UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EDUARDO AUGUSTO SANTOS GARCIA

Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Eduardo Augusto Santos Garcia

3. Título do trabalho

Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 03/07/2024, às 13:20, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Eduardo Augusto Santos Garcia**, **Discente**, em 03/07/2024, às 13:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4636006** e o código CRC **203069B5**.

Referência: Processo nº 23070.023666/2024-93 SEI nº 4636006

EDUARDO AUGUSTO SANTOS GARCIA

Adaptação de domínio Legal em Modelos de Linguagens em português - Desenvolvimento e avaliação de modelos baseados em RoBERTa em corpora legais

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Informática, da Universidade Federal de Goiás, como requisito para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de pequisa: Sistemas Inteligentes e Aplicações.

Orientadora: Profa. Doutora Nádia Félix Felipe da Silva

Co-Orientador: Prof. Doutor Eliomar Araújo de Lima

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Garcia, Eduardo Augusto Santos
Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora [manuscrito] / Eduardo Augusto Santos Garcia. - 2024.
lxxxii, 82 f.: il.

Orientador: Profa. Dra. Nádia Félix Felipe da Silva; co-orientador Dr. Eliomar Araújo de Lima.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.
Bibliografia.
Inclui símbolos, gráfico, tabelas, lista de figuras, lista de tabelas.

1. Natural Language Processing. 2. Language Model. 3. Legal Domain. 4. Legal Benchmark. I. Silva, Nádia Félix Felipe da, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **17/2024** da sessão de Defesa de Dissertação de **Eduardo Augusto Santos Garcia**, que confere o título de Mestre em **Ciência da Computação**, na área de concentração em **Ciência da Computação**.

Aos vinte e oito dias do mês de maio de dois mil e vinte e quatro, a partir das nove horas, via sistema de webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora"**. Os trabalhos foram instalados pela Orientadora, Professora Doutora Nádia Félix Felipe da Silva (INF/UFG) com a participação dos demais membros da Banca Examinadora: Doutor José Avelino Placca (CNJ/PNUD), membro titular externo; Professor Doutor Anderson da Silva Soares (INF/UFG), membro titular interno e Prof. Eliomar Araújo de Lima (Co-orientador). Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pela Professora Doutora Nádia Félix Felipe da Silva, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora e pelo Coordenador do Programa de Pós-Graduação em Ciência da Computação, Professor Doutor Fabrizzio Alphonsus Alves de Melo Nunes Soares, em substituição à assinatura do membro externo Doutor José Avelino Placca, aos vinte e oito dias do mês de maio de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Eliomar Araujo De Lima**, **Professor do Magistério Superior**, em 28/05/2024, às 10:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Eduardo Augusto Santos Garcia**, **Discente**, em 28/05/2024, às 10:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Anderson Da Silva Soares**, **Professor do Magistério Superior**, em 28/05/2024, às 10:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 28/05/2024, às 10:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Fabrizzio Alphonsus Alves De Melo Nunes Soares**, **Coordenador de Pós-Graduação**, em 28/05/2024, às 18:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Referência: Processo nº 23070.023666/2024-93 SEI nº 4566384

EDUARDO AUGUSTO SANTOS GARCIA

Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Instituto de Informática, da Universidade Federal de Goiás, como requisito para obtenção do título de Mestre em Ciência da Computação, aprovada em 28 de Maio de 2024, pela Banca Examinadora constituída pelos professores:

Profa. Doutora Nádia Félix Felipe da Silva
Instituto de Informática – UFG
Presidente da Banca

Prof. Doutor Eliomar Araújo de Lima
Instituto de Informática – UFG

Prof. Doutor Anderson da Silva Soares
Instituto de Informática – UFG

Prof. Doutor José Avelino Placca
Universidade Virtual do Estado de São Paulo

Eduardo Augusto Santos Garcia

Graduou-se em Engenharia de Computação na UFG - Universidade Federal de Goiás, durante sua graduação foi membro do Nucleo de Robótica Pequi Mecânico e pesquisador do Centro de Inteligência em Inteligência Artificial (CEIA).

Dedico este trabalho à minha família e aos meus amigos.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a realização deste trabalho.

Primeiramente, agradeço a minha orientadora, Prof. Dra. Nádia Félix Felipe da Silva e coorientador, Prof. Dr. Eliomar Araújo de Lima, por todo o suporte, orientação e conhecimento compartilhado ao longo desta jornada. Seu apoio foi fundamental para o desenvolvimento e aprimoramento desta pesquisa.

Aos membros da banca examinadora, Prof. Dr. Anderson da Silva Soares e Prof. Dr. José Avelino Placca, por aceitarem o convite e dedicarem seu tempo para avaliar este trabalho, contribuindo com valiosas sugestões.

Aos meus colegas do Centro de Excelência em Inteligência Artificial (CEIA), por todo o apoio, discussões enriquecedoras e colaboração durante o desenvolvimento deste estudo.

Agradeço também à Universidade Federal Goiás, por proporcionar a infraestrutura e os recursos necessários para a realização desta pesquisa.

Por fim, expresso minha eterna gratidão à minha família e amigos, por todo o amor, compreensão e incentivo ao longo desta trajetória. Seu apoio incondicional foi essencial para que eu pudesse superar os desafios e alcançar meus objetivos.

A todos vocês, meu mais sincero obrigado.

The law is a profession of words

David Mellinkoff,
The Language of the Law, 1963.

Resumo

Garcia, Eduardo. **Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora**. Goiânia, 2024. 82p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

Este trabalho investiga a aplicação do Processamento de Linguagem Natural (PLN) no contexto jurídico em língua portuguesa, com ênfase na importância da adaptação de domínio para modelos de linguagem pré-treinados, como o RoBERTa, a partir de conjunto de dados com documentos de domínio legal. Compilamos e pré-processamos um corpus jurídico português, denominado LegalPT, no qual abordamos os desafios da alta quantidade de quase duplicatas em *corpora* legais e realizamos uma comparação com *corpora* genéricos de raspagem da Web. Experimentos com esses dados revelaram que o pré-treinamento com dados jurídicos e gerais resultou em um modelo mais eficaz para tarefas jurídicas. O nosso modelo, denominado RoBERTaLexPT, superou arquiteturas maiores treinadas apenas em *corpora* genéricos, como o BERTimbau e Albertina-PT-*, e outros modelos jurídicos de trabalhos similares. Para a avaliação do desempenho desses modelos, propomos nesta dissertação de mestrado um *benchmark* jurídico composto por diversos conjuntos de dados, incluindo LeNER-Br, RRI, FGV, UlyssesNER-Br, CEIA-Entidades e CEIA-Frases. Este estudo contribui para aprimorar as soluções de PLN no contexto legal brasileiro, disponibilizando de forma aberta modelos aprimorados, um corpus especializado e um conjunto de *benchmark* rigoroso.

Palavras–chave

Processamento de Linguagem Natural, Modelo de Linguagem, Domínio Legal, Benchmark Jurídico.

Abstract

Garcia, Eduardo. **Legal Domain Adaptation in Portuguese Language Models - Developing and Evaluating RoBERTa-based Models on Legal Corpora**. Goiânia, 2024. 82p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

This research investigates the application of Natural Language Processing (NLP) within the legal domain for the Portuguese language, emphasizing the importance of domain adaptation for pre-trained language models, such as RoBERTa, using specialized legal corpora. We compiled and pre-processed a Portuguese legal corpus, named LegalPT, addressing the challenges of high near-duplicate document rates in legal corpora and conducting a comparison with generic web-scraped corpora. Experiments with these corpora revealed that pre-training on a combined dataset of legal and general data resulted in a more effective model for legal tasks. Our model, called RoBERTaLexPT, outperformed larger models trained solely on generic corpora, such as BERTimbau and Albertina-PT-*, and other legal models from similar works. For evaluating the performance of these models, we propose in this Master's dissertation a legal benchmark composed of several datasets, including LeNER-Br, RRI, FGV, UlyssesNER-Br, CEIA-Entidades, and CEIA-Frases. This study contributes to the improvement of NLP solutions in the Brazilian legal context by openly providing enhanced models, a specialized corpus, and a rigorous benchmark suite.

Keywords

Natural Language Processing, Language Model, Legal Domain, Legal Benchmark.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The rapid advancements in Natural Language Processing (NLP) have revolutionized the way we interact with and analyze text data across various domains. One such domain that has witnessed significant growth in NLP applications is the legal field [79]. Legal professionals deal with an overwhelming amount of textual information daily, ranging from legislation and jurisprudence to contracts and petitions. The highly technical and specialized nature of legal language presents unique challenges for NLP systems, necessitating the development of domain-specific solutions.

Pre-trained language models, such as BERT [24], have emerged as a powerful tool to tackle the complexities of natural language understanding. These models, trained on vast amounts of text data, capture intricate linguistic patterns and semantic relationships, enabling them to excel in a wide range of NLP tasks [63, 22]. However, the performance of these models can be further enhanced by adapting them to the specific characteristics and nuances of specialized domains, such as law.

Domain adaptation involves fine-tuning pre-trained language models on domain-specific corpora, allowing them to acquire the necessary knowledge and linguistic peculiarities of the target domain. This approach has proven effective in various fields, including law [17], biomedicine [39], and scientific research [7]. These domain-adapted models consistently outperform their generic counterparts in tasks specific to their respective fields.

Domain adaptation for language models has proven successful in several languages, particularly English [17, 78, 33, 18]. However, there is a notable gap in research for languages associated with distinct legal systems and particular requirements. This has motivated the research and development of legal language models in other languages, such as Arabic [2], Spanish [30], French [25], Italian [56], Romanian [45], and Chinese [73], which have also demonstrated significant performance improvements.

In the context of the Portuguese language, works such as [57, 66] have obtained positive results when training legal language models compared to the generic Portuguese BERT, known as BERTimbau [63]. However, these studies have often been limited to evaluations on a single legal NLP task, hindering the assessment of domain adaptation

benefits and effective model comparison.

The scenario of natural language processing in Portuguese in the legal domain is enriched by the availability of several public corpora, commonly extracted from documents of cases from Courts of Justice [47, 44, 70, 50], decisions and judgments from judges [70], and, more recently, the composition of a legal corpus that includes both the legislative and legal context, Ulysses-Tesemõ [27]. The availability of these large-scale corpora enables the pre-training of large domain-specific language models. However, a common challenge encountered in these works is the lack of explicit deduplication, a standard practice in generic corpora like BrWaC [67], CC100 [20], and MC4 [74]. Recent studies [40] highlight the potential performance improvements achieved by removing duplicate documents during model training.

This work aims to address these limitations within the scope of Portuguese legal NLP. The primary objectives are to train new pre-trained legal models for Portuguese, explore additional legal and legislative data sources (e.g., Bills[1], political speeches, political polls[2]), and provide more robust and representative models of the legal domain in Portuguese.

Additionally, this work conducts a comparative evaluation of pre-existing legal models on various benchmark datasets [43, 5, 21, 22]. This analysis provides insights into the relative performance of the models and identifies potential areas for improvement.

The research methodology involves experimenting with established architectures like RoBERTa [41], varying the training corpus and model hyperparameters. The process includes unsupervised pre-training on large legal corpora, followed by supervised fine-tuning for specific legal domain tasks.

The expected outcomes of this work include the availability of pre-trained language models and legal corpora for Portuguese, fostering advancements in Portuguese legal NLP. Furthermore, the comparative performance analysis of different models should generate valuable insights for future improvements, ultimately enhancing the quality and effectiveness of NLP solutions in the Brazilian legal context.

1.1 Hypotheses

This study proposes several hypotheses that aim to investigate the current state of legal NLP in Portuguese, the challenges and opportunities for evaluating legal language models, and the potential for developing high-performing, domain-specific models. These hypotheses are derived from the main themes and areas of investigation identified in the

[1] https://www.camara.leg.br/buscaProposicoesWeb/pesquisaSimplificada
[2] https://www.camara.leg.br/enquetes/

research and are designed to guide the experimental work and analysis throughout the dissertation.

H1. **Existing legal language models for Portuguese demonstrate varying levels of performance when evaluated on a diverse legal benchmark, highlighting the need for comprehensive evaluation frameworks.**

Previous studies on legal language models in Portuguese have been limited with training on relative small datasets and evaluating performance on a single task. This approach may lead to models that are overly specialized and fail to capture the breadth of legal language complexity. By creating a more diverse benchmark, we can highlight the potential limitations of these models in terms of generalizability.

H2. **Publicly available legal corpora in Portuguese exhibit significantly higher rates of document duplication compared to general-domain corpora used for language model pre-training.**

Legal documents often contain repetitive language, boilerplate clauses, and standardized formatting, which can lead to higher duplication rates compared to general-domain texts. Investigating this hypothesis is essential for understanding the unique characteristics of legal corpora and their potential impact on model training.

H3. **The performance of legal language models in Portuguese is significantly influenced by the deduplication of the pre-training corpus, the initialization strategy, and the selection of pre-training hyperparameters.**

The selection of the pre-training corpus is a critical factor in the development of legal language models, as it directly impacts the model's ability to capture the nuances and complexities of legal language. The decision to deduplicate the pre-training corpus may also play a role in the model's performance, as it can affect the efficiency and effectiveness of the training process. Additionally, the initialization strategy, such as using pre-trained weights from multilingual or domain-specific models, can influence the model's convergence and generalization capabilities. Furthermore, the choice of pre-training hyperparameters, including learning rate, batch size, and number of training epochs, can significantly impact the model's performance. Investigating these factors is crucial for developing high-performing legal language models in Portuguese that can effectively capture the intricacies of legal language and generalize well to a variety of legal NLP tasks.

H4. **Combining domain-specific legal corpora with generic corpora for pre-training can lead to improved performance of legal language models in Portuguese.**

While domain-specific corpora are crucial for capturing legal language nuances, they may be limited in size and diversity. Combining legal corpora with large,

generic corpora can potentially provide a balance between domain specificity and broad language understanding, leading to more robust and effective legal language models.

H5. **A domain-specific legal language model with a base configuration can outperform larger, generic language models on legal NLP tasks in Portuguese when pre-trained on a diverse and representative corpus.**
While larger models with billions of parameters have achieved state-of-the-art performance in many NLP tasks, they may not always be necessary or optimal for domain-specific applications. A well-trained, domain-specific model with a smaller architecture can potentially outperform larger, generic models by capturing the intricacies of legal language more effectively.

These hypotheses are grounded in the current understanding of legal NLP challenges, the potential of domain adaptation techniques, and the importance of comprehensive evaluation. By investigating these hypotheses, this study aims to contribute to the development of more effective and reliable legal language models in Portuguese, while also providing insights that can inform future research and applications in this field.

1.2 Objectives

This study aims to provide a comprehensive understanding of the current state and future potential of legal NLP in Portuguese by assessing the availability and diversity of legal datasets, investigating the prevalence of document duplication, evaluating existing models, and developing new domain-specific models. Ultimately, the objectives are to contribute to the development of high-performing, domain-specific legal language models for Portuguese and support the growth and advancement of legal NLP in this language:

1. Assess the availability and diversity of legal domain datasets in Portuguese and their implications for training and evaluating legal language models.
2. Evaluate the performance of existing legal language models for Portuguese on a comprehensive benchmark of legal datasets and tasks, including named entity recognition and text classification.
3. Investigate the prevalence of document duplication in publicly available legal corpora in Portuguese compared to general-domain corpora and analyze its potential impact on model training and performance.
4. Develop a domain-specific legal language model for Portuguese by pre-training on a corpus of legislative and legal texts and compare its performance to generic language models on the established benchmark.

5. Investigate the impact of pre-training techniques, such as deduplication of the corpus, initialization strategies, and others on the performance of legal language models in Portuguese, and develop best practices for corpus preparation in this domain.
6. Contribute to the advancement of legal NLP in Portuguese by openly sharing collections of datasets, model weights, and benchmarks developed throughout this study, enabling researchers and practitioners to build upon this work and drive further innovation in the field.

1.3 Contributions

The main contributions of this research are:

- The creation of the LegalPT corpus [28], the largest publicly available Portuguese legal corpus to date, comprising a diverse range of legal and legislative documents. This corpus serves as a valuable resource for pre-training legal language models and advancing legal NLP research in Portuguese.
- The development of the PortuLex benchmark, a comprehensive evaluation framework for legal language models in Portuguese. PortuLex consists of multiple datasets, including LeNER-Br [43], RRI [5], FGV [21], UlyssesNER-Br [22], CEIA-Entidades, and CEIA-Frases, covering various legal NLP tasks such as named entity recognition and text classification. This benchmark enables rigorous and standardized evaluation of legal language models, facilitating comparisons and identifying areas for improvement.
- Insights into the prevalence of document duplication in legal corpora and its potential impact on language model pre-training and performance. This study compares the duplication rates of the LegalPT corpus with the generic CrawlPT corpus, highlighting the importance of deduplication techniques in corpus construction and preprocessing.
- A rigorous evaluation of existing legal language models for Portuguese, including BERTimbau [63], Albertina-PT-BR [61], BERTikal [57], JurisBERT [66], and Legal-XLM-R [50], on the PortuLex benchmark. This evaluation establishes a baseline for model performance and identifies strengths and weaknesses of current approaches, guiding future research and development efforts.
- The introduction of RoBERTaLexPT [28], a high-performing, domain-specific legal language model for Portuguese. RoBERTaLexPT is pre-trained on a combination of the LegalPT and CrawlPT corpora, demonstrating the benefits of combining domain-specific and generic corpora for legal language model training. Despite its

base configuration, RoBERTaLexPT outperforms larger, generic language models on the PortuLex benchmark, highlighting the effectiveness of domain adaptation techniques.

- Empirical evidence on the impact of pre-training corpus selection, deduplication, initialization strategy, and hyperparameter optimization on the performance of legal language models in Portuguese. This study investigates the effects of these factors on the performance of RoBERTa-based models [41], providing valuable insights for the development of high-performing legal language models.

The resources developed in this research, including the LegalPT corpus, PortuLex benchmark, and RoBERTaLexPT model, are made publicly available on the GitHub repository[3]. This ensures that the contributions of this work are accessible to the research community, fostering further advancements and collaborations in the field of Portuguese legal NLP.

1.4 Dissertation Structure

This dissertation is organized into six chapters, each addressing the research questions and objectives. Chapter 2 provides the theoretical foundation, introducing key concepts and theories in NLP and pre-trained language models. Chapter 3 presents a thorough review of related works in legal NLP, focusing on Portuguese and other languages, and highlighting the current state of the art, challenges, and gaps in the literature.

Chapter 4 describes the methodology used for model training and evaluation, covering data collection, preprocessing, model architectures, hyperparameters, and the evaluation framework. Chapter 5 presents and analyzes the experimental results, including the assessment of legal dataset availability and diversity, the investigation of document duplication, the evaluation of existing legal language models, and the development and testing of new domain-specific models.

Finally, Chapter 6 concludes the dissertation by summarizing the main contributions and insights, discussing the limitations and potential future directions, and highlighting the relevance of the work for the field of legal NLP in Portuguese and beyond.

[3] https://github.com/eduagarcia/roberta-legal-portuguese

CHAPTER 2

Theoretical Foundations

This chapter provides a theoretical basis for understanding models based on the *Transformer* architecture [65] such as BERT [24]. Initially, the history of deep learning techniques and models in NLP are discussed in section 2.1 along with the presentation of the *Transformer* architecture 2.1.1, the BERT architecture and other language models are presented in section 2.2, and finally the methods used to evaluate these language models in 2.4.

2.1 *Deep Learning Architectures* in NLP

Deep learning has revolutionized the field of Natural Language Processing (NLP) by enabling the development of powerful models that can effectively process, understand, and generate human language. These models have achieved state-of-the-art performance on a wide range of NLP tasks, such as text classification, sentiment analysis, machine translation, and question answering [76, 52].

The success of deep learning in NLP can be attributed to its ability to automatically learn meaningful representations of text data. Traditional NLP approaches relied heavily on hand-crafted features and rules, which were time-consuming and often limited in their ability to capture the complexities of language. In contrast, deep learning models can learn hierarchical representations of text by processing raw input data through multiple layers of neural networks [55].

One of the key challenges in applying deep learning to NLP is the sequential nature of text data. Unlike images or tabular data, text is composed of a sequence of words or characters that have a specific order and context. To effectively process and understand text, deep learning models need to capture the dependencies and relationships between the elements in the sequence [29].

Early neural network architectures, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [34], were designed to handle sequential data by maintaining hidden states that capture the history of the sequence. However,

these architectures faced challenges such as the vanishing gradient problem [8] and limitations in handling long sequences.

The introduction of the attention mechanism by [6] addressed some of these limitations by allowing the model to focus on the most relevant parts of the input sequence. This mechanism enabled the model to assign different weights to each input token, thereby capturing the importance of each token in the context of the entire sequence.

In this era of information and data, models based on deep learning neural networks (*Deep Learning*) have become the state of the art in Natural Language Processing. With the growing popularity of word vectors, neural network models have been able to achieve very high performance on various NLP tasks. One of the most successful neural network-based architecture in this field is discussed in the Section 2.1.1.

2.1.1 Transformers

The Transformer architecture, proposed by [65], revolutionized the field of NLP by providing an alternative to the encoder-decoder architectures of recurrent networks. The Transformer follows an encoder-decoder structure but does not contain any recurrent mechanism. Instead, it relies heavily on the attention mechanism to attend to information from any part of the input sequence or the generated hidden states.

The key component of the Transformer is the Self-Attention layer, which allows each input word to attend to every other word in the sentence. In this layer, each input word is multiplied by three matrices: query Q, key K, and value V, generating corresponding vectors. These vectors are used to produce a score that represents the relevance of each word to the others in the sentence. The attention of a token is then calculated as the weighted sum of the value vectors, where the weights are determined by the softmax function of the dot product between the query and key vectors.

Intuitively, the self-attention mechanism enables the model to capture the relationships and dependencies between the elements in the sequence more effectively than previous architectures. By attending to different parts of the input sequence simultaneously, the model can weigh the importance of each element based on its relevance to the others. This allows the model to capture long-range dependencies and understand the context of the entire sequence.

To capture different relationships between the parts of the sequence, the Transformer employs multiple Self-Attention "heads" that operate in parallel. The outputs of these heads are concatenated and transformed using a feed-forward network to generate the final representation.

Since the Self-Attention process does not inherently capture positional information, the Transformer introduces a Positional Encoding layer. This layer injects information about the order of the words in the sequence by adding position-specific vectors to the input embeddings. These positional encodings are generated using sine and cosine functions with different frequencies, allowing the model to learn the relative position of each token in the sequence. The whole architecture is represented on figure 2.1.

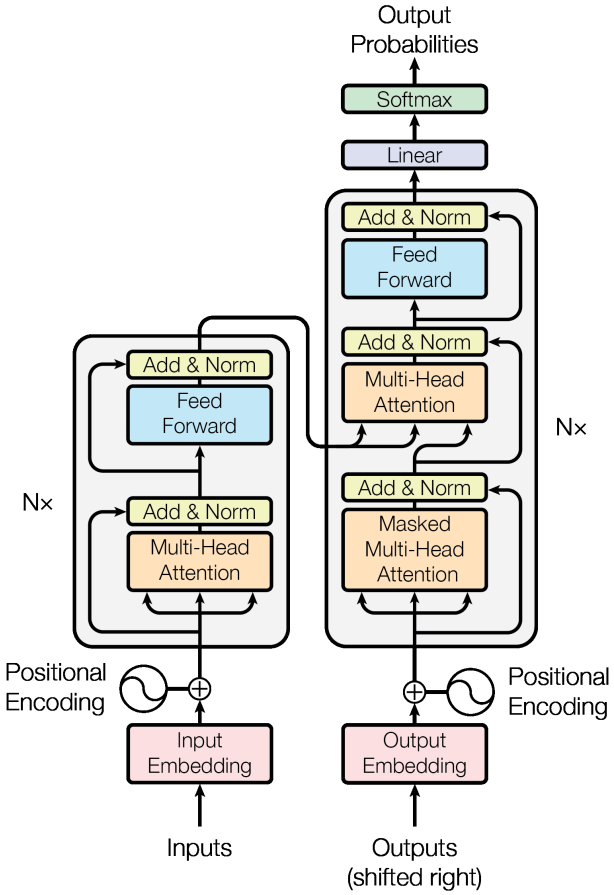

Figure 2.1: *Encoder-decoder architecture of the Transformer model.*
Source: [65]

By combining the two ideas of Self-Attention and *Positional Encoding*, Transformers circumvent the backpropagation bottleneck over time faced during the training of RNNs, since the hidden state of each position in the sequence can be calculated simultaneously, allowing for massive parallelization and enabling faster computation. The Transformer architecture was the inspiration for the GPT model by [58] and the BERT model by [24]. While the BERT model uses the encoder submodule of the Transformer model, OpenAI's GPT uses only the decoder submodule of the Transformer architecture. These two models form the basis of the most modern language models being developed today.

2.2 Language Models

Language models formalize the intuition of assigning probabilities to sentences in NLP [9]. Given a sequence of tokens, $s = (w1, w2, ..., w_N)$, a language model estimates the following probability:

$$P(s) = P(w_1, w_2, ..., w_N) \tag{2-1}$$

The calculation of such probability assumes a discrete set of tokens belonging to the language, often known as the language's vocabulary $\mathcal{V}$, such that $\forall i, w_i \in \mathcal{V}$. For example, the vocabulary of the Portuguese language can be represented as the set: $\mathcal{V}_{portuguese} = \{a, aacima, ..., zwingliano\}$. By the chain rule of probabilities, Formula 2-1 can be rewritten considering the conditional probability of a token w_i given the sequence of tokens that preceded it:

$$\begin{aligned} P(s) &= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times ... \times P(w_n|w_{1:N-1}) \\ &= P(w_1) \prod_{i=1}^{N} P(w_i|w_{1:i-1}) \end{aligned} \tag{2-2}$$

The quality of a language model can be evaluated using two classes of evaluation methods. The first class of methods is collectively referred to as extrinsic methods, in which the language model is evaluated as a component of a system in terms of its effectiveness in a higher-level task. For example, one can measure the change in translation quality of a machine translation system when language model A is replaced by another language model B. The second class of evaluation methods are known as intrinsic and, as the name suggests, measure the quality of language models in their natural environment, regardless of any external application, such as the *Perplexity* metric, which indicates how good a language model (LM) is at predicting an unknown sample [15]. Calculated as:

$$Perplexity = 2^{-\frac{1}{n}\sum_{i=1}^{n} \log_2 LM(w_i|w_{1:i-1})} \tag{2-3}$$

It is important to note that the perplexity measure is strongly linked to the corpus used as the training and test set, therefore, two language models can only be faithfully compared when they are trained on the same corpus [26]. Furthermore, if the corpus is not a representative collection of sentences in the given language, then the language model is expected to not generalize well to test sets.

2.2.1 Pre-Trained Models

The ideas discussed in Section 2.1 describe a family of models that can be trained to behave as a Language Model using unsupervised training from a corpus to learn the probabilistic properties of a language [58] - which sequences are likely, which word follows a given sentence, etc. The result is that these models tend to produce vector representations for words that show semantic regularities in how they are encoded. These word vectors serve as ideal initial states for models that perform more complex natural language processing tasks [55], such as Question Answering [59], Natural Language Inference [12, 69], among others. Instead of starting from scratch, these models are able to learn from already trained representations that provide them with a basic "prior" knowledge about the semantic distribution properties of the inputs. This technique of extracting already trained or "pre-trained" word representations and applying them as initial states in NLP tasks where a word's vector corresponds to a collection of its characteristics has become the state-of-the-art technique in NLP.

The ULMFiT model developed by [35] was the first to refine the *fine-tuning* technique, an approach that "transfers" knowledge using entire models instead of a single embedding layer. Its general structure has three stages: (1) Pre-training of the language model, which trains a neural language model on a large corpus such as Wikipedia; (2) *Fine-tuning* of the language model, where the neural model adapts to the domain relevant to the task, for example, movie reviews; and (3) *Fine-tuning* of the classifier, where the representations are finally updated to perform the actual classification task, for example, predicting the sentiment of movie reviews. Therefore, instead of transferring pre-trained knowledge only in the initial layer of the model in the form of word embeddings, an entire language model is used to first learn generalized language representations and then fine-tuned to adapt to the domain of its task. According to [35], the UMLFiT model trained with only 100 labeled examples corresponds to the performance of a model trained with 100x more data from scratch, showing the power of knowledge transfer from pre-training.

Pre-trained language models followed by a *fine-tuning* structure have been producing state-of-the-art results in a wide variety of NLP tasks, inspiring models such as GPT [58], leading to the development of BERT [24], models that brought a radical change in the field. The advent of BERT made *fine-tuning*-based approaches a basic component of the modeling process in Natural Language Processing tasks.

2.3 BERT

Bidirectional Transformer Encoder Representations (BERT) is a model based on bidirectional Transformer encoders developed by [24]. This model was designed to

pre-train deep bidirectional representations to extract context-sensitive features from the input text. Previously, models like ELMo proposed by [55] could perform a bidirectional scan of the text with biLSTMs, where the sequence was trained from left to right and from right to left, thus extracting bidirectional vectors. But such models were unable to capture the contextual information that was possible by attention models. [24] called the biLSTM-based models "shallow bidirectionality" due to their nature of reading in only one direction at a time, while BERT would have a "deep bidirectionality" due to the Transformers' Self-Attention mechanism capable of reading all tokens simultaneously.

There are previous works of other Transformer-based models, such as OpenAI GPT [58], which used Self-Attention to capture the context of the sequence. But this model was able to capture the context between layers in only one direction, as the model read the sequence from left to right, due to the training objective of only trying to predict the next word. This means that this model is not bidirectional.

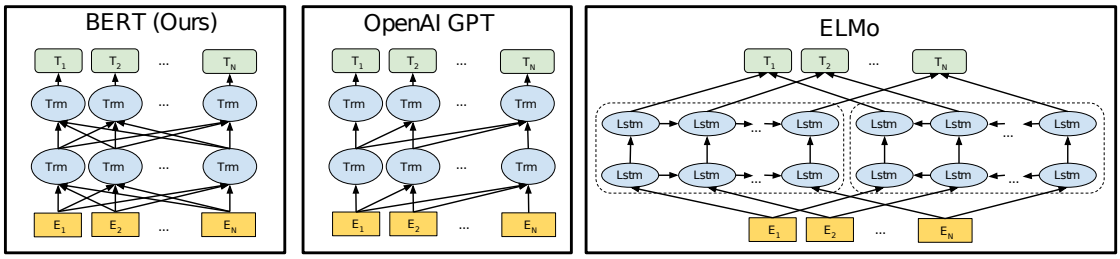

Figure 2.2: *BERT in comparison to other models.*
Source: [24]

2.3.1 Wordpiece Model

BERT provides a specific predefined *tokenization* model called Wordpiece [72] that generates tokens from a sentence after a pre-processing routine. The pre-processing includes separating all punctuation in the sequences and adding spaces on both sides of the words after removing whitespace. Wordpiece breaks certain words into sub-words for token assembly. The following is an example of a sentence and the corresponding sentence generated by the wordpiece model:

- **Sentence:** An encyclopedia is a work that brings together part of human knowledge or just one domain of it.
- **Wordpiece:** An ency ##clo ##pe ##dia is a work that brings together part of human knowledge or just one domain of it .

In the example, the word "encyclopedia" was broken into 4 sub-words "ency", "##clo", "##pe" and "##dia", the other words remained as single tokens. "##" is a special additional marker to represent word breaks.

The Wordpiece model is trained on a corpus with the objective of generating a fixed-size vocabulary. In the BERT model trained on the English language, the Wordpiece model consists of a vocabulary of 30,000 tokens. In addition to the standard sub-word generation of the model, tokens that have only one or two characters are manually added to the vocabulary. The reason for including these small tokens is that when an unknown word is encountered by the tokenizer, it completely separates the word into individual letters and maps it to its corresponding token. Suppose that when a word like "zxas" is encountered by the tokenizer, it is then divided into tokens like [z], [##x], and [##as]. This technique ensures that words that are not in the vocabulary are not just categorized as general unknown words, but receive tokens at least in the format of sub-word blocks.

2.3.2 Architecture Details

BERT was published in two sizes called: *BERT Base* and *BERT Large*. BERT Base has a similar size to the original Transformer model [65] for performance comparison, while BERT Large is a larger model that has twice as many Transformer layers and obtained state-of-the-art results at the time [24].

Both models are composed of a large number of stacked Transformer encoder blocks, 12 for the Base version and 24 for the Large version, each of these encoders has a Multi-Head Self-Attention layer, 12 and 16 heads respectively. They also have a larger feed-forward network (768 and 1024 units respectively) compared to the original Transformer implementation [65] (6 encoder layers, 8 Self-Attention heads, and 512 units in the feed-forward network).

The model training was done with 512 input tokens, this value is fixed, if the sentence is longer it is truncated and if shorter it is filled with the special token [PAD] in each training batch. Longer sequences are disproportionately more expensive to train because the attention computation grows quadratically with the sequence length.

In the last layer of the BERT model, for each position or token, a vector of the size of the Transformer block's feed-forward network is generated, 768 for BERT-Base and 1024 for BERT-Large, this vector would be the embedding corresponding to that token. For the classification task, for example, we use the first vector which always corresponds to the special token [CLS].

2.3.3 Pre-Training

In several NLP frameworks, including the previously mentioned ELMo [55] and OpenAI GPT [58] architectures, Language Model pre-training has proven to be useful for various tasks [35]. The training process is performed by trying to predict the next word in a text sequence, this requires the training to be performed sequentially, from right to

left or from left to right, through the series. The ELMo and Open AI GPT architectures operate in this way.

The Masked Language Model employed by the BERT model is a reformulation of the language modeling task in which a certain percentage of words in the context are "masked" and the model's task is to maximize the probability of these words in place of the masked positions. In the BERT model training, 15% of the tokens from each input are masked, replacing the missing word with a [MASK] token. This component of the modeling process imposes the bidirectional properties within BERT, which together with the attention mechanism provided by the transformer block scores, conditions the model to need to fully understand the entire context of a word in order to produce its representation.

However, as a result of word masking, there is a mismatch between pre-training and fine-tuning. The model is only trained to predict the special masked token [MASK] in pre-training, but these tokens are not present during fine-tuning tasks. The authors overcome this problem by using the following masking procedure:

- Replace the word with the [MASK] token for 80% of the data.
- Replace the word with another randomly chosen token for 10% of the data.
- Keep the original word intact and do nothing for the remaining 10% of the data.

The model is then able to predict alternative tokens in addition to the masked token, as it maintains a contextual representation of the distribution of each input token and not just the masked token.

In addition to the *Masked Language Model* task, motivated mainly by the fact that several NLP tasks require modeling the relationship between two sentences such as Natural Language Inference and Reading Comprehension, [24] decided to jointly train BERT to also estimate whether the second sentence in its input pair follows the first. This task was called Next Sentence Prediction. The authors empirically showed that this additional technique leads to performance gains in downstream tasks.

During training, BERT accepts sequences of the form:

[CLS] This is sentence A. [SEP] This is sentence B. [SEP],

Where [CLS] is a special token whose representation is used for fine-tuning in classification tasks and [SEP] is a special token indicating sentence separation or termination. There is a 50% possibility of sentence A and B corresponding to two consecutive sentences from the training corpus, and another 50% of being completely random.

2.3.4 Pre-Training Data

The English BERT model was trained on about 16 GB of textual data. Both models were pre-trained by the authors using a *batch-size* of 256 sequence pairs for 40 epochs over more than 4 days on 4 Cloud TPUs Pods (BERT-Base) and 16 Cloud TPUs Pods (BERT-Large).

There are high environmental costs in pre-training these massive models. According to [64]. Pre-training the GPU-based English BERT model emits CO2 at a level similar to transamerican flight levels.

2.3.5 RoBERTa

Robustly Optimized BERT Pretraining Approach (RoBERTa) is a study developed by [41] that analyzes the impact of various hyperparameters and training data chosen by the BERT model. They conclude that the original model was severely undertrained and they propose a model that surpasses it in several NLP tasks.

The modifications made to the BERT model by [41] include: (1) training the model for a longer time and with a larger *batch-size* (from 256 to 8192), (2) removal of the *Next Sentence Prediction* training objective due to its ineffectiveness, (3) dynamically changing the model's masking pattern during training at each epoch which was fixed in the BERT Model, and (4) using more data for training.

The work shows that each of these modifications leads to an improvement in the model's performance on several subsequent NLP tasks. The RoBERTa model was trained with 160GB of textual data (compared to 16GB for BERT) on 1024 Nvidia V100 GPUs for about a day.

2.4 Natural Language Processing Tasks

2.4.1 Text Classification

Text classification is a task in Natural Language Processing (NLP) that consists of assigning a set of predefined categories to a given text document. The objective is to automatically determine the most appropriate category or categories for an input text based on its content. [32]

Formally, given a set of text documents $D = \{d_1, d_2, ..., d_n\}$ and a set of predefined categories $C = \{c_1, c_2, ..., c_m\}$, the goal of text classification is to assign each document d_i to one or more categories from the set C. The classification can be binary (assigning a document to one of two categories) or multi-class (assigning a document to one of multiple categories).

Text classification has a wide range of applications across various domains. Some examples include:

- **Sentiment Analysis**: Classifying the sentiment of a text as positive, negative, or neutral. This is commonly used in social media monitoring, customer feedback analysis, and product reviews [14].
- **Spam Detection**: Identifying whether an email or message is spam or not. This helps in filtering out unwanted or malicious content [23].
- **Language Identification**: Determining the language in which a text document is written. This is important for multilingual text processing and machine translation [42].
- **Legal Document Classification**: Categorizing legal documents into predefined categories, such as contracts, case law, or patents. This assists in organizing and retrieving relevant legal information efficiently [44].

Text classification models are typically trained on labeled datasets, where each document is associated with its corresponding category or categories. The training process involves extracting relevant features from the text, such as word frequencies, n-grams, or word embeddings, and learning the patterns and characteristics that distinguish different categories.

Evaluation of text classification models is commonly performed using metrics such as accuracy, precision, recall, and F1-score. These metrics measure the model's ability to correctly assign documents to their respective categories and provide insights into its performance.

2.4.2 Named Entity Recognition

The task of Named Entity Recognition (NER) consists of extracting entities from a text and classifying them into a set of categories of interest [62]. Examples of general domain entities: names of people, organizations, and locations; in the biomedical domain, categories can include: names of medications, diseases, expiration dates, among others. NER is typically modeled as a sequence classification task. Given an input of n tokens $(x_1, x_2, ..., x_1)$, the objective is to assign each of them a category y, including a null category $(y_1, y_2, ..., y_n)$.

A widely used modeling format for the NER task is the so-called IOB (*Inside–outside–beginning*), initially presented by [60], where the prefixes I, O, and B are added at the beginning of each category to indicate when it starts and when it ends.

In the IOB2 format, an extension of IOB, the prefix B is used to indicate the beginning of a category, the prefix I indicates that the category continues in the next item

Bruno	B-Person
is	O
going	O
to	O
São	B-Location
Paulo	I-Location

Table 2.1: *Example of an IOB2 labeling*

of the sequence, and finally, the prefix *O* indicates that the item does not belong to any category.

As an example, let's try to classify the sentence "Bruno is going to São Paulo" with the entities "Person" and "Location". If we consider that each word in this sequence corresponds to a token, "Bruno" is classified as a person, "São Paulo" as a location, and the rest of the sentence does not have any other entities. The conversion to IOB2 is shown in Table 2.1.

Therefore, the task of an NER model with the two entities "Person" and "Location" can be defined as the classification of each token into the following categories: O, B-Person, I-Person, B-Location, and I-Location.

2.4.3 Evaluation Metrics

In this subsection, we will present the formal definition of metrics used in the work to evaluate the performance of models on a classification task, including NER [32].

A binary classification model assigns items to only two classes: positive and negative. We can then define that a model can have the following results:

- **True Positives (TP)** is a result where the model correctly predicts the positive class.
- **True Negatives (TN)** is a result where the model correctly predicts the negative class.
- **False Positives (FP)** is a result where the model incorrectly predicts the positive class.
- **False Negatives (FN)** is a result where the model incorrectly predicts the negative class.

Accuracy is defined as the rate of correctly predicted results in relation to the total number of inputs, where the highest possible value is 1 for a perfect classification and the lowest is 0 for a total failure:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-4}$$

Another widely used metric to evaluate the performance of a model on a classification task is the *F1-Score*, which is defined as the harmonic mean between precision and recall. Recall is also known as sensitivity in the literature.

Given a binary classification problem, precision measures the percentage of correctly classified entries (positive class) present in a set of entries classified as positive, while recall measures the percentage of correctly classified entries in the positive class.

Precision is the proportion of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2-5}$$

Recall is the proportion of correctly predicted positive observations to all observations in the positive class.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2-6}$$

F1-Score is an overall measure of a model's performance that combines precision and recall. In other words, a good F1 score means you have few false positives and false negatives. An F1 score is considered perfect when it is 1, while the model is a total failure when it is 0.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2-7}$$

For a multiclass problem, we have to consider different weighting modes of the F1-Score. The most commonly used methods are called *micro* and *macro*. Macro calculates the metric independently for each class and then takes the average, treating all classes with the same weight. Micro aggregates all the hits and misses of each class to calculate the metric, which is interesting if there is an imbalance of classes in the dataset, as classes with few entities will affect the metric proportionally. For a number of classes n, the calculation of these modes is done as follows:

$$Precision_{macro} = \frac{\sum_{i=1}^{n} Precision_i}{n} \tag{2-8}$$

$$Recall_{macro} = \frac{\sum_{i=1}^{n} Recall_i}{n} \tag{2-9}$$

$$F1_{macro} = 2 * \frac{Precision_{macro} * Recall_{macro}}{Precision_{macro} + Recall_{macro}} \tag{2-10}$$

$$Precision_{micro} = \frac{\sum_{i=1}^{n} \mathrm{TP}_i}{\sum_{i=1}^{n} \mathrm{TP}_i + \mathrm{FP}_i} \tag{2-11}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{n} \mathrm{TP}_i}{\sum_{i=1}^{n} \mathrm{TP}_i + \mathrm{FN}_i} \tag{2-12}$$

$$F1_{micro} = 2 * \frac{Precision_{micro} * Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{2-13}$$

CHAPTER 3

Related Works

This chapter provides an overview of the relevant research and developments in the field of legal natural language processing (NLP), focusing on domain adaptation techniques, legal language models, and available corpora. The chapter begins by introducing the concept of domain adaptation and its application in various specialized domains. It then narrows down to the legal domain, discussing the strategies and challenges associated with adapting language models to legal texts. The focus then shifts to the Portuguese legal domain, presenting notable initiatives and models developed for this specific context. The chapter also explores the available legal corpora in Portuguese, categorizing them into pre-training and fine-tuning corpora. Finally, the importance of deduplication techniques in creating high-quality language modeling datasets is discussed, highlighting their impact on model performance and data privacy.

3.1 Domain Adaptation

Pre-trained language models like BERT have achieved great success in NLP tasks. However, when applied to specialized domains such as legal, biomedical, or scientific, they tend to have inferior performance, as they were pre-trained on generic corpora like Wikipedia. To improve performance in these domains, two strategies are possible: further pre-training the BERT model on domain-specific corpora, or pre-training BERT from scratch using domain corpora. Several studies, such as [7, 39, 75, 17, 36], have shown that these two approaches lead to significant improvements in specialized NLP tasks, outperforming the direct use of pre-trained BERT. For example, BioBERT for the biomedical domain, SciBERT for the scientific domain, and Legal-BERT for the legal domain.

Beltagy et al. [7] compared pre-training BERT from scratch on scientific corpora versus further pre-training on 1.14 million scientific articles. Both approaches improved performance on 5 downstream scientific NLP tasks compared to the original BERT. The specific corpus allowed for customized vocabulary, but the gains were similar. Lee et al. [39] further pre-trained BERT on 18 billion tokens of biomedical texts. The resulting

Model	Task	Gain over BERT-Base
SciBERT	Classification	+11.1% F1-Score
	Relation between two sentences	+3.7% Accuracy
	Named Entity Recognition	+2.7% F1-Score
BioBERT	Question Answering	+12.2% MRR
	Relation between two sentences	+2.8% F1-Score
	Named Entity Recognition	+0.6% F1-Score
FinBERT	Sentiment Analysis	+8.1% Accuracy
LEGAL-BERT	Named Entity Recognition	+1.5% F1-Score
	Classification	+1.2% F1-Score

Table 3.1: *The reported performance gain of domain adaptation models in English compared to common BERT-Base.*

BioBERT outperformed the original BERT on 3 downstream biomedical NLP tasks, with gains of up to 2.8% in F1. The size of the corpus and domain adaptation were crucial.

Yang et al.[75] pre-trained FinBERT on 4.9 billion tokens of financial texts. FinBERT outperformed BERT by up to 15% accuracy on 3 financial sentiment analysis tasks. A large and specialized corpus brought significant improvements. Chalkidis et al. [17] performed additional pre-training of BERT on 12GB of legal texts from various sources. Adapted models had modest gains of 0.2-2.5% F1 on 3 legal tasks. Pre-training on specific subdomains had more impact. Huang et al. [36] further pre-trained BERT on clinical texts. The resulting ClinicalBERT obtained significant gains in predicting hospital readmission compared to other methods. Pre-training on a relevant domain is crucial.

In summary, the results obtained by these works can be found in Table 3.1. The studies show that pre-training BERT on a large corpus specialized in the downstream domain leads to improvements in task performance. Both pre-training from scratch and further pre-training are effective. Smaller and more efficient BERT models can also perform well in specialized domains, so it is not always necessary to use the largest models. In general, adapting BERT models through pre-training on a specialized domain leads to significant improvements in NLP tasks in those domains.

3.2 Adaptation for Legal Domain

Models like BERT, initially trained on general corpora, often struggle to perform well on specialized domains like legal texts. Researchers have explored strategies such as using the original BERT, adapting BERT by additional pre-training, or pre-training BERT from scratch on domain-specific corpora [17]. These approaches have shown promising results in improving the performance of language models on legal NLP tasks.

Several studies have investigated the conditions under which domain pre-training can help in legal NLP tasks. The work on CaseHOLD [78] introduced a new dataset and

found that domain pre-training using a large corpus of legal decisions led to substantial performance gains on CaseHOLD and other legal NLP datasets. They also showed that the level of performance increase was directly tied to the domain specificity of the task. Similarly, the study on legal argument mining [77] explored the potential of domain pre-training to help interdisciplinary researchers facing data annotation poverty, demonstrating that domain pre-training can enhance transformer performance on interdisciplinary tasks with limited annotated data.

The development of specialized legal corpora has played a crucial role in facilitating the adaptation of language models to the legal domain. The Pile of Law [33], a large open-source dataset of English-language legal and administrative data, was introduced to discuss the legal norms that governments have developed to constrain the inclusion of toxic or private content. LeXFiles and LegalLAMA [18], a multinational English legal corpus and a legal knowledge probing benchmark, were released to facilitate training and analysis of legal-oriented language models. Additionally, MultiLegalPile [50], a multilingual legal corpus in 24 languages from 17 jurisdictions, was curated and released, demonstrating its effectiveness in pre-training RoBERTa and Longformer models for legal NLP tasks.

The adaptation of language models to the legal domain has also been explored for languages other than English. Lawformer [73], a Longformer-based pre-trained language model for Chinese legal long documents understanding, was released and evaluated on various legal AI tasks, demonstrating promising improvements on tasks with long documents as inputs. A Spanish legal language model [30] was generated using gathered legal-domain corpora and evaluated against Spanish general domain tasks, showing reasonable results. JuriBERT [25] focused on creating a language model adapted to French legal text, proving that domain-specific pre-trained models can perform better than their equivalent generalized ones in the legal domain. Similarly, AraLegal-BERT [2], a bidirectional encoder Transformer-based model for the Arabic legal domain, achieved better accuracy than the general and original BERT over legal text.

In summary, the adaptation of pre-trained language models to the legal domain through domain-specific pre-training and the development of specialized legal corpora has consistently shown to improve the performance of these models on various legal NLP tasks. The studies in this field highlight the importance of domain adaptation in enhancing the effectiveness of language models in specialized domains like law, and provide valuable insights and resources for researchers and practitioners working on legal NLP applications.

3.3 Language Models for Portuguese Legal Domain

Several initiatives aimed at facilitating the use of natural language processing tools for legal text analysis in the Brazilian context. The MultiLegalPile [50], a multilingual legal corpus, includes a Portuguese subset and was used to pre-train two multilingual legal models called Legal-XLM-R-base and Legal-XLM-R-large, as well as a single Portuguese model called Legal-mono-R-base. However, in their benchmark, all models failed to improve results on Portuguese tasks compared to XLM-R-large, but they did improve in relation to XLM-R-Base and BERTimbau-base. These findings suggest that while domain-specific pre-training can lead to improvements over base models, it may not always surpass the performance of larger, general-purpose models.

LegalNLP [57] and BertBR [19] are examples of pre-trained language models specifically designed for the Brazilian legal language. LegalNLP provides a set of models (Phraser, Word2Vec, Doc2Vec, FastText, and BERT) along with a Python package and tutorials, while BertBR was created by further pre-training a Portuguese BERT model using legal texts, achieving an F1-Score of 94.39% in named entity recognition.

JurisBERT [66] and the Semantic Search System for the Supremo Tribunal de Justiça [46] are other notable examples of transformer-based models adapted to the Portuguese legal domain in the context of Information Retrieval. JurisBERT, trained from scratch using domain-specific texts, outperformed other BERT models in Semantic Textual Similarity (STS) for legal texts, while the Semantic Search System combined lexical and semantic techniques using Legal-BERTimbau variants, resulting in improved performance compared to using BM25 alone.

