Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Reconhecimento de Entidades Nomeadas em Documentos Legais

Fernando Hurias Lopes Neto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Luis Paulo Faina Garcia

Brasília
2026

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Reconhecimento de Entidades Nomeadas em Documentos Legais

Fernando Hurias Lopes Neto

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Luis Paulo Faina Garcia (Orientador)
PPGI/UnB

Prof.a Dr.a Nádia Felix Felipe da Silva
INF/UFG

Prof. Dr. Geraldo Pereira Rocha Filho
PPGI/UnB

Prof.a Dr.a Cláudia Nalon
Coordenadora do Programa de Pós-graduação em Informática

Brasília, 20 de Março de 2026

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Lopes Neto, Fernando Hurias
L846r Reconhecimento de Entidades Nomeadas em Documentos Legais / Fernando Hurias Lopes Neto; orientador Luis Paulo Faina Garcia. Brasília, 2026.
80 p.

Tese(Mestrado em Informática) Universidade de Brasília, 2026.

1. Reconhecimento de Entidades Jurídicas. 2. Abordagem Híbrida. 3. Domínio Jurídico Brasileiro. 4. Modelos de Linguagem. I. Faina Garcia, Luis Paulo, orient. II. Título.

Dedicatória

À minha avó/mãe, Antonia Moreira (in memoriam). Não há conquista minha que não carregue um pouco de você. Foi a sua fé simples e inabalável que me ensinou a continuar quando o caminho ficava difícil. Durante todo o mestrado, nos momentos de cansaço e de dúvida, era a sua voz que eu ouvia me dizendo para não desistir. Esta dissertação é minha, mas a força que a sustentou sempre foi sua. Com amor eterno.

Agradecimentos

O agradecimento mais importante que gostaria de fazer é ao meu orientador, Prof. Dr. Luis Paulo Faina Garcia, que me acompanhou com dedicação e paciência ao longo de toda essa jornada. Mais do que orientação técnica sobre Reconhecimento de Entidades Nomeadas em Documentos Legais, ele me ensinou a enxergar a pesquisa com rigor, responsabilidade e propósito, e isso faz toda a diferença.

Gostaria de agradecer também à Prof.ª Nádia Felix Felipe da Silva (INF/UFG) e ao Prof. Dr. Geraldo Pereira Rocha Filho (PPGI/UnB), membros da banca de qualificação e de defesa, que generosamente ofereceram sugestões e orientações fundamentais para o amadurecimento desta pesquisa. A colaboração de vocês foi essencial para que este trabalho chegasse aonde chegou.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro por meio da concessão de bolsa de estudos. Sem esse suporte, esta pesquisa simplesmente não teria sido possível.

À minha mãe, Maria do Socorro, e à minha tia Andreia, obrigado por nunca terem deixado de acreditar em mim, mesmo quando eu duvidei. À minha irmã Enmilly Kamylle e ao meu irmão Antonio Vinicius, pela cumplicidade de sempre. E a toda a minha família, que foi meu porto seguro durante esses anos: vocês são a razão de tudo.

Aos meus amigos e companheiros de mestrado, Lays Santana, Laylson Sampaio, João Tavares, Marcus Pontes, Jaciara Saraiva, Arthur Santos, Juliana Vale, Diogo Nava, Kailane, Mayara e Thiago que tornaram esses dois anos muito mais leves e significativos. Compartilhar essa jornada com vocês foi um privilégio. As conversas, os desabafos e as risadas nos corredores ficam.

A todos que, de alguma forma, fizeram parte desta caminhada, minha sincera gratidão.

Resumo

O sistema judiciário brasileiro é caracterizado por um extenso volume de documentos e linguagem altamente técnica, o que demanda métodos eficientes para automatizar a análise de textos jurídicos. O Reconhecimento de Entidades Nomeadas (NER), que identifica elementos como leis, tribunais e atores jurídicos, representa um desafio significativo nesse contexto. Embora o NER em português ainda seja pouco explorado, diversas abordagens, incluindo modelos estatísticos, redes neurais e *transformers*, têm sido investigadas. No entanto, esses estudos frequentemente carecem de comparações abrangentes entre diferentes modelos e conjuntos de dados. Este estudo aborda essa lacuna apresentando uma avaliação comparativa abrangente de modelos NER no domínio jurídico brasileiro, analisando os conjuntos de dados públicos disponíveis: LeNER-BR (acórdãos), CDJUR-BR (sentenças criminais) e UlyssesNER-BR (textos legislativos). Foram avaliados 23 modelos no total, considerando suas versões puras e variações combinadas com camadas BiLSTM, CRF, CNN e IDCNN. O conjunto abrange abordagens clássicas (CRF, BiLSTM), baseados em *transformers* (BERT, XLNet, ELECTRA) e híbridas. O estudo avaliou o desempenho usando métricas de *Precision*, *Recall* e *F1-Score*, combinadas com testes estatísticos de Friedman e Nemenyi para identificar diferenças significativas. Os resultados indicam que abordagens híbridas superam consistentemente os modelos isolados, com destaques específicos por conjunto: BERT-BiLSTM-CRF alcançou o *F1-Score* de 0,9440 no LeNER-BR, BERT-CRF atingiu 0,8020 no UlyssesNER-BR e BERT-BiLSTM obteve 0,8800 no CDJUR-BR. Testes estatísticos confirmaram diferenças significativas ($p < 0,05$) entre os modelos híbridos e suas contrapartes clássicas. A análise por entidade revelou que *transformers* puros se destacam em entidades semanticamente ricas, enquanto as abordagens híbridas são superiores para categorias com dependências sequenciais, demonstrando que a escolha da arquitetura deve considerar a natureza específica de cada corpus jurídico.

Palavras-chave: Reconhecimento de Entidades Jurídicas, Abordagem Híbrida, Domínio Jurídico Brasileiro, Modelos de Linguagem

Abstract

The Brazilian judicial system is characterized by a vast volume of documents and highly technical language, demanding efficient methods to automate the analysis of legal texts. Named Entity Recognition (NER), which identifies elements such as laws, courts, and legal actors, represents a significant challenge in this context. Although NER in Portuguese is still relatively unexplored, several approaches, including statistical models, neural networks, and transformers, have been investigated. However, these studies often lack comprehensive comparisons between different models and datasets. This study addresses this gap by presenting a comprehensive comparative evaluation of NER models in the Brazilian legal domain, analyzing the publicly available datasets: LeNER-BR (court decisions), CDJUR-BR (criminal sentences), and UlyssesNER-BR (legislative texts). A total of 23 models were evaluated, considering their pure versions and variations combined with BiLSTM, CRF, CNN, and IDCNN layers. The set encompasses classical (CRF, BiLSTM), transformer-based (BERT, XLNet, ELECTRA), and hybrid approaches. The study evaluated performance using Precision, Recall, and F1-Score metrics, combined with Friedman and Nemenyi statistical tests to identify significant differences. The results indicate that hybrid approaches consistently outperform stand-alone models, with specific highlights per set: BERT-BiLSTM-CRF achieved an F1-Score of 0.9440 on LeNER-BR, BERT-CRF reached 0.8020 on UlyssesNER-BR, and BERT-BiLSTM obtained 0.8800 on CDJUR-BR. Statistical tests confirmed significant differences ($p < 0.05$) between the hybrid models and their classical counterparts. Entity analysis revealed that pure transformers excel in semantically rich entities, while hybrid approaches are superior for categories with sequential dependencies, demonstrating that the choice of architecture should consider the specific nature of each legal corpus.

Keywords: Legal Entity Recognition, Hybrid Approach,Brazilian legal domain, Language Models

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ALBERT A LITE BERT.

BERT Bidirectional Encoder Representation from Transformers.

BERTimbau Pretrained BERT Models for Brazilian Portuguese.

BILOU Beginning, Inside, Last, Outside, Unit.

BiLSTM Bidirectional Long Short-Term Memory.

BIOES Beginning, Inside, Outside, End, Single.

CD Diferença Crítica.

CDJUR-BR Coleção Dourada do Judiciário Brasileiro.

CNN Convolutional Neural Network.

CoNLL-03 Conference on Computational Natural Language Learning.

CRF Conditional Random Fields.

ELMo Embeddings from Language Model.

EN Entidades Nomeadas.

FAIR Facebook AI Research.

HMM Hidden Markov Model.

IA Inteligência Artificial.

IDCNN Iterated Dilated Convolutional Neural Network.

IOB Inside-Outside-Beginning.

LeNER-BR A Dataset for Named Entity Recognition in Brazilian Legal Text.

LER Legal Entity Recognition.

LSTM Long Short-Term Memory.

MEMM Maximum Entropy Markov Models.

MLM Masked Language Model.

MPs Ministérios Públicos.

NER Named Entity Recognition.

NSP Next Sentence Prediction.

OOV out-of-vocabulary.

PLN Processamento Linguagem Natural.

POS Past-of-speech.

RNN Rede Neural Recorrentes.

RoBERTa Robustly optmized BERT approach.

STF Supremo Tribunal Federal.

STJ Superior Tribunal de Justiça.

TCU Tribunal de Contas da União.

TJCE Tribunal de Justiça do Ceará.

TJMG Tribunal de Justiça de Minas Gerais.

Capítulo 1

Introdução

O sistema de justiça do Brasil é complexo, composto por várias instâncias e tribunais, cuja função é assegurar a implementação da Constituição, das leis e dos direitos básicos dos cidadãos. O sistema é organizado de maneira hierarquizada, com cortes como o Supremo Tribunal Federal (STF)[1], o Superior Tribunal de Justiça (STJ)[2], os tribunais de Justiça estaduais e entidades especializadas como o Tribunal de Contas da União (TCU)[3] e os Ministérios Públicos (MPs)[4], que exercem funções vitais na supervisão e controle da gestão pública. A justiça no Brasil se baseia nos princípios da transparência, acessibilidade e segurança jurídica, visando garantir que todos os cidadãos possam ter acesso à justiça e compreender a operação das instituições [3].

A eficiência e a agilidade no manejo de informações jurídicas são indispensáveis diante do crescente volume de casos e da complexidade das questões legais, que demandam ferramentas capazes de aprimorar o acesso à justiça [4]. Nesse cenário, a aplicação de tecnologias baseadas em Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) destaca-se como uma solução promissora, especialmente na análise de grandes volumes de dados, como sentenças judiciais, pareceres e normativas [5].

Entre as diversas técnicas de PLN, o *Named Entity Recognition* (NER) destaca-se, no domínio jurídico, por sua capacidade de identificar e classificar automaticamente elementos-chave em documentos legais, como nomes de partes, artigos de lei, datas e precedentes jurídicos [6]. Ao automatizar essas tarefas, o NER não apenas aprimora a pesquisa e o monitoramento de questões legais, mas também promove um acesso mais eficiente a informações, beneficiando profissionais do Direito, como advogados, magistrados e promotores, bem como à sociedade em geral [6, 7].

[1] https://portal.stf.jus.br/
[2] https://www.stj.jus.br/sites/portalp/Inicio
[3] https://sites.tcu.gov.br/
[4] https://www.cnmp.mp.br/portal/

O uso de abordagens clássicas, como o *Conditional Random Fields* (CRF) [8], as Redes *Long Short-Term Memory* (LSTM)[9] e Bidirecional LSTM (BiLSTM) [10], para a tarefa NER, tem mostrado eficiência em gerenciar a complexidade e o contexto dos textos legais. Essas tecnologias possibilitam a identificação de entidades em textos extensos e complexos, como as sentenças emitidas pelo STF ou pelo STJ, de maneira precisa e contextualizada [11]. Ademais, modelos de linguagem baseados em *transformers*, como o *Bidirectional Encoder Representations from Transformers* (BERT), conseguem identificar sutilezas e conexões entre as entidades, aprimorando a compreensão de textos legais e a implementação da lei [6]. Mais recentemente, variantes e adaptações desses modelos, como o *BERTimbau* [12] e o *Legal-BERT* [13], além de arquiteturas como o *XLNet* [14] e *ELECTRA* [2], têm sido exploradas para o português jurídico, demonstrando avanços significativos na captura de especificidades linguísticas e contextuais. Paralelamente, abordagens baseadas em redes neurais convolucionais (CNN) e *Iterated Dilated Convolutional Neural Networks* (IDCNN) [1] têm mostrado eficácia na extração de características locais e globais em sequências textuais, oferecendo um equilíbrio entre custo computacional e capacidade de modelagem de contexto.

A introdução do NER no cenário jurídico do Brasil tem um efeito notável, principalmente em relação à transparência e rapidez na supervisão de decisões. Por exemplo, decisões do STF sobre questões constitucionais ou do STJ sobre assuntos infraconstitucionais podem ser examinadas de forma mais ágil, identificando os precedentes que orientarão julgamentos e decisões subsequentes. Esta automatizaçã simplifica a identificação de artigos legais relevantes e precedentes pertinentes, aprimorando a atividade de advogados, magistrados e até mesmo dos MPs [13, 15].

Através do uso de abordagens de NER, podemos obter automaticamente informações sobre casos, identificando, por exemplo, as ações de proteção tomadas e os precedentes pertinentes. Este tipo de abordagem simplifica o acesso à justiça para as vítimas e possibilita um monitoramento mais eficiente das políticas governamentais de proteção [4]. Assim, a implementação do NER no sistema judicial do Brasil não só acelera e simplifica o acesso à justiça, como também reforça a transparência e a efetividade das sentenças judiciais. O uso de modelos como BERT, LSTM e CRF não apenas aprimora a análise de grandes volumes de dados jurídicos, mas também auxilia na construção de um sistema judicial mais acessível, ágil e equitativo para todos os cidadãos [15, 7].

A aplicação e validação das técnicas de NER discutidas dependem fundamentalmente da utilização de bases de dados jurídicas adequadas, que sejam representativas da complexidade do judiciário brasileiro. Para este fim, o trabalho se apoiará em *dataset* consolidados no domínio jurídico. Entre eles, destacam-se o LeNER-BR, que oferece textos judiciais focados em decisões do STJ e STF; o UlyssesNER-Br, focado especificamente

em documentos legislativos; e o CDJUR-BR, que amplia o escopo com acórdãos de diferentes instâncias. A escolha desses recursos anotados é crucial, pois permite não apenas o treinamento supervisionado e a avaliação rigorosa de modelos como BERT e LSTM, mas também assegura que os resultados obtidos tenham relevância prática e validade para a análise do sistema de justiça, fortalecendo a busca por maior transparência e eficiência.

1.1 Justificativa

Apesar dos avanços observados em âmbitos internacionais (sobretudo na língua inglesa), a tarefa de NER em língua portuguesa ainda enfrenta desafios severos, conforme apontado por Silveira et al. (2023) [13]. Historicamente, os modelos aplicados ao português enfrentam dificuldades para atingir o estado da arte devido à escassez de dados anotados em volume suficiente e à carência de ferramentas altamente especializadas para o idioma [11]. Embora existam pesquisas relevantes voltadas ao cenário nacional [16, 17, 7, 15, 18, 19], uma parcela considerável restringe-se ao uso de abordagens isoladas ou tradicionais que demonstram limitações para capturar as ricas nuances do jargão jurídico brasileiro.

Além disso, o contexto jurídico brasileiro exige uma compreensão mais profunda das entidades e das relações descritas nos textos, o que aumenta a complexidade da tarefa. Embora existam pesquisas voltadas ao NER em português [16, 17, 7, 15, 18, 19], muitas delas ainda utilizam abordagens clássicas. Apesar de oferecerem resultados promissores em algumas situações, essas abordagens demonstram limitações ao capturar as nuances e especificidades do idioma.

Historicamente, a tarefa de NER tem sido amplamente explorada por meio de abordagens clássicas, como os modelos de CRF e Redes Neurais Recorrentes (RNN), incluindo variantes como LSTM e BiLSTM. Essas técnicas se destacam pela robustez e capacidade de capturar padrões sequenciais em diferentes domínios. Mais recentemente, os modelos baseados em *transformers*, como o BERT e suas variantes (e.g., *BERTimbau*, *Legal-BERT*, *XLNet*, *ELECTRA*), têm redefinido o estado da arte em NER. Esses modelos são notáveis por sua capacidade de capturar contextos complexos e relações linguísticas profundas, superando as limitações das abordagens clássicas. No contexto brasileiro, modelos ajustados para o português, como o *BERTimbau*, têm sido empregados com sucesso em tarefas de PLN [12]. Além dos *transformers*, arquiteturas baseadas em CNN e IDCNN têm se destacado pela eficiência computacional e pela capacidade de modelar dependências de longo alcance com camadas convolucionais dilatadas, sendo particularmente úteis em cenários com restrições de recursos [1].

No domínio jurídico em português, a literatura ainda apresenta lacunas importantes, sendo a principal delas a ausência de comparações sistemáticas entre abordagens clás-

sicas, abordagens estado da arte baseadas em *transformers*, arquiteturas convolucionais modernas (CNN, IDCNN) e abordagens híbridas que combinam essas famílias. Documentos legais brasileiros apresentam desafios específicos, como a terminologia altamente técnica e o grande volume de texto não estruturado, o que pode afetar a eficácia de diferentes técnicas de NER. Diante disso, este trabalho propõe uma avaliação sistemática de múltiplas configurações de modelos: (i) abordagens clássicas isoladas (CRF, BiLSTM, BiLSTM-CRF); (ii) abordagens baseadas em *transformers* isoladas (BERT, *BERTimbau*, *Legal-BERT*, *XLNet*, *ELECTRA*); (iii) abordagens baseadas em CNN e IDCNN; e (iv) abordagens híbridas que integram representações contextuais de *transformers* com decodificação estruturada por CRF ou camadas recorrentes/convolucionais (e.g., *BERT*-CRF, *BERT*-BiLSTM-CRF, *XLNet*-CRF, *ELECTRA*-IDCNN-CRF). A integração nessas arquiteturas híbridas busca explorar o melhor de múltiplos mundos: a capacidade dos *transformers* de captar contextos profundos e informações semânticas ricas, a eficiência na extração de características locais e a eficácia dos modelos estruturados em modelar dependências de saída e transições entre rótulos. Com isso, espera-se superar as limitações de cada abordagem de forma isolada e avançar na aplicação em português brasileiro.

A viabilidade de conduzir a avaliação sistemática que este trabalho propõe, entre abordagens clássicas, baseadas em *transformers*, convolucionais e híbridas, só se tornou possível graças ao surgimento de *dataset* jurídicos robustos para o português. A existência de bases como o LeNER-BR [7], o UlyssesNER-Br [16, 19] e o CDJUR-BR [15] viabiliza a análise comparativa de diferentes arquiteturas de modelos nos contextos judicial e legislativo, abrangendo suas nuances e estilos textuais específicos. Mais do que apenas fornecer dados, a variedade desses conjuntos, que englobam desde peças processuais variadas até acórdãos específicos do STF, cria o ambiente controlado necessário para uma análise comparativa rigorosa. Portanto, ao empregar esses recursos, este estudo consegue endereçar a lacuna central da literatura: a falta de uma análise comparativa aprofundada e multiarquitetural, fundamentada em um espectro representativo de documentos legais brasileiros, para determinar o real estado da arte da tarefa de NER.

1.2 Objetivos

Este trabalho tem como objetivo principal realizar uma avaliação sistemática e comparativa do desempenho de diferentes abordagens para a tarefa de NER em textos jurídicos em língua portuguesa, abrangendo abordagens clássicos, baseados em *transformers* e híbridas que combinam essas famílias. A análise busca identificar quais configurações são mais adequadas para lidar com as especificidades e desafios do domínio jurídico brasileiro, contribuindo para o avanço do estado da arte.

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realizar um estudo aprofundado das principais abordagens clássicas (CRF, BiLSTM, BiLSTM-CRF), baseadas em *transformers* (BERT, *BERTimbau*, *Legal-BERT*, *XL-Net*, *ELECTRA*) e convolucionais (CNN, IDCNN) aplicadas à tarefa de NER, identificando suas capacidades, limitações e adequação ao domínio jurídico em português.
- Propor e implementar abordagens híbridas que combinem representações contextuais de modelos *transformers* ou convolucionais com decodificadores estruturados (CRF) ou camadas recorrentes, tais como *BERT*-CRF, *BERT*-BiLSTM-CRF, *XL-Net*-CRF, *ELECTRA*-IDCNN e CRF.
- Avaliar e comparar, de forma sistemática e com rigor metodológico, o desempenho de todas as abordagens clássicas, *transformers* e híbridas, utilizando *dataset* de referência do domínio jurídico brasileiro LeNER-BR, UlyssesNER-Br e CDJUR-BR, empregando métricas consolidadas e análise estatística para verificar a significância das diferenças observadas.
- Analisar a influência das características específicas de cada *datasets* (tipo de documento, domínio, estilo de anotação) no desempenho dos diferentes modelos, fornecendo *insights* sobre a generalização e a robustez das abordagens investigadas.

1.3 Hipótese

A aplicação de abordagens híbridas, que combinam modelos baseados em *transformers* (como *BERT*, *XLNet* e *ELECTRA*) com decodificadores estruturados por CRF, permite alcançar melhores desempenhos na tarefa de NER em língua portuguesa no domínio jurídico em relação às abordagens clássicas isoladamente. Adicionalmente, espera-se que arquiteturas que incorporam camadas intermediárias (BiLSTM ou IDCNN) entre o encoder e o CRF apresentem ganhos adicionais em *dataset* com maior complexidade estrutural.

1.4 Estrutura do Trabalho

Os demais capítulos estão organizados da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica do trabalho, abordando os principais conceitos relacionados à tarefa de NER, e o Capítulo 3 apresenta os trabalhos relacionados; o Capítulo 4 trata da proposta de pesquisa, contendo a proposta da pesquisa, a metodologia