	RRI	5.e-5	16	1000	1.9	84.33
RoBERTaLexPT$_{large}$	LeNER-Br	1.e-5	16	3750	7.6	89.64
RoBERTaLexPT$_{large}$	UlyssesNER-Br T.	2.5e-5	16	2000	14.0	89.81
RoBERTaLexPT$_{large}$	UlyssesNER-Br C.	5.e-5	32	3250	45.1	88.83
RoBERTaLexPT$_{large}$	FGV-STF Amplo	5.e-5	32	1500	8.9	75.87
RoBERTaLexPT$_{large}$	RRI	7.5e-6	32	500	1.9	83.77
RoBERTaLexPT+$_{large}$	LeNER-Br	7.5e-6	16	1750	3.6	90.37
RoBERTaLexPT+$_{large}$	UlyssesNER-Br T.	2.5e-5	16	1000	7.0	89.62
RoBERTaLexPT+$_{large}$	UlyssesNER-Br C.	1.e-5	16	3500	24.5	90.12
RoBERTaLexPT+$_{large}$	FGV-STF Amplo	5.e-5	16	4000	11.9	75.51
RoBERTaLexPT+$_{large}$	RRI	2.5e-5	32	500	1.9	84.78

Tabela 12: Melhores configurações de hiperparâmetros para modelos na fase de ajuste fino do PortuLex. O F1-*score* corresponde ao Macro F1-*Score* no conjunto de validação de cada dataset.

Modelo	***Dataset***	***Learning Rate***	***Batch Size***	**Passos**	**Épocas**	**Pontuação (Validação)**
BERTimbau$_{base}$	ASSIN2 RTE	1.e-5	16	3250	8.0	97.20
BERTimbau$_{base}$	ASSIN2 STS	2.5e-5	16	5750	14.1	96.87
BERTimbau$_{base}$	HAREM Total	5.e-5	32	750	46.9	74.21
BERTimbau$_{base}$	HAREM Seletivo	5.e-5	32	500	31.2	86.05
BERTimbau$_{large}$	ASSIN2 RTE	2.5e-5	32	750	3.7	96.40
BERTimbau$_{large}$	ASSIN2 STS	5.e-5	32	5000	24.5	97.38
BERTimbau$_{large}$	HAREM Total	5.e-5	16	1000	32.3	75.75
BERTimbau$_{large}$	HAREM Seletivo	5.e-5	16	750	24.2	87.82
Albertina-PT-BR$_{base}$	ASSIN2 RTE	7.5e-6	32	2500	12.2	96.40
Albertina-PT-BR$_{base}$	ASSIN2 STS	2.5e-5	16	7250	17.8	97.32
Albertina-PT-BR$_{base}$	HAREM Total	2.5e-5	16	1250	27.2	74.19
Albertina-PT-BR$_{base}$	HAREM Seletivo	5.e-5	16	1500	32.6	85.30
Albertina-PT-BR$_{xlarge}$	ASSIN2 RTE	7.5e-6	16	1500	3.7	97.40
Albertina-PT-BR$_{xlarge}$	ASSIN2 STS	1.e-5	32	5250	25.7	97.88
Albertina-PT-BR$_{xlarge}$	HAREM Total	1.e-5	32	1000	47.6	80.08
Albertina-PT-BR$_{xlarge}$	HAREM Seletivo	2.5e-5	32	1000	47.6	91.33
RoBERTaTimbau$_{base}$	ASSIN2 RTE	2.5e-5	32	2000	9.8	96.40
RoBERTaTimbau$_{base}$	ASSIN2 STS	2.5e-5	32	5750	28.2	96.99
RoBERTaTimbau$_{base}$	HAREM Total	2.5e-5	16	1250	43.1	74.34
RoBERTaTimbau$_{base}$	HAREM Seletivo	5.e-5	32	750	50.0	83.15
RoBERTaCrawPT$_{base}$	ASSIN2 RTE	2.5e-5	32	3250	15.9	96.60
RoBERTaCrawPT$_{base}$	ASSIN2 STS	5.e-5	32	4500	22.1	97.03
RoBERTaCrawPT$_{base}$	HAREM Total	2.5e-5	16	1000	34.5	77.67
RoBERTaCrawPT$_{base}$	HAREM Seletivo	5.e-5	16	750	25.9	86.30
RoBERTaLegalPT$_{base}$	ASSIN2 RTE	7.5e-6	16	5500	13.5	95.60
RoBERTaLegalPT$_{base}$	ASSIN2 STS	5.e-5	32	4000	19.6	96.96
RoBERTaLegalPT$_{base}$	HAREM Total	5.e-5	16	750	24.2	74.90
RoBERTaLegalPT$_{base}$	HAREM Seletivo	5.e-5	16	1000	32.3	83.91
RoBERTaLexPT+$_{base}$	ASSIN2 RTE	1.e-5	32	1750	8.6	97.00
RoBERTaLexPT+$_{base}$	ASSIN2 STS	5.e-5	32	4000	19.6	96.90
RoBERTaLexPT+$_{base}$	HAREM Total	5.e-5	16	1250	43.1	79.02
RoBERTaLexPT+$_{base}$	HAREM Seletivo	5.e-5	16	1250	43.1	91.12
RoBERTaLexPT+$_{large}$	ASSIN2 RTE	7.5e-6	32	1750	8.6	97.00
RoBERTaLexPT+$_{large}$	ASSIN2 STS	1.e-5	32	3250	15.9	97.57
RoBERTaLexPT+$_{large}$	HAREM Total	1.e-5	16	1250	43.1	78.87
RoBERTaLexPT+$_{large}$	HAREM Seletivo	1.e-5	16	500	17.2	87.28

Tabela 13: Melhores hiperparâmetros para os modelos na fase de ajuste fino dos *benchmarks* ASSIN2 e HAREM. A pontuação corresponde à métrica principal: coeficiente de correlação de Pearson para ASSIN2 STS e Macro F1-Score para os demais datasets.