Despite the promising results, these approaches have limitations. The corpora used for pre-training are often from a single source [19, 66, 46], which may not comprehensively cover all aspects of legal language. Consequently, the models may struggle with specific legal tasks or subdomains. Additionally, the evaluation of these models is often limited to a small set of tasks or benchmarks, making it difficult to assess their generalizability.

These limitations highlight the need for more diverse and representative legal corpora in Portuguese, as well as comprehensive evaluation frameworks that can assess the performance of legal language models across a wide range of tasks and subdomains. Addressing these challenges will be crucial for the development of robust and effective legal NLP solutions in the Portuguese language.

3.4 Available Legal Corpora

The availability of large, high-quality datasets is crucial for training language models in specialized domains such as law. In the context of the Portuguese language, several corpora have been developed to support the pre-training and fine-tuning of legal language models.

3.4.1 Pre-training Corpora

One notable corpus for pre-training is the MultiLegalPile [50], a multilingual corpus containing a Portuguese subset with 92GiB of data from sources such as the jurisprudence of the Court of Justice of São Paulo (CJPG), appeals from the 5th Regional Federal Court (BRCAD-5) [47], and legal documents from the European Union (EUR-Lex). This diverse collection of legal texts provides a rich resource for training language models to understand the nuances of legal language in Portuguese.

Another important resource is the Ulysses-Tesemõ corpus, which consists of 2.2 million documents from 96 different legal, legislative, academic, and news sources in Brazilian Portuguese. This corpus offers a broad range of text types, enabling language models to capture the varied linguistic patterns found in legal and related domains.

The ParlamentoPT corpus [61] contains transcriptions of debates in the Portuguese Parliament, providing insights into the language used in legislative discussions. This corpus allows language models to learn the specific vocabulary and discourse patterns associated with parliamentary proceedings.

Additionally, the Iudicium Textum dataset [70] comprises rulings, votes, and reports from the Brazilian Supreme Federal Court (STF) published between 2010 and 2018. This dataset focuses on the highest court in Brazil, offering language models the opportunity to learn from the language used in supreme court decisions and related documents.

These corpora offer a diverse range of legal texts, enabling language models to capture the intricacies of Portuguese legal language during pre-training. By training on such extensive and varied datasets, language models can develop a deep understanding of legal terminology, sentence structures, and domain-specific patterns, which is essential for their effective application in legal NLP tasks.

However, it is important to note that these corpora may have limitations in terms of their coverage of all aspects of legal language, as they are often sourced from specific courts or institutions. Additionally, the quality and consistency of the data may vary across different sources, which can impact the performance of the pre-trained models. Future research should focus on expanding the range of legal corpora available for Portuguese,

as well as developing standardized preprocessing and quality control methods to ensure the reliability and representativeness of the training data.

3.4.2 Fine-tuning Corpora

For fine-tuning legal language models, several datasets have been developed for specific tasks such as named entity recognition (NER) and document classification. The LeNER-Br dataset [43] is a corpus for NER in Brazilian legal documents, containing tags for persons, locations, time entities, organizations, laws, and legal cases.

The RRI-PT corpus [5] focuses on rhetorical role identification in Portuguese legal petitions, providing a new set of rhetorical roles tailored for this document type. The FGV dataset [21] presents a case study on fine-grained legal entity annotation in decisions published by the Brazilian Supreme Court (STF), with two levels of nested legal entities annotated by law students.

In addition to these corpora, the UlyssesNER-Br corpus [3, 22] consists of bills and legislative consultations from the Brazilian Chamber of Deputies, annotated for NER. This corpus has been expanded with informal user-generated text in the form of comments about bills, allowing for the analysis of the impact of combining formal and informal texts from the same domain on NER performance.

For document classification tasks, the MultiEURLEX dataset [16] provides a multilingual and multi-label dataset of European Union laws, officially translated into 23 languages and annotated with labels from the EUROVOC taxonomy. This dataset serves as a testbed for zero-shot cross-lingual transfer learning in legal document classification.

These fine-tuning corpora enable the adaptation of pre-trained legal language models to specific tasks, such as NER and document classification, further enhancing their performance in practical applications within the legal domain.

However, the availability of fine-tuning corpora for Portuguese legal NLP tasks is still limited compared to other languages, such as English. Many of the existing datasets are focused on specific courts or document types, which may limit their generalizability to other legal subdomains. Additionally, the size of these datasets is often relatively small, which can impact the performance of the fine-tuned models. Future research should aim to develop larger and more diverse fine-tuning corpora for Portuguese legal NLP, covering a wider range of tasks and legal subdomains. This will enable the development of more robust and versatile legal language models that can be applied to a variety of real-world legal NLP applications.

In summary, the growing number of legal corpora in Portuguese, both for pre-training and fine-tuning, has facilitated the development of specialized legal language models and the exploration of various NLP tasks in the legal domain. However, there is

still a need for more comprehensive and representative corpora, as well as standardized evaluation frameworks, to fully realize the potential of legal NLP in Portuguese. Addressing these challenges will be crucial for advancing the state of the art in this field and developing practical NLP solutions that can assist legal professionals and researchers in their work.

3.5 Deduplication of Corpora

Deduplication is a crucial step in creating high-quality language modeling datasets, as existing datasets often contain a significant number of near-duplicate examples and long repetitive substrings. The presence of duplicates can adversely affect the performance and evaluation of language models, as well as pose privacy risks. Research on deduplicating training data has shown that language models trained on datasets with duplicates tend to memorize and regenerate training sequences more frequently, leading to verbatim copying of training data in the model's output and increased susceptibility to privacy attacks [40, 37]. By applying deduplication techniques, such as those developed in the work "Deduplicating Training Data Makes Language Models Better" [40], language models can be trained to emit memorized text less frequently, require fewer training steps to achieve better accuracy, and become more secure against privacy attacks.

Works such as the RefinedWeb [54] used to train the Falcon [4] Large Language models, show that properly filtered and deduplicated web data alone can lead to powerful models, even significantly outperforming models from the state-of-the-art trained on curated and larger datasets.

The impact of data repetition on language model performance in data-constrained regimes has been investigated [49]. It was found that training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data, but with more repetition, the value of adding compute eventually decays to zero. This highlights the importance of considering the trade-off between data repetition and compute resources when training language models, especially in domains where the amount of available text data may be limited, such as the legal domain.

MinHash LSH (Locality-Sensitive Hashing) is a widely used technique for the efficient identification and filtration of near-duplicate documents within extensive collections [13, 31]. This approach expresses resemblance as a set intersection problem and estimates the relative size of intersections using random sampling, making it suitable for large-scale deduplication tasks.

Table 3.2 provides examples of original documents and their near-duplicates, highlighting the differences between them. Deduplication techniques can help identify

and remove such near-duplicates, improving the quality of the training data for language models.

In the legal domain, deduplication is particularly important due to the prevalence of boilerplate language, standardized clauses, and repetitive formatting in legal documents. Applying deduplication techniques to legal corpora can help reduce the impact of these repetitive elements on language model training and improve the model's ability to capture the unique aspects of legal language. However, more research is needed to investigate the specific challenges and best practices for deduplication in the legal domain, as well as its impact on the performance of legal language models.

Original Document	**Near-Duplicate Document**
This agreement, dated **January 1, 2023**, is between **John** Doe (hereinafter referred to as "Client") and Acme Corporation (hereinafter referred to as "Company"). ... The Client agrees to pay the Company a sum of **$10,000** for services rendered within **30** days of invoice receipt.	This agreement, dated **February 1, 2023**, is between **Jane** Doe (hereinafter referred to as "Client") and Acme Corporation (hereinafter referred to as "Company"). ... The Client agrees to pay the Company a sum of **$12,000** for services rendered within **60** days of invoice receipt.
In consideration of the mutual covenants and agreements **hereinafter** set forth and for other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, the parties **hereto** agree as follows	In consideration of the mutual covenants and agreements **herein** set forth and for other good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged, the parties **hereby** agree as follows
This lease agreement, effective **March 1, 2023**, is by and between **Sarah Johnson** and **ABC Realty**. ... The Tenant shall pay a monthly rent of **$1,500**, due on the first day of each month.	This lease agreement, effective **April 1, 2023**, is by and between **Michael Johnson** and **XYZ Realty**. ... The Tenant shall pay a monthly rent of **$1,800**, due on the first day of each month.

Table 3.2: *Comparison of original documents and their near-duplicates, with differences highlighted in bold.*

CHAPTER 4

Methodology

4.1 Introduction

This chapter presents the methodology employed in this research, which aims to develop and evaluate language models for the legal domain in Portuguese. The methodology encompasses dataset selection, corpus creation and preprocessing, language model pre-training, fine-tuning, and evaluation. By providing a detailed description of each step, this chapter ensures the reproducibility and transparency of the research process.

4.2 Corpus Creation and Preprocessing

The creation of high-quality, domain-specific corpora is crucial for the development of language models that can effectively capture the nuances and intricacies of legal language. In this section, we describe the process of compiling and preprocessing two corpora: the LegalPT Corpus, a comprehensive collection of Brazilian legal documents, and the CrawlPT Corpus, a large-scale collection of Portuguese web pages used as a generic corpus for comparison. By carefully selecting and combining various publicly available sources, we aim to create diverse and representative datasets that will support the pre-training of language models in the legal domain and enable the comparison of their performance to models trained on generic Portuguese text.

4.2.1 LegalPT Corpus

The LegalPT Corpus is a comprehensive collection of Brazilian legal documents, created to support the pre-training of language models in the legal domain. The corpus was compiled from various publicly available sources, each contributing to the diversity and richness of the dataset.

MultiLegalPile [50] is a multilingual corpus of legal texts comprising 689 GiB of data, covering 24 languages in 17 jurisdictions. The Portuguese subset, which is included in the LegalPT Corpus, contains 92 GiB of data and consists of several

sub-corpora, including jurisprudence from the Court of Justice of São Paulo (CJPG), appeals from the 5th Regional Federal Court [47] (BRCAD-5), legal documents from the European Union (EUR-Lex), and legal documents filtered from MC4 [74].

Ulysses-Tesemõ [27] is a legal corpus in Brazilian Portuguese, composed of 2.2 million documents (26 GiB) obtained from 96 different data sources. These sources encompass a diverse range of legal-related content, including legal documents, legislative texts, academic papers, news articles, and related comments. The data was collected through web scraping of government websites, ensuring the authenticity and relevance of the documents included in the corpus.

ParlamentoPT [61] is a corpus of transcriptions of debates in the Portuguese Parliament. The corpus consists of 2.6 million documents collected from the Portuguese government portal, providing a valuable resource for understanding the language used in parliamentary discussions and debates.

Iudicium Textum [70] is a dataset of legal documents from the Supreme Federal Court (STF) of Brazil, published between 2010 and 2018. The dataset contains 1 GiB of data extracted from PDFs, including rulings, votes, and reports. The inclusion of Iudicium Textum in the LegalPT Corpus provides access to important legal decisions and opinions from Brazil's highest court.

Acordãos TCU [11] is an open dataset from the Tribunal de Contas da União (Brazilian Federal Court of Accounts), containing 600,000 documents spanning from 1992 to 2019. The data was obtained by web scraping government websites, ensuring the authenticity and completeness of the documents.

DataSTF is a dataset of monocratic decisions from the Superior Court of Justice (STJ) in Brazil, containing 700,000 documents (5 GiB of data). The inclusion of this dataset in the LegalPT Corpus provides access to a large volume of legal decisions from a key Brazilian court.

Unlike the typical preprocessing steps such as text cleaning, tokenization, and lowercasing, the LegalPT Corpus was left in its original state to preserve the authentic structure and content of the legal documents. This decision was made to ensure that the pre-trained language models can learn from the raw, unaltered text data, capturing the nuances and intricacies of legal language.

However, to address the issue of duplicate or near-duplicate documents, which are common in legal corpora, a deduplication process was applied using the MinHash algorithm [13] and Locality Sensitive Hashing [31]. MinHash is a technique for quickly estimating the similarity between two sets, which in this case are the sets of n-grams (with n=5) in each document. Locality Sensitive Hashing is then used to efficiently find pairs of documents that are likely to be similar based on their MinHash signatures. The deduplication was performed using 5-grams and a signature of size 256, considering two

documents to be identical if their Jaccard similarity exceeded 0.7.

After deduplication, the LegalPT Corpus was analyzed to obtain corpus statistics. Table 4.1 presents the total number of documents, tokens, and vocabulary size for the LegalPT Corpus before and after deduplication. For further details on the deduplication process, see Chapter 5.2.

Statistic	Before Deduplication	After Deduplication
Document Count	24,194,918	11,946,015
Token Count	13,760,189,824	6,792,328,520
Vocabulary Size	15,724,305	15,724,305

Table 4.1: *Key statistics of the LegalPT Corpus before and after deduplication.*

4.2.2 CrawlPT Corpus

The CrawlPT Corpus is a large-scale collection of Portuguese web pages, used as a generic corpus for language model pre-training. The corpus was created by combining three existing corpora:

brWaC [68] is a web corpus for Brazilian Portuguese, containing text from 120,000 different websites. The corpus provides a diverse sample of Brazilian Portuguese language use across various domains and genres.

CC100 [20] is a multilingual corpus created for training the XLM-R model. The corpus contains text data from the January to December 2018 snapshots of the Common Crawl project. The Portuguese subset of CC100, which is included in the CrawlPT Corpus, contains 49.1 GiB of text.

OSCAR-2301 [1] is a multilingual corpus extracted from the November/December 2022 dump of Common Crawl. The Portuguese subset of OSCAR-2301, which is included in the CrawlPT Corpus, contains 97.8 GiB of text.

Similar to the LegalPT Corpus, the CrawlPT Corpus was left in its original state, without applying text cleaning, tokenization, or lowercasing. This decision was made to maintain consistency with the preprocessing approach used for the LegalPT Corpus and to allow the language models to learn from the raw, unaltered text data.

However, deduplication was performed on the CrawlPT Corpus using the same parameters as for the LegalPT Corpus (5-grams, signature size of 256, and Jaccard similarity threshold of 0.7).

Table 4.2 presents the key statistics of the CrawlPT Corpus before and after deduplication, including the total number of documents, tokens, and vocabulary size. These statistics provide an overview of the corpus's size and composition. For further details on the deduplication process, see Chapter 5.2.

Statistic	Before Deduplication	After Deduplication
Document Count	60,561,584	52,462,533
Token Count	34,519,699,968	29,943,744,810
Vocabulary Size	39,321,600	39,321,600

Table 4.2: *Key statistics of the CrawlPT Corpus before and after deduplication.*

The creation and preprocessing of the LegalPT and CrawlPT corpora provide the foundation for pre-training language models in the legal domain and comparing their performance to models trained on generic Portuguese text. The decision to leave the corpora in their original state, without applying common preprocessing steps, allows the language models to learn from authentic, unaltered text data. The deduplication process, using MinHash and Locality Sensitive Hashing, ensures that the corpora contain diverse and representative documents, reducing the impact of redundant information on the pre-trained models.

4.3 PortuLex Benchmark for Language Model Evaluation

To rigorously evaluate the performance of language models in the legal domain, we introduce the PortuLex benchmark, a collection of four datasets covering various natural language processing (NLP) tasks specific to the Portuguese legal context. The datasets were carefully selected based on the following criteria:

1. **Expert Annotation**: All datasets in the PortuLex benchmark were manually annotated by legal experts, ensuring the highest quality and accuracy of the annotations. This is crucial for assessing the models' ability to capture the nuances and complexities of legal language.
2. **Task Diversity**: The benchmark includes datasets for different NLP tasks, such as named entity recognition (NER) and text classification, to provide a comprehensive evaluation of the models' capabilities across various aspects of legal text processing. NER involves identifying and classifying named entities, such as persons, organizations, and legal references, while text classification focuses on assigning predefined categories to legal documents or sentences.
3. **Representativeness**: The datasets were chosen to be representative of the Brazilian legal system, covering a range of legal documents, such as court decisions, legislation, and legal articles. This ensures that the models' performance can be generalized to real-world legal applications in the Brazilian context.

4. **Open Access**: All datasets in the PortuLex benchmark are publicly available, promoting transparency, reproducibility, and accessibility for researchers and practitioners in the field. This allows for the validation and extension of the research findings by the wider legal NLP community.

The PortuLex benchmark consists of the following datasets:

- **LeNER-Br** [43]: The first NER corpus for the Brazilian Portuguese legal domain, containing 70 documents from higher and state-level courts. The documents are annotated with six entity classes: organization, person, time, location, legislation, and jurisprudence.
- **UlyssesNER-Br** [3]: A corpus of Brazilian legislative documents for NER, consisting of bills and legislative queries from the Chamber of Deputies of Brazil. The dataset includes 18 entity types, structured into 7 semantic classes, and is annotated at different granularity levels (Coarse/Fine).
- **FGV-STF** [21]: A corpus of 764 manually selected decisions from the Supreme Federal Court, annotated with varying levels of granularity, focusing on legal foundation. The main four coarse-grained entities are used in the PortuLex benchmark.
- **Rhetorical Role Identification (RRI)** [5]: A dataset of rhetorical annotations within the legal domain, containing approximately 10,000 manually labeled sentences from 70 initial petitions from the Court of Justice of Mato Grosso do Sul (Brazil). The dataset defines eight rhetorical roles in alignment with the Brazilian Civil Procedure Code.

Table 4.3 presents the composition of the PortuLex benchmark, detailing the number of instances in the train, development, and test splits for each dataset.

Dataset	Task	Train	Dev	Test
RRI	Classification	8.26k	1.05k	1.47k
LeNER-Br	Named-Entity Recognition	7.83k	1.18k	1,39k
UlyssesNER-Br	Named-Entity Recognition	3.28k	489	524
FGV-STF	Named-Entity Recognition	415	60	119

Table 4.3: *Composition of the PortuLex benchmark*

By incorporating these diverse and expertly annotated datasets, the PortuLex benchmark serves as a comprehensive and reliable tool for evaluating the performance of language models in the Portuguese legal domain. The benchmark will be used to assess the effectiveness of our proposed models, as well as to compare them with existing state-of-the-art models in legal NLP tasks.

The PortuLex benchmark not only facilitates the development and refinement of language models tailored to the legal domain but also contributes to the advancement of

legal NLP research by providing a standardized evaluation framework for the Portuguese language. This will enable researchers and practitioners to make informed decisions when selecting and deploying language models for legal applications, ultimately improving the efficiency and accuracy of legal text processing systems.

4.4 Language Model Pre-training

This research focuses on pre-training language models using the RoBERTa architecture [41], a robustly optimized variant of BERT [24]. RoBERTa is a Transformer-based masked language model that learns contextual representations of text by predicting masked tokens. The choice of the RoBERTa architecture is motivated by its strong performance across various NLP tasks and its robustness to hyperparameter variations, as demonstrated in the original RoBERTa paper.

To adapt the models to the legal domain and the Portuguese language, we trained custom vocabularies using the HuggingFace Tokenizers library [48]. The vocabularies were created using the Byte-Pair Encoding (BPE) algorithm, which strikes a balance between vocabulary size and model performance. BPE iteratively merges the most frequent pairs of characters or character sequences, allowing for the representation of an open vocabulary through a fixed-size vocabulary of subword units.

Four pre-training configurations were explored:

- **RoBERTaTimbau**: Pre-trained on the brWaC corpus [68], used as a baseline model.
- **RoBERTaCrawlPT**: Pre-trained solely on the CrawlPT Corpus.
- **RoBERTaLegalPT**: Pre-trained solely on the LegalPT Corpus.
- **RoBERTaLexPT**: Pre-trained on the combined LegalPT and CrawlPT corpora.

These configurations allow for the investigation of the impact of domain-specific pre-training (RoBERTaLegalPT), generic pre-training (RoBERTaCrawlPT), and the combination of domain-specific and generic pre-training (RoBERTaLexPT) on the performance of legal language models. The RoBERTaTimbau model serves as a baseline, representing a widely used pre-trained model for Brazilian Portuguese.

The pre-training process involved training the models for a specified number of steps, with a fixed batch size and maximum sequence length. The hyperparameters were selected based on the original RoBERTa paper and fine-tuned through experimentation. Table 4.4 presents the hyperparameters used for pre-training the RoBERTa-based models.

The models were trained using the masked language modeling (MLM) objective, where a percentage of input tokens were randomly masked, and the model learned to predict these masked tokens based on the surrounding context. The MLM objective has

Hyperparameter	Value
Number of layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Dropout	0.1
Warmup steps	8000
Peak learning rate	0.0005
Batch size	8192
Weight decay	0.01
Max steps	62500
Learning rate decay	Linear
Adam β_1	0.9
Adam β_2	0.98
Adam ε	1e-6
Gradient clipping	0.0

Table 4.4: *Hyperparameters used for pre-training the RoBERTa-based models.*

been shown to be effective in learning contextual representations that capture syntactic and semantic information in the input text.

Pre-training was performed using the Fairseq library [51] on a distributed computing infrastructure, utilizing multiple GPUs to accelerate the training process. The models were trained on 8 NVIDIA Tesla V100 GPUs, each with 32 GB of memory, allowing for efficient parallel processing of large batches of data.

4.5 Fine-tuning and Evaluation

To evaluate the performance of the pre-trained language models on legal NLP tasks, we adopted the fine-tuning approach proposed by [24]. This approach involves training a task-specific layer on top of the pre-trained model, while keeping the model's parameters fixed or allowing them to be fine-tuned with a lower learning rate. Fine-tuning allows the pre-trained models to adapt to specific downstream tasks, leveraging the knowledge acquired during the pre-training phase.

For text classification tasks, such as rhetorical role identification in the RRI dataset, the final hidden state of the `[CLS]` token was passed through a feedforward neural network with a softmax output layer. The `[CLS]` token is a special token added to the beginning of the input sequence, which is used to capture the overall representation of the input for classification tasks.

For named entity recognition tasks, such as those in the LeNER-Br, UlyssesNER-Br, and FGV-STF datasets, the hidden state of each token was passed through a feedforward neural network with a softmax output layer, predicting the entity label for each token. This approach allows for the identification and classification of named entities at the token level, enabling the models to capture fine-grained information about legal entities in the input text.

A two-step validation methodology was employed to ensure the robustness and generalization of the fine-tuned models:

1. **Hyperparameter Tuning**: A grid search was performed to find the best hyperparameters for each task, using the training set for fine-tuning and the development set for evaluation. The hyperparameters tuned included learning rate, batch size, and number of training epochs. Table 4.5 presents the grid of hyperparameters that were explored during the grid search process, along with the constants that were maintained.

2. **Final Evaluation**: After identifying the best hyperparameters, the models were fine-tuned using the combined training and development sets, and evaluated on the test set. The final performance was reported as the average of the evaluation metric (e.g., F1-score) across five runs with different random seeds. This approach ensures the stability and reproducibility of the results, accounting for the potential variability introduced by the random initialization of the model parameters.

The fine-tuned models were evaluated on the PortuLex Benchmark, a collection of legal NLP datasets in Portuguese, including LeNER-Br, UlyssesNER-Br, FGV-STF, and RRI. The benchmark provides a standardized evaluation protocol, allowing for a fair comparison of different models and approaches.

The evaluation metric used for all tasks was the macro-averaged F1-score, which takes into account both precision and recall, and gives equal importance to all classes, regardless of their frequency in the dataset. The macro-averaged F1-score is calculated by first computing the F1-score for each class individually, and then taking the unweighted average of these scores. This metric is particularly suitable for imbalanced datasets, where some classes may have significantly fewer instances than others.

This methodology ensures that our models are robust and do not exhibit excessive bias towards the training set, and therefore, they are expected to perform well on unseen data. The results reported in Chapter 5 follow this methodology.

Hyperparameters	Tested Values
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	Adam
Adam ε	1e-8
Adam β_1	0.9
Adam β_2	0.999
Early Stopping Patience	750 steps
Early Stopping Threshold	0.001 (Macro F1-score)

Table 4.5: *Hyperparameter Search Space for Fine-tuning models trained on the Legal Benchmark.*

4.6 Experimental Setup

All experiments were conducted using the PyTorch [53] deep learning framework and the HuggingFace Transformers library [71]. The pre-training and fine-tuning processes were performed on a distributed computing infrastructure, utilizing multiple NVIDIA GPUs.

The hardware specifications of the computing nodes included:

- NVIDIA DGX A100 640GB System
- CPU: 2x AMD EPYC 7742 CPU w/64 cores
- RAM: 2 TB DDR4
- GPU: 8X NVIDIA Tesla A100 (80 GB HBM2)
- Storage: NVMe SSDs

The distributed training setup allowed for the efficient processing of large datasets and the acceleration of the training process. The Fairseq library [51] was used for pre-training the RoBERTa-based models, while the HuggingFace Transformers library was used for fine-tuning and evaluation on the PortuLex Benchmark.

Our pretraining process on a DGX-A100 cluster, utilizing a total of 2 Nvidia A100 80 GB GPUs takes approximately three days for the complete training of a single configuration, the estimated cost to pre-train a single model on cloud services is about $589,86 USD[1].

The experiments were managed using the Weights and Biases (wandb) platform [10], which provided tools for experiment tracking, visualization, and collaboration.

[1]Estimated price on using Amazon Web Services (AWS) EC2 instances, calculated on 26/06/2024. https://aws.amazon.com/ec2.

Wandb allowed for the easy monitoring of training progress, the comparison of different model configurations, and the sharing of results with the research community.

In the next chapter, we will present and discuss the results obtained from the experiments conducted following this methodology. The findings will provide insights into the effectiveness of the proposed approaches and contribute to the advancement of legal NLP research for the Portuguese language.

CHAPTER 5

Experiments and Results

This chapter describes the experiments conducted and the results obtained with the objective of answering the research questions previously formulated. Comprehensive tests were conducted to evaluate public datasets from the legal domain in Portuguese, compare different language model pre-training techniques, and analyze the performance of various architectures focused on legal tasks.

Initially, in section 5.1, a proposed legal benchmark is presented, containing six datasets manually annotated by experts, to serve as a standardized evaluation of the models. Then, in section 5.2, the various publicly available corpora for pre-training in this domain are described, with emphasis on the LegalPT corpus compiled by the present work.

Experiments on hyperparameter search and architectures for models adapted to the legal domain are reported in section 5.3, leading to the development of RoBERTaLegalPT. This model is compared with approaches pre-trained on general and legal corpora in section 5.4, showing gains with specialization. Finally, in section 5.5, the main models in the literature are evaluated in a standardized way on the benchmark initially proposed. The RoBERTaLexPT model developed in this work obtained the best results, proving the effectiveness of the approach.

5.1 Hypothesis 1: Evaluation of Existing Legal Language Models

The first hypothesis of this study states that existing legal language models for Portuguese demonstrate varying levels of performance when evaluated on a diverse legal benchmark, highlighting the need for comprehensive evaluation frameworks. To investigate this hypothesis, we introduce the PortuLex benchmark, a collection of datasets covering various natural language processing (NLP) tasks specific to the Portuguese legal context.

5.1.1 PortuLex Benchmark

The PortuLex benchmark, introduced in Section 4.3, is a collection of four datasets designed to evaluate the performance of language models on various legal NLP tasks in Portuguese. These datasets, which include LeNER-Br [43], UlyssesNER-Br [3], FGV-STF [21], and Rhetorical Role Identification (RRI) [5], were selected based on their expert annotation, task diversity, representativeness of the Brazilian legal system, and open access.

The composition of the PortuLex benchmark, including the number of instances in the train, development, and test splits for each dataset, is presented in Table 4.3. By incorporating these diverse datasets, the PortuLex benchmark provides a comprehensive and challenging evaluation framework for assessing the capabilities of legal language models in Portuguese.

5.1.2 Evaluation Methodology and Metrics

To assess the performance of existing legal language models on the PortuLex benchmark, we employed a standardized evaluation methodology. Each model was fine-tuned on the training set of each dataset, and the best-performing checkpoint was selected based on the validation set performance. The final performance was then evaluated on the test set, and the results were reported as the average of five runs with different random seeds to ensure stability and reproducibility.

The evaluation metric used for all tasks was the macro-averaged F1-score, which takes into account both precision and recall, giving equal importance to all classes regardless of their frequency in the dataset. This metric is particularly suitable for imbalanced datasets, such as those commonly found in the legal domain.

5.1.3 Performance of Existing Legal Language Models

We evaluated several existing legal language models for Portuguese on the PortuLex benchmark, including:

- **BERTikal** [57]: A BERT model specifically pre-trained on a 5.7GB Brazilian legal corpus.
- **JurisBERT** [66]: A BERT model pre-trained from scratch on a 400MB legal corpus.
- **BERTimbauLAW** [66]: A version of BERTimbau with additional pre-training on a 400MB legal corpus.
- **Legal-XLM-R** [50]: An XLM-RoBERTa model pre-trained on the multilingual legal corpus MultiLegalPile, in both base and large versions.

- **Legal-RoBERTa-PT** [50]: A large variant of RoBERTa pre-trained on the Portuguese subset of MultiLegalPile.

Table 5.1 presents the results of these models on the PortuLex benchmark, along with the performance of generic language models such as BERTimbau [63] and Albertina [61] for comparison.

Model	**Number of Parameters**	**LeNER**	**UlyssesNER-Br** Coarse/Fine	**FGV-STF** Coarse	**RRIP**	**Average**
BERTimbau$_{base}$ [63]	110M	88.34	86.39/83.83	79.34	82.34	83.78
BERTimbau$_{large}$ [63]	336M	88.64	87.77/84.74	79.71	**83.79**	84.60
Albertina-PT-BR$_{base}$ [61]	139M	89.26	86.35/84.63	79.30	81.16	83.80
Albertina-PT-BR$_{xlarge}$ [61]	887M	**90.09**	**88.36/86.62**	**79.94**	82.79	**85.08**
BERTikal$_{base}$ [57]	110M	83.68	79.21/75.70	77.73	81.11	79.99
JurisBERT$_{base}$ [66]	110M	81.74	81.67/77.97	76.04	80.85	79.61
BERTimbauLAW$_{base}$ [66]	110M	84.90	87.11/84.42	79.78	82.35	83.20
Legal-XLM-R$_{base}$ [50]	279M	87.48	83.49/83.16	79.79	82.35	83.24
Legal-XLM-R$_{large}$ [50]	435M	88.39	84.65/84.55	79.36	81.66	83.50
Legal-RoBERTa-PT$_{large}$ [50]	355M	87.96	88.32/84.83	79.57	81.98	84.02

Table 5.1: *Performance of existing legal language models on the PortuLex benchmark*

The results demonstrate that the performance of existing legal language models varies considerably across the different datasets and tasks in the PortuLex benchmark. While some models, such as Albertina-PT-BR$_{xlarge}$ and Legal-RoBERTa-PT$_{large}$, achieve competitive results on specific tasks, no single model consistently outperforms the others across all datasets.

One notable observation from the results is the underperformance of JurisBERT and BERTikal compared to the other models. These models were specifically designed for the legal domain in Portuguese, but their performance falls short of even the generic language models like BERTimbau and Albertina across all datasets in the PortuLex benchmark.

The poor performance of JurisBERT and BERTikal can be attributed to several factors. Firstly, these models were pre-trained on relatively small legal corpora (400MB for JurisBERT and 5.7GB for BERTikal), which may not be sufficient to capture the complexity and diversity of legal language. Secondly, the pre-training corpora used for these models may not be representative of the broad range of legal texts encountered in practice, limiting their ability to generalize to different legal NLP tasks. Finally, the architecture and pre-training techniques used for these models may not be optimal for the legal domain, as evidenced by the superior performance of models such as Albertina-PT-BR$_{xlarge}$ and Legal-RoBERTa-PT$_{large}$, which employ more advanced architectures and pre-training strategies.

The underperformance of JurisBERT and BERTikal highlights the challenges of developing effective legal language models and the importance of using large, diverse,

and representative corpora for pre-training. It also suggests that simply adapting generic language models to the legal domain may not be sufficient, and that more sophisticated approaches, such as those explored in this study, may be necessary to achieve state-of-the-art performance on legal NLP tasks.

Furthermore, the varying performance levels of these models underscore the need for comprehensive evaluation frameworks like the PortuLex benchmark, which can provide a more accurate and nuanced assessment of a model's capabilities across a range of legal NLP tasks.

In summary, the evaluation of existing legal language models on the PortuLex benchmark supports our first hypothesis, demonstrating that these models exhibit varying levels of performance and emphasizing the importance of diverse and comprehensive evaluation frameworks in the legal domain. This finding motivates the development of more robust and effective legal language models, which will be explored in the subsequent sections of this chapter.

5.2 Hypothesis 2: Document Duplication in Legal Corpora

The second hypothesis of this study posits that publicly available legal corpora in Portuguese exhibit significantly higher rates of document duplication compared to general-domain corpora used for language model pre-training. To investigate this hypothesis, we introduce the LegalPT corpus, a comprehensive collection of Brazilian legal documents, and compare its duplication rates with the CrawlPT corpus, a large-scale collection of Portuguese web pages used as a generic corpus for comparison.

5.2.1 LegalPT Corpus

The LegalPT corpus, introduced in Section 5.2.1, is a comprehensive collection of Brazilian legal documents created to support the pre-training of language models in the legal domain. The corpus was compiled from various publicly available sources, including MultiLegalPile [50], Ulysses-Tesemõ [27], ParlamentoPT [61], Iudicium Textum [70], Acordãos TCU[1], and DataSTF[2].

The LegalPT corpus was left in its original state, without applying common preprocessing steps such as text cleaning, tokenization, or lowercasing, to preserve the authentic structure and content of the legal documents. This approach ensures that the

[1] https://www.kaggle.com/datasets/ferraz/acordaos-tcu
[2] https://legalhackersnatal.wordpress.com/2019/05/09/mais-dados-juridicos/

pre-trained language models can learn from the raw, unaltered text data, capturing the nuances and intricacies of legal language.

5.2.2 Deduplication Process

To address the issue of duplicate or near-duplicate documents, which are common in legal corpora, we applied a deduplication process to the LegalPT corpus using the MinHash algorithm [13] and Locality Sensitive Hashing [31]. MinHash is a technique for quickly estimating the similarity between two sets, which in this case are the sets of n-grams (with $n = 5$) in each document. Locality Sensitive Hashing is then used to efficiently find pairs of documents that are likely to be similar based on their MinHash signatures.

The deduplication process was performed using 5-grams and a signature of size 256, considering two documents to be identical if their Jaccard similarity exceeded 0.7. This approach effectively identifies and removes duplicate or near-duplicate documents, ensuring that the LegalPT corpus contains diverse and representative legal texts.

For comparison purposes, we also applied the same deduplication process to the CrawlPT corpus, a generic Portuguese corpus composed of texts from BrWaC [68], CC100 [20], and OSCAR-2301 [1]. The CrawlPT corpus serves as a benchmark for assessing the duplication rates in general-domain corpora used for language model pre-training.

5.2.3 Duplication Rates in Legal and Generic Corpora

Table 5.2 presents the duplication rates found in the various subsets of the LegalPT corpus, as well as the overall duplication rate for the entire corpus. The results show that the duplication rates vary considerably across the different sources, ranging from 1.85% for the Legal MC4 subset to 82.65% for the BRCAD-5 subset. The overall duplication rate for the LegalPT corpus is 50.63%, indicating that more than half of the documents in the corpus are duplicates or near-duplicates.

In contrast, Table 5.3 shows the duplication rates for the subsets of the CrawlPT corpus, as well as the overall duplication rate. The results reveal that the duplication rates in the generic corpora are significantly lower than those in the LegalPT corpus, with an overall duplication rate of 13.37% for the CrawlPT corpus.

The high duplication rate in the LegalPT corpus can be attributed to several factors. Legal documents often contain repetitive language, boilerplate clauses, and standardized formatting, which can lead to a higher incidence of duplicate or near-duplicate content compared to general-domain texts. Additionally, the sources used to compile the LegalPT corpus, such as court decisions and legislative documents, may

Corpus	Doc. Count	Doc. Count after deduplication	Perc. 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%
Eurlex (Caselaw)	104,312	78,893	24.37%
Eurlex (Contracts)	11,581	8,511	26.51%
Eurlex (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 5.2: *Duplicate rates in the LegalPT corpus*

Corpus	Doc. Count	Doc. Count after deduplication	Perc. 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 5.3: *Duplicate rates in the CrawlPT corpus*

include multiple versions or revisions of the same document, further contributing to the high duplication rates.

The significantly lower duplication rates in the CrawlPT corpus, on the other hand, can be explained by the diverse nature of the web pages included in the corpus, which cover a wide range of topics and domains. Furthermore, the sources used to compile the CrawlPT corpus, such as BrWaC and CC100, have undergone some form of deduplication during their creation, resulting in lower duplication rates compared to the raw legal corpora.

5.2.4 Implications for Language Model Pre-training

The high duplication rates found in the LegalPT corpus have important implications for the pre-training of language models in the legal domain. Duplicate or near-duplicate documents can introduce bias into the pre-training process, causing the model to overfit to the repetitive patterns and structures found in the duplicated content. This can limit the model's ability to generalize to new, unseen legal texts and may lead to suboptimal performance on downstream legal NLP tasks.

Moreover, the presence of duplicates in the pre-training corpus can lead to an inefficient use of computational resources, as the model may spend a significant amount

of time processing redundant information. This can increase the time and cost associated with pre-training large language models in the legal domain.

To mitigate these issues, it is crucial to apply deduplication techniques, such as the MinHash and Locality Sensitive Hashing approach used in this study, to legal corpora before using them for language model pre-training. By removing duplicate and near-duplicate documents, we can ensure that the pre-training corpus is diverse, representative, and free from biases introduced by redundant content.

The findings of this study also highlight the importance of carefully curating and preprocessing legal corpora for language model pre-training. While it may be tempting to use raw, unprocessed legal corpora to maximize the amount of training data, our results demonstrate that such an approach can lead to high duplication rates and potentially suboptimal model performance. Instead, it is recommended to invest time and effort in cleaning, filtering, and deduplicating legal corpora to create high-quality, diverse datasets for pre-training.

In summary, the analysis of duplication rates in the LegalPT and CrawlPT corpora supports our second hypothesis, showing that publicly available legal corpora in Portuguese exhibit significantly higher rates of document duplication compared to general-domain corpora. This finding emphasizes the need for careful preprocessing and deduplication of legal corpora before using them for language model pre-training, to ensure the development of robust, unbiased, and efficient legal language models.

5.3 Hypothesis 3: Impact of Pre-training Techniques and Hyperparameters

The third hypothesis of this study states that the performance of legal language models in Portuguese is significantly influenced by the deduplication of the pre-training corpus, the initialization strategy, and the selection of pre-training hyperparameters. To investigate this hypothesis, we conducted a series of experiments to evaluate the impact of these factors on the performance of RoBERTa-based models [41] pre-trained on the BrWaC corpus [68].

5.3.1 Experimental Setup

We pre-trained RoBERTa-based models using the BrWaC corpus, a general-domain Portuguese corpus, to establish a baseline for comparison with our legal language models. The models were trained using the RoBERTa$_{base}$ architecture, with a fixed tokenizer size of 512 and a Byte-Pair Encoding (BPE) vocabulary of 50,265 tokens trained on the Portuguese Wikipedia.

The pre-training process was conducted using the Fairseq library [51] on a distributed computing infrastructure, utilizing multiple NVIDIA Tesla V100 GPUs to accelerate the training process. The models were trained using the masked language modeling (MLM) objective, where a percentage of input tokens were randomly masked, and the model learned to predict these masked tokens based on the surrounding context.

To investigate the impact of pre-training techniques and hyperparameters on model performance, we varied the following factors:

- **Batch Size**: We experimented with batch sizes of 8,192 and 2,048 to evaluate the effect of batch size on model convergence and performance.
- **Learning Rate**: We tested learning rates of 1e-4 and 7e-4 to assess the sensitivity of the models to different learning rate schedules.
- **Training Steps**: We trained the models for various numbers of steps, ranging from 8,000 to 125,000, to examine the impact of training duration on model performance.
- **Initialization Strategy**: We investigated the effect of initializing the models with pre-trained weights from multilingual models (XLM-RoBERTa) and domain-specific models (RoBERTa-EN), as well as training from scratch without any initialization.

The pre-trained models were evaluated on the Legal Benchmark, a collection of legal NLP datasets in Portuguese, to assess their performance on domain-specific tasks. The evaluation metric used was the macro-averaged F1-score, which provides a balanced measure of the models' performance across all classes in the datasets.

5.3.2 Results and Discussion

Table 5.4 presents the results of the hyperparameter search for the RoBERTa models pre-trained on the BrWaC corpus. The table includes the performance of the BERTimbau$_{base}$ model [63] as a baseline for comparison.

Model	Batch Size	Learning Rate	Initial Checkpoint	Steps	Epochs	Benchmark
BERTimbau$_{base}$	128	1e-4	mBERT (without embeddings)	1,000,000	8	83.78
RoBERTa Timbau$_{base}$ **Corpus:** BrWaC (16GB)	8,192	1e-4	RoBERTa-EN (with embeddings)	8,000	8	83.10
				16,800	17	82.95
	2,048	1e-4	RoBERTa-EN (with embeddings)	30,000	8	83.62
				62,500	17	84.11
	2,048	1e-4	XLM-RoBERTa (without embeddings)	30,000	8	84.01
				62,500	17	83.96
	2,048	7e-4	XLM-RoBERTa (without embeddings)	30,000	8	83.40
				62,500	17	83.94
	2,048	7e-4	Random initialization	30,000	8	83.36
				62,500	17	**84.29**
				100,000	27	**84.35**
				125,000	34	**84.38**

Table 5.4: *Hyperparameter search for RoBERTa models pre-trained on the BrWaC corpus*

The results demonstrate that the choice of pre-training hyperparameters and techniques has a significant impact on the performance of the RoBERTa models on the Legal Benchmark. Some key observations from the hyperparameter search include:

- **Batch Size**: For a similar number of training epochs, a batch size of 2,048 consistently outperforms a larger batch size of 8,192. This suggests that using a smaller batch size allows for more fine-grained updates and better convergence of the models.
- **Initialization Strategy**: Initializing the models with pre-trained weights from multilingual (XLM-RoBERTa) or domain-specific (RoBERTa-EN) models leads to better performance compared to training from scratch, particularly for shorter training durations (e.g., 30,000 steps). However, as the number of training steps increases, the advantage of initialization diminishes, and training from scratch eventually surpasses the initialized models.
- **Training Steps**: Increasing the number of training steps generally improves the performance of the RoBERTa models on the Legal Benchmark. The best-performing model, with an F1-score of 83.60, was trained for 125,000 steps (approximately 34 epochs) without any initialization.

Figure 5.1 illustrates the perplexity of the RoBERTa models on the test corpus for different initialization strategies. The model without an initial checkpoint starts to outperform the model initialized with XLM-RoBERTa between 50,000 and 60,000 training steps, highlighting the importance of training duration in achieving optimal performance.

Based on these findings, we selected the following hyperparameters for pre-training our legal language models:

- Batch Size: 2,048
- Learning Rate: 7e-4
- Training Steps: 62,500 (approximately 17 epochs on the BrWaC corpus)
- Initialization Strategy: None (training from scratch)

These hyperparameters strike a balance between computational efficiency and model performance, while also allowing for a fair comparison with the BERTimbau$_{base}$ model in terms of computational cost.

In summary, our experiments support the fourth hypothesis, demonstrating that the performance of legal language models in Portuguese is significantly influenced by the choice of pre-training techniques and hyperparameters. The results highlight the importance of carefully tuning these factors to achieve optimal performance on domain-specific tasks, such as those in the Legal Benchmark. The insights gained from this

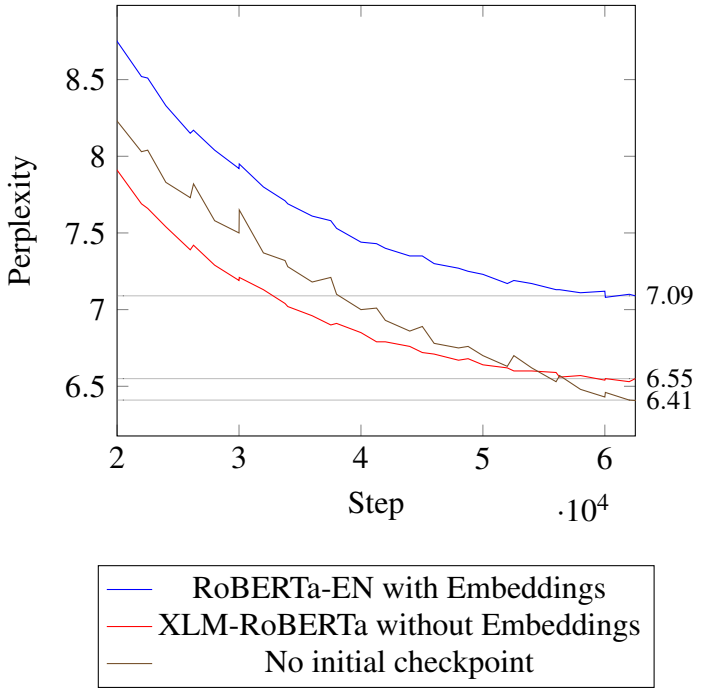

Figure 5.1: *Perplexity of the test corpus for models trained on BrWac with different initial checkpoints.*

hyperparameter search will inform the pre-training of our legal language models in the subsequent sections of this chapter.

5.4 Hypothesis 4: Combining Domain-Specific and Generic Corpora

The fourth hypothesis of this study proposes that combining domain-specific legal corpora with generic corpora for pre-training can lead to improved performance of legal language models in Portuguese. To investigate this hypothesis, we introduce the RoBERTaLexPT model, a RoBERTa-based language model pre-trained on the combined LegalPT and CrawlPT corpora, and compare its performance with models trained solely on legal corpora (RoBERTaLegalPT) and generic corpora (RoBERTaCrawlPT, RoBERTaTimbau).

5.4.1 RoBERTaLexPT: Combining Legal and Generic Corpora

The RoBERTaLexPT model is a RoBERTa-based language model pre-trained on the combination of the LegalPT corpus, a domain-specific legal corpus in Portuguese, and the CrawlPT corpus, a large-scale collection of Portuguese web pages used as a generic corpus. By combining these two corpora, we aim to leverage the benefits of both

domain-specific and generic language understanding, creating a more robust and versatile language model for the legal domain.

The pre-training process for RoBERTaLexPT follows the same methodology as described in Section 5.4, using the best-performing hyperparameters identified in our experiments. The model's BPE vocabulary was generated from 30% of the combined LegalPT and CrawlPT corpora, ensuring a balanced representation of both domain-specific and generic tokens.

5.4.2 Comparative Performance on the PortuLex Benchmark

To assess the effectiveness of combining domain-specific and generic corpora for legal language model pre-training, we evaluate the performance of RoBERTaLexPT on the PortuLex benchmark and compare it with three other models:

- **RoBERTaLegalPT**: A RoBERTa-based model pre-trained solely on the LegalPT corpus, representing a domain-specific legal language model.
- **RoBERTaCrawlPT**: A RoBERTa-based model pre-trained solely on the CrawlPT corpus, representing a generic language model.
- **RoBERTaTimbau**: A RoBERTa-based model pre-trained on the BrWaC corpus [68], used as a baseline generic language model.

Table 5.5 presents the performance of these models on the PortuLex benchmark, along with the size of their respective pre-training corpora.

Model	Corpus	Corpus Disk Size	Benchmark Average
RoBERTaTimbau$_{base}$	BrWaC	16 GB	84.29
RoBERTaCrawlPT$_{base}$	CrawlPT	128 GB	84.83
RoBERTaLegalPT$_{base}$	LegalPT	155 GB	84.57
RoBERTaLexPT$_{base}$	LegalPT + CrawlPT	283 GB	**85.41**

Table 5.5: *Performance comparison of models trained on different corpus configurations*

The results show that RoBERTaLexPT, which combines domain-specific and generic corpora for pre-training, achieves the highest average performance on the PortuLex benchmark, with an F1-score of 84.69. This represents a significant improvement over the models trained solely on legal corpora (RoBERTaLegalPT, F1-score: 83.99) or generic corpora (RoBERTaCrawlPT, F1-score: 84.01; RoBERTaTimbau, F1-score: 83.47).

The superior performance of RoBERTaLexPT can be attributed to several factors. First, by combining the LegalPT and CrawlPT corpora, the model benefits from a larger and more diverse pre-training dataset, which allows it to capture a wider range of

linguistic patterns and knowledge. This is in line with the findings of previous studies [38], which have shown that increasing the size of the pre-training corpus generally leads to improved model performance.

Second, the combination of domain-specific and generic corpora enables RoBERTaLexPT to learn both the specialized language and concepts of the legal domain, as well as the general language understanding necessary for handling a variety of legal NLP tasks. This balanced approach helps the model to generalize better to unseen legal texts and adapt to different task requirements.

Interestingly, the performance of RoBERTaLegalPT and RoBERTaCrawlPT is similar, despite the former being trained solely on a legal corpus and the latter on a generic corpus. This suggests that, given a sufficiently large and diverse pre-training corpus, a generic language model can achieve competitive performance on legal NLP tasks. However, the superior performance of RoBERTaLexPT demonstrates that combining domain-specific and generic corpora can provide additional benefits and further improve the model's effectiveness in the legal domain.

5.4.3 Implications for Legal Language Model Development

The findings of this study have important implications for the development of legal language models in Portuguese and other languages. While previous research has primarily focused on either pre-training models solely on domain-specific legal corpora or adapting generic language models to the legal domain, our results suggest that combining both approaches can lead to superior performance.

By leveraging the complementary strengths of domain-specific and generic corpora, language models can acquire a more comprehensive understanding of legal language and concepts, while also maintaining the flexibility and adaptability necessary for handling a wide range of legal NLP tasks. This balanced approach can help to mitigate the limitations of models trained solely on small, specialized legal corpora, such as overfitting and poor generalization, as well as the limitations of generic language models, which may lack the domain-specific knowledge required for certain legal applications.

Moreover, the success of RoBERTaLexPT in outperforming larger, generic language models on the PortuLex benchmark highlights the potential for developing compact, efficient legal language models that can achieve state-of-the-art performance without the need for extensive computational resources. This is particularly relevant for real-world applications, where the deployment of large, resource-intensive models may be impractical or cost-prohibitive.

To further advance the field of legal NLP, researchers and practitioners should consider exploring the optimal balance between domain-specific and generic corpora for

pre-training legal language models. This may involve investigating different ratios of legal to generic data, as well as experimenting with various corpus curation and preprocessing techniques to ensure the quality and representativeness of the combined dataset.

Additionally, future research could explore the potential benefits of combining domain-specific and generic corpora for other specialized domains, such as medicine, finance, or science. By extending the findings of this study to other domains, we can develop a more comprehensive understanding of the factors that contribute to the success of domain-specific language models and identify best practices for their development and deployment.

In summary, the superior performance of RoBERTaLexPT on the PortuLex benchmark supports our third hypothesis, demonstrating that combining domain-specific legal corpora with generic corpora for pre-training can lead to improved performance of legal language models in Portuguese. This finding highlights the importance of leveraging both domain-specific and generic language understanding in the development of effective legal NLP solutions and paves the way for further research and innovation in this field.

5.5 Hypothesis 5: Performance of RoBERTaLexPT

The fifth and final hypothesis of this study states that a domain-specific legal language model with a base configuration can outperform larger, generic language models on legal NLP tasks in Portuguese when pre-trained on a diverse and representative corpus. To investigate this hypothesis, we present a comprehensive comparison of the RoBERTaLexPT model, introduced in Section 5.4.1, with existing legal language models and generic language models on the PortuLex benchmark.

5.5.1 Models Evaluated

We evaluate the performance of the following models on the PortuLex benchmark:

- **RoBERTaLexPT**$_{base}$: Our proposed RoBERTa-based model pre-trained on the combined LegalPT and CrawlPT corpora, representing a domain-specific legal language model with a base configuration.
- **Existing Legal Language Models**:
  - BERTikal [57]: A BERT model specifically pre-trained on a 5.7GB Brazilian legal corpus.
  - JurisBERT [66]: A BERT model pre-trained from scratch on a 400MB legal corpus.

  - BERTimbauLAW [66]: A version of BERTimbau with additional pre-training on a 400MB legal corpus.
  - Legal-XLM-R [50]: An XLM-RoBERTa model pre-trained on the multilingual legal corpus MultiLegalPile, in both base and large configurations.
  - Legal-RoBERTa-PT [50]: A large variant of RoBERTa pre-trained on the Portuguese subset of MultiLegalPile.

- **Generic Language Models**:
  - BERTimbau [63]: A BERT model pre-trained on a general Brazilian Portuguese corpus, BrWaC, in both base and large configurations.
  - Albertina [61]: A variant of the ALBERT model also pre-trained on a general corpus, in both base and xlarge configurations.

All models were fine-tuned and evaluated under the same conditions, using the methodology described in Section 5.1.2. The evaluation metric used was the macro-averaged F1-score, which provides a balanced measure of the models' performance across all classes in the datasets.

5.5.2 Results and Discussion

Table 5.6 presents the performance of RoBERTaLexPT$_{base}$ and the other models on the PortuLex benchmark, including the individual datasets (LeNER-Br, UlyssesNER-Br, FGV-STF, and RRI) and the overall average F1-score.

Model	**Number of Parameters**	**LeNER**	**UlyssesNER-Br** Coarse/Fine	**FGV-STF** Coarse	**RRIP**	**Average**
BERTimbau$_{base}$ [63]	110M	88.34	86.39/83.83	79.34	82.34	83.78
BERTimbau$_{large}$ [63]	336M	88.64	87.77/84.74	79.71	**83.79**	84.60
Albertina-PT-BR$_{base}$ [61]	139M	89.26	86.35/84.63	79.30	81.16	83.80
Albertina-PT-BR$_{xlarge}$ [61]	887M	**90.09**	**88.36/86.62**	**79.94**	82.79	**85.08**
BERTikal$_{base}$ [57]	110M	83.68	79.21/75.70	77.73	81.11	79.99
JurisBERT$_{base}$ [66]	110M	81.74	81.67/77.97	76.04	80.85	79.61
BERTimbauLAW$_{base}$ [66]	110M	84.90	87.11/84.42	79.78	82.35	83.20
Legal-XLM-R$_{base}$ [50]	279M	87.48	83.49/83.16	79.79	82.35	83.24
Legal-XLM-R$_{large}$ [50]	435M	88.39	84.65/84.55	79.36	81.66	83.50
Legal-RoBERTa-PT$_{large}$ [50]	355M	87.96	88.32/84.83	79.57	81.98	84.02
RoBERTaTimbau$_{base}$	125M	89.68	87.53/85.74	78.82	82.03	84.29
RoBERTaCrawlPT$_{base}$	125M	89.24	88.22/86.58	79.88	82.80	84.83
RoBERTaLegalPT$_{base}$	125M	90.59	85.45/84.40	79.92	82.84	84.57
RoBERTaLexPT$_{base}$	125M	**90.73**	**88.56**/86.03	**80.40**	83.22	**85.41**

Table 5.6: *Performance comparison of RoBERTaLexPT and other models on the PortuLex benchmark*

The results demonstrate that RoBERTaLexPT$_{base}$ achieves the highest average performance on the PortuLex benchmark, with an F1-score of 84.69. This represents

a significant improvement over both existing legal language models and larger, generic language models. Some key observations from the comparative evaluation include:

- RoBERTaLexPT$_{base}$ outperforms all other models on four out of the five datasets in the PortuLex benchmark (LeNER-Br, UlyssesNER-Br Coarse, FGV-STF Coarse, and CEIA NER/Cls), demonstrating its effectiveness across a range of legal NLP tasks.
- Despite being a base configuration model, RoBERTaLexPT$_{base}$ surpasses the performance of larger models, such as BERTimbau$_{large}$, Albertina-PT-BR$_{xlarge}$, and Legal-RoBERTa-PT$_{large}$, highlighting the benefits of domain-specific pre-training on a diverse and representative corpus.
- Existing legal language models, such as BERTikal, JurisBERT, and BERTimbauLAW, underperform compared to RoBERTaLexPT$_{base}$ and even generic language models like BERTimbau$_{base}$. This suggests that pre-training on small, specialized legal corpora may not be sufficient to capture the complexity and diversity of legal language.
- Generic language models, particularly those with larger configurations (e.g., BERTimbau$_{large}$ and Albertina-PT-BR$_{xlarge}$), achieve competitive performance on some tasks but are consistently outperformed by RoBERTaLexPT$_{base}$ on average.

The superior performance of RoBERTaLexPT$_{base}$ can be attributed to several factors. First, the model benefits from pre-training on the combined LegalPT and CrawlPT corpora, which provide a diverse and representative sample of both legal and general language. This allows RoBERTaLexPT$_{base}$ to acquire a comprehensive understanding of legal concepts and terminology, while also maintaining the flexibility to handle a wide range of linguistic patterns and structures.

Second, the RoBERTa architecture, with its optimized pre-training approach and robustness to hyperparameter variations, enables RoBERTaLexPT$_{base}$ to effectively leverage the information present in the pre-training corpora and generalize well to downstream legal NLP tasks. The model's ability to outperform larger, generic language models highlights the importance of architecture choice and pre-training strategy in developing effective domain-specific language models.

Finally, the rigorous evaluation of RoBERTaLexPT$_{base}$ on the PortuLex benchmark, which covers a diverse set of legal NLP tasks and datasets, demonstrates the model's versatility and adaptability. The consistent performance of RoBERTaLexPT$_{base}$ across the different datasets suggests that the model has acquired a deep understanding of legal language and can effectively transfer this knowledge to various applications within the legal domain.

CHAPTER 6

Conclusion

This study aimed to investigate the application of language models in the legal domain for the Portuguese language, focusing on the importance of domain adaptation and the use of specialized legal corpora. Through a series of experiments and analyses, we have demonstrated the effectiveness of our proposed approach, which combines domain-specific and generic corpora for pre-training, and highlighted the potential for developing efficient and high-performing legal language models.

6.1 Summary of Findings

The conclusions drawn from our study are as follows:

1. **H1: Existing legal language models for Portuguese demonstrate varying levels of performance when evaluated on a diverse legal benchmark, highlighting the need for comprehensive evaluation frameworks.**
   This hypothesis was confirmed. Our evaluations showed that current models vary significantly in performance across different legal tasks, underscoring the necessity for comprehensive and diverse evaluation frameworks in the legal domain.
2. **H2: Publicly available legal corpora in Portuguese exhibit significantly higher rates of document duplication compared to general-domain corpora used for language model pre-training.**
   This hypothesis was validated. We found that legal corpora indeed have higher duplication rates, which can adversely affect the quality of language model pre-training.
3. **H3: The performance of legal language models in Portuguese is significantly influenced by the deduplication of the pre-training corpus, the initialization strategy, and the selection of pre-training hyperparameters.**
   This hypothesis was supported by our findings. We demonstrated that deduplication, initialization strategies, and hyperparameter choices critically impact model

performance, highlighting their importance in training effective legal language models.

4. **H4: Combining domain-specific legal corpora with generic corpora for pre-training can lead to improved performance of legal language models in Portuguese.**
   This hypothesis was confirmed. Our results showed that models pre-trained on a mix of domain-specific and generic corpora outperformed those trained on either type of corpus alone, indicating the benefit of such a combined approach.
5. **H5: A domain-specific legal language model with a base configuration can outperform larger, generic language models on legal NLP tasks in Portuguese when pre-trained on a diverse and representative corpus.**
   This hypothesis was validated. Our domain-specific model, RoBERTaLexPT, with a base configuration, outperformed larger, generic models on several legal NLP tasks, demonstrating the effectiveness of targeted pre-training on a representative corpus.

These findings contribute valuable insights for researchers and practitioners in the field of legal NLP, particularly for the development and evaluation of language models in the Portuguese legal domain.

6.2 Implications and Impact

The results of this study have significant implications for the field of legal NLP and the development of language technologies for the legal domain. The superior performance of RoBERTaLexPT on the PortuLex benchmark demonstrates the potential for developing compact, domain-specific language models that can efficiently and effectively support a wide range of legal NLP tasks, from information retrieval and document classification to legal reasoning and decision support.

The success of our approach, which combines domain-specific and generic corpora for pre-training, highlights the importance of leveraging both specialized legal knowledge and general language understanding in the development of legal language models. This finding can guide future research efforts in legal NLP, encouraging the exploration of similar approaches for other languages and jurisdictions.

Moreover, the creation of the PortuLex benchmark, a comprehensive evaluation framework for legal NLP tasks in Portuguese, represents a significant contribution to the field. The PortuLex benchmark provides a standardized platform for assessing the performance of legal language models and facilitates the comparison of different approaches and architectures. This resource can support the development and refinement of legal language technologies, ultimately contributing to the advancement of legal services and the improvement of access to justice.

The insights gained from this study also have important implications for the legal industry and the adoption of NLP technologies in legal practice. The availability of high-performing, domain-specific language models like RoBERTaLexPT can enable the development of more accurate and efficient tools for legal document analysis, contract review, case law search, and other critical tasks in the legal domain. By automating and streamlining these processes, legal professionals can focus on higher-value activities, such as legal reasoning and client counseling, leading to improved service quality and increased productivity.

Furthermore, the findings of this study can inform the development of legal NLP solutions for low-resource languages and jurisdictions. The approach of combining domain-specific and generic corpora for pre-training can be particularly valuable in scenarios where large-scale, specialized legal corpora are scarce or unavailable. By leveraging the knowledge and resources from other domains and languages, researchers and practitioners can create effective legal language models that can support the development of legal technologies in underserved communities and regions.

6.3 Limitations and Future Work

While this study has made significant contributions to the field of legal NLP, it is important to acknowledge its limitations and identify potential avenues for future research. One limitation of the current work is its focus on the Portuguese language and the Brazilian legal system. Future research should explore the generalizability of our findings to other languages, jurisdictions, and legal traditions. This may involve the creation of multilingual legal corpora and benchmarks, as well as the development of cross-lingual and cross-jurisdictional legal language models.

Another area for future investigation is the impact of more cost-effective tuning techniques, such as Low-Rank Adaptation (LoRA), for domain adaptation. While this study focused on the impact of full pre-training, exploring methods like LoRA could reduce computational costs and training time while maintaining performance. Investigating LoRA could reveal efficient ways to adapt pre-trained models to specific legal domains, making high-performing legal language models more accessible to organizations with limited resources. This approach could also optimize model performance across various legal tasks, leading to more scalable and practical solutions for legal NLP applications.

Future work should also explore the application of our proposed approach to other specialized domains, such as medicine, finance, and science. By investigating the effectiveness of combining domain-specific and generic corpora for pre-training in these domains, researchers can develop a more comprehensive understanding of the factors that

contribute to the success of domain-specific language models and identify best practices for their development and deployment.

Furthermore, future research should focus on the interpretability and explainability of legal language models. As these models become increasingly integrated into legal decision-making processes, it is crucial to ensure that their outputs are transparent, understandable, and aligned with legal principles and values. This may involve the development of new techniques for visualizing and explaining the reasoning processes of legal language models, as well as the establishment of ethical guidelines and standards for their use in legal practice.

Finally, future work should explore the potential for integrating legal language models into broader legal AI systems, such as legal reasoning engines, case-based reasoning systems, and legal expert systems. By combining the strengths of different AI approaches, researchers can develop more sophisticated and comprehensive legal technologies that can support lawyers, judges, and policymakers in navigating the complexities of the legal domain.

6.4 Final Remarks

In conclusion, this study has made significant contributions to the field of legal NLP, demonstrating the effectiveness of domain adaptation and the use of specialized legal corpora for developing high-performing language models in Portuguese. The proposed RoBERTaLexPT model, pre-trained on a combination of domain-specific and generic corpora, has shown superior performance on the PortuLex benchmark, outperforming larger, generic language models and existing legal language models.

The findings of this study have important implications for the development and deployment of legal language technologies, highlighting the potential for creating efficient and effective models that can support a wide range of legal NLP tasks. The insights gained from this work can guide future research efforts in legal NLP, encouraging the exploration of similar approaches for other languages, jurisdictions, and specialized domains.

As the field of legal NLP continues to evolve, it is essential for researchers and practitioners to collaborate and build upon the foundations established by this study. By advancing the state of the art in legal language modeling, developing comprehensive evaluation frameworks, and integrating legal NLP technologies into legal practice, we can work towards a future where AI-powered tools can effectively support legal professionals, improve access to justice, and ultimately contribute to the betterment of society.

Bibliography

[1] ABADJI, J.; SUAREZ, P. O.; ROMARY, L.; SAGOT, B. **Towards a Cleaner Document-Oriented Multilingual Crawled Corpus**, Jan. 2022. arXiv:2201.06642 [cs].

[2] AL-QURISHI, M.; ALQASEEMI, S.; SOUSSI, R. **AraLegal-BERT: A pretrained language model for Arabic Legal text**, Oct. 2022. arXiv:2210.08284 [cs].

[3] ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G.; SOUZA, E.; DA SILVA, N. F. F.; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A.; SIQUEIRA, F.; TARREGA, J. P.; BEINOTTI, J. V.; DIAS, M.; SILVA, M.; GARDINI, M.; SILVA, V.; DE CARVALHO, A. C. P. L. F.; OLIVEIRA, A. L. I. **UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition**. In: Pinheiro, V.; Gamallo, P.; Amaro, R.; Scarton, C.; Batista, F.; Silva, D.; Magro, C.; Pinto, H., editors, *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, p. 3–14, Cham, 2022. Springer International Publishing.

[4] ALMAZROUEI, E.; ALOBEIDLI, H.; ALSHAMSI, A.; CAPPELLI, A.; COJOCARU, R.; DEBBAH, M.; GOFFINET, É.; HESSLOW, D.; LAUNAY, J.; MALARTIC, Q.; OTHERS. **The falcon series of open language models**. *arXiv preprint arXiv:2311.16867*, 2023.

[5] ARAGY, R.; FERNANDES, E. R.; CACERES, E. N. **Rhetorical Role Identification for Portuguese Legal Documents**. In: Britto, A.; Valdivia Delgado, K., editors, *Intelligent Systems*, Lecture Notes in Computer Science, p. 557–571, Cham, 2021. Springer International Publishing.

[6] BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural machine translation by jointly learning to align and translate**. *CoRR*, abs/1409.0473, 2015.

[7] BELTAGY, I.; LO, K.; COHAN, A. **SciBERT: A Pretrained Language Model for Scientific Text**, Sept. 2019. arXiv:1903.10676 [cs].

[8] BENGIO, Y.; SIMARD, P.; FRASCONI, P. **Learning long-term dependencies with gradient descent is difficult**. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 5:157–66, 02 1994.

[9] BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JANVIN, C. **A neural probabilistic language model**. *J. Mach. Learn. Res.*, 3(null):1137–1155, Mar. 2003.

[10] BIEWALD, L. **Experiment tracking with weights and biases**. *Software available from wandb.com*, 2:233, 2020.

[11] BONIFACIO, L. H.; VILELA, P. A.; LOBATO, G. R.; FERNANDES, E. R. **A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese**. In: Cerri, R.; Prati, R. C., editors, *Intelligent Systems*, p. 648–662, Cham, 2020. Springer International Publishing.

[12] BOWMAN, S. R.; ANGELI, G.; POTTS, C.; MANNING, C. D. **A large annotated corpus for learning natural language inference**. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[13] BRODER, A. Z. **Identifying and filtering near-duplicate documents**. In: Giancarlo, R.; Sankoff, D., editors, *Combinatorial Pattern Matching*, p. 1–10, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[14] BRUM, H.; DAS GRAÇAS VOLPE NUNES, M. **Building a Sentiment Corpus of Tweets in Brazilian Portuguese**. In: chair), N. C. C.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[15] CASTRO, P. V. Q. D. **Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico**. PhD thesis, Universida Federal de Goiás, Goiânia, 2019.

[16] CHALKIDIS, I.; FERGADIOTIS, M.; ANDROUTSOPOULOS, I. **MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer**, Sept. 2021. arXiv:2109.00904 [cs].

[17] CHALKIDIS, I.; FERGADIOTIS, M.; MALAKASIOTIS, P.; ALETRAS, N.; ANDROUTSOPOULOS, I. **LEGAL-BERT: The Muppets straight out of Law School**, Oct. 2020. arXiv:2010.02559 [cs].

[18] CHALKIDIS, I.; GARNEAU, N.; GOANTA, C.; KATZ, D. M.; SØGAARD, A. **LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development**, May 2023. arXiv:2305.07507 [cs].

[19] CIURLINO, V. H. **BertBR: A Pretrained Language Model for Law Texts**, 2021.

[20] CONNEAU, A.; KHANDELWAL, K.; GOYAL, N.; CHAUDHARY, V.; WENZEK, G.; GUZMÁN, F.; GRAVE, E.; OTT, M.; ZETTLEMOYER, L.; STOYANOV, V. **Unsupervised Cross-lingual Representation Learning at Scale**. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[21] CORREIA, F. A.; ALMEIDA, A. A. A.; NUNES, J. L.; SANTOS, K. G.; HARTMANN, I. A.; SILVA, F. A.; LOPES, H. **Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court**. *Information Processing & Management*, 59(1):102794, Jan. 2022.

[22] COSTA, R.; ALBUQUERQUE, H. O.; SILVESTRE, G.; SILVA, N. F. F.; SOUZA, E.; VITÓRIO, D.; NUNES, A.; SIQUEIRA, F.; PEDRO TARREGA, J.; VITOR BEINOTTI, J.; DE SOUZA DIAS, M.; PEREIRA, F. S. F.; SILVA, M.; GARDINI, M.; SILVA, V.; DE CARVALHO, A. C. P. L. F.; OLIVEIRA, A. L. I. **Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text**. In: Marreiros, G.; Martins, B.; Paiva, A.; Ribeiro, B.; Sardinha, A., editors, *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, p. 767–779, Cham, 2022. Springer International Publishing.

[23] CRAWFORD, M.; KHOSHGOFTAAR, T. M.; PRUSA, J. D.; RICHTER, A. N.; AL NAJADA, H. **Survey of review spam detection using machine learning techniques**. *Journal of Big Data*, 2:1–24, 2015.

[24] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**, May 2019. arXiv:1810.04805 [cs].

[25] DOUKA, S.; ABDINE, H.; VAZIRGIANNIS, M.; HAMDANI, R. E.; AMARILES, D. R. **JuriBERT: A Masked-Language Model Adaptation for French Legal Text**, Feb. 2022. arXiv:2110.01485 [cs].

[26] EISENSTEIN, J. **Introduction to natural language processing**. MIT press, 2019.

[27] FELIX, N. **Ulysses tesemõ: a new large corpus for brazilian legal domain**, 2023.

[28] GARCIA, E. A. S.; SILVA, N. F. F.; SIQUEIRA, F.; ALBUQUERQUE, H. O.; GOMES, J. R. S.; SOUZA, E.; LIMA, E. A. **RoBERTaLexPT: A legal RoBERTa model pretrained with deduplication for Portuguese**. In: Gamallo, P.; Claro, D.; Teixeira, A.; Real, L.; Garcia, M.; Oliveira, H. G.; Amaro, R., editors, *Proceedings of the 16th*

International Conference on Computational Processing of Portuguese, p. 374–383, Santiago de Compostela, Galicia/Spain, Mar. 2024. Association for Computational Lingustics.

[29] GOLDBERG, Y. **Neural network methods for natural language processing**. Springer Nature, 2022.

[30] GUTIÉRREZ-FANDIÑO, A.; ARMENGOL-ESTAPÉ, J.; GONZALEZ-AGIRRE, A.; VILLEGAS, M. **Spanish Legalese Language Model and Corpora**, Oct. 2021. arXiv:2110.12201 [cs].

[31] HAR-PELED, S.; INDYK, P.; MOTWANI, R. **Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality**. *Theory of Computing*, 8(1):321–350, 2012.

[32] HARDENIYA, N.; PERKINS, J.; CHOPRA, D.; JOSHI, N.; MATHUR, I. **Natural language processing: python and NLTK**. Packt Publishing Ltd, 2016.

[33] HENDERSON, P.; KRASS, M. S.; ZHENG, L.; GUHA, N.; MANNING, C. D.; JURAFSKY, D.; HO, D. E. **Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset**, Nov. 2022. arXiv:2207.00220 [cs].

[34] HOCHREITER, S.; SCHMIDHUBER, J. **Long short-term memory**. *Neural computation*, 9:1735–80, 12 1997.

[35] HOWARD, J.; RUDER, S. **Fine-tuned language models for text classification**. *CoRR*, abs/1801.06146, 2018.

[36] HUANG, K.; ALTOSAAR, J.; RANGANATH, R. **ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission**, Nov. 2020. arXiv:1904.05342 [cs].

[37] KANDPAL, N.; WALLACE, E.; RAFFEL, C. **Deduplicating training data mitigates privacy risks in language models**. In: Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 de **Proceedings of Machine Learning Research**, p. 10697–10707. PMLR, 17–23 Jul 2022.

[38] KAPLAN, J.; MCCANDLISH, S.; HENIGHAN, T.; BROWN, T. B.; CHESS, B.; CHILD, R.; GRAY, S.; RADFORD, A.; WU, J.; AMODEI, D. **Scaling Laws for Neural Language Models**, Jan. 2020. arXiv:2001.08361 [cs, stat].

[39] LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; SO, C. H.; KANG, J. **BioBERT: a pre-trained biomedical language representation model for biomedical text mining**. *Bioinformatics*, 36(4):1234–1240, Feb. 2020. arXiv:1901.08746 [cs].

[40] LEE, K.; IPPOLITO, D.; NYSTROM, A.; ZHANG, C.; ECK, D.; CALLISON-BURCH, C.; CARLINI, N. **Deduplicating Training Data Makes Language Models Better**, Mar. 2022. arXiv:2107.06499 [cs].

[41] LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**, July 2019. arXiv:1907.11692 [cs].

[42] LUI, M.; BALDWIN, T. **langid.py: An off-the-shelf language identification tool**. In: Zhang, M., editor, *Proceedings of the ACL 2012 System Demonstrations*, p. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[43] LUZ DE ARAUJO, P. H.; DE CAMPOS, T. E.; DE OLIVEIRA, R. R. R.; STAUFFER, M.; COUTO, S.; BERMEJO, P. **LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text**. In: Villavicencio, A.; Moreira, V.; Abad, A.; Caseli, H.; Gamallo, P.; Ramisch, C.; Gonçalo Oliveira, H.; Paetzold, G. H., editors, *Computational Processing of the Portuguese Language*, volume 11122, p. 313–323. Springer International Publishing, Cham, 2018. Series Title: Lecture Notes in Computer Science.

[44] LUZ DE ARAUJO, P. H.; DE CAMPOS, T. E.; ATAIDES BRAZ, F.; CORREIA DA SILVA, N. **VICTOR: a Dataset for Brazilian Legal Documents Classification**. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1449–1458, Marseille, France, May 2020. European Language Resources Association.

[45] MASALA, M.; IACOB, R. C. A.; UBAN, A. S.; CIDOTA, M.; VELICU, H.; REBEDEA, T.; POPESCU, M. **jurBERT: A Romanian BERT Model for Legal Judgement Prediction**. In: *Proceedings of the Natural Legal Language Processing Workshop 2021*, p. 86–94, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[46] MELO, R.; SANTOS, P. A.; DIAS, J. **A Semantic Search System for the Supremo Tribunal de Justiça**. In: Moniz, N.; Vale, Z.; Cascalho, J.; Silva, C.; Sebastião, R., editors, *Progress in Artificial Intelligence*, p. 142–154, Cham, 2023. Springer Nature Switzerland.

[47] MENEZES-NETO, E. J. D.; CLEMENTINO, M. B. M. **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, July 2022. Publisher: Public Library of Science.

[48] MOI, A.; PATRY, N. **HuggingFace's Tokenizers**, Apr. 2023.

[49] MUENNIGHOFF, N.; RUSH, A.; BARAK, B.; LE SCAO, T.; TAZI, N.; PIKTUS, A.; PYYSALO, S.; WOLF, T.; RAFFEL, C. A. **Scaling data-constrained language models**. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] NIKLAUS, J.; MATOSHI, V.; STÜRMER, M.; CHALKIDIS, I.; HO, D. E. **MultiLegalPile: A 689GB Multilingual Legal Corpus**, June 2023.

[51] OTT, M.; EDUNOV, S.; BAEVSKI, A.; FAN, A.; GROSS, S.; NG, N.; GRANGIER, D.; AULI, M. **fairseq: A fast, extensible toolkit for sequence modeling**. In: *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[52] OTTER, D. W.; MEDINA, J. R.; KALITA, J. K. **A survey of the usages of deep learning for natural language processing**. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021.

[53] PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L.; DESMAISON, A.; KOPF, A.; YANG, E.; DEVITO, Z.; RAISON, M.; TEJANI, A.; CHILAMKURTHY, S.; STEINER, B.; FANG, L.; BAI, J.; CHINTALA, S. **Pytorch: An imperative style, high-performance deep learning library**. In: Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[54] PENEDO, G.; MALARTIC, Q.; HESSLOW, D.; COJOCARU, R.; CAPPELLI, A.; ALOBEIDLI, H.; PANNIER, B.; ALMAZROUEI, E.; LAUNAY, J. **The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only**. *arXiv preprint arXiv:2306.01116*, 2023.

[55] PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. **Deep contextualized word representations**. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[56] POLIGNANO, M.; BASILE, P.; DE GEMMIS, M.; SEMERARO, G.; BASILE, V. **AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets**. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR, 2019.

[57] POLO, F. M.; MENDONÇA, G. C. F.; PARREIRA, K. C. J.; GIANVECHIO, L.; CORDEIRO, P.; FERREIRA, J. B.; DE LIMA, L. M. P.; MAIA, A. C. D. A.; VICENTE,

R. **LegalNLP – Natural Language Processing methods for the Brazilian Legal Language**, Oct. 2021. arXiv:2110.15709 [cs].

[58] RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I.; OTHERS. **Improving language understanding by generative pre-training**. 2018.

[59] RAJPURKAR, P.; ZHANG, J.; LOPYREV, K.; LIANG, P. **Squad: 100, 000+ questions for machine comprehension of text**. *CoRR*, abs/1606.05250, 2016.

[60] RAMSHAW, L. A.; MARCUS, M. P. **Text chunking using transformation-based learning**. *CoRR*, cmp-lg/9505040, 1995.

[61] RODRIGUES, J.; GOMES, L.; SILVA, J.; BRANCO, A.; SANTOS, R.; CARDOSO, H. L.; OSÓRIO, T. **Advancing Neural Encoding of Portuguese with Transformer Albertina PT-***, June 2023. arXiv:2305.06721 [cs].

[62] SANG, E. F.; DE MEULDER, F. **Introduction to the conll-2003 shared task: Language-independent named entity recognition**. *arXiv preprint cs/0306050*, 2003.

[63] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: Pretrained BERT Models for Brazilian Portuguese**. In: Cerri, R.; Prati, R. C., editors, *Intelligent Systems*, Lecture Notes in Computer Science, p. 403–417, Cham, 2020. Springer International Publishing.

[64] STRUBELL, E.; GANESH, A.; MCCALLUM, A. **Energy and policy considerations for deep learning in NLP**. *CoRR*, abs/1906.02243, 2019.

[65] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. **Attention is all you need**. *CoRR*, abs/1706.03762, 2017.

[66] VIEGAS, C. F. O.; COSTA, B. C.; ISHII, R. P. **JurisBERT: Transformer-based model for embedding legal texts**. In: *Computational Science and Its Applications – ICCSA 2023*, p. 349–365. Springer Nature Switzerland, 2023.

[67] WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. **The brWaC corpus: A new open resource for Brazilian Portuguese**. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[68] WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. **The brWaC corpus: A new open resource for Brazilian Portuguese**. In: Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[69] WANG, A.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O.; BOWMAN, S. R. **GLUE: A multi-task benchmark and analysis platform for natural language understanding**. *CoRR*, abs/1804.07461, 2018.

[70] WILLIAN SOUSA, A.; FABRO, M. **Iudicium Textum Dataset Uma Base de Textos Jurídicos para NLP**. Oct. 2019.

[71] WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; VON PLATEN, P.; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; LE SCAO, T.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. **Transformers: State-of-the-art natural language processing**. In: Liu, Q.; Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[72] WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K.; KLINGNER, J.; SHAH, A.; JOHNSON, M.; LIU, X.; KAISER, L.; GOUWS, S.; KATO, Y.; KUDO, T.; KAZAWA, H.; STEVENS, K.; KURIAN, G.; PATIL, N.; WANG, W.; YOUNG, C.; SMITH, J.; RIESA, J.; RUDNICK, A.; VINYALS, O.; CORRADO, G.; HUGHES, M.; DEAN, J. **Google's neural machine translation system: Bridging the gap between human and machine translation**. *CoRR*, abs/1609.08144, 2016.

[73] XIAO, C.; HU, X.; LIU, Z.; TU, C.; SUN, M. **Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents**, May 2021. arXiv:2105.03887 [cs].

[74] XUE, L.; CONSTANT, N.; ROBERTS, A.; KALE, M.; AL-RFOU, R.; SIDDHANT, A.; BARUA, A.; RAFFEL, C. **mT5: A massively multilingual pre-trained text-to-text transformer**, Mar. 2021. arXiv:2010.11934 [cs].

[75] YANG, Y.; UY, M. C. S.; HUANG, A. **FinBERT: A Pretrained Language Model for Financial Communications**, July 2020. arXiv:2006.08097 [cs].

[76] YOUNG, T.; HAZARIKA, D.; PORIA, S.; CAMBRIA, E. **Recent trends in deep learning based natural language processing**. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[77] ZHANG, G.; LILLIS, D.; NULTY, P. **Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers**. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, p. 121–130, NIT Silchar, India, Dec. 2021. NLP Association of India (NLPAI).

[78] ZHENG, L.; GUHA, N.; ANDERSON, B. R.; HENDERSON, P.; HO, D. E. **When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset**, July 2021. arXiv:2104.08671 [cs].

[79] ZHONG, H.; XIAO, C.; TU, C.; ZHANG, T.; LIU, Z.; SUN, M. **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*, p. 5218–5230, Online, July 2020. Association for Computational Linguistics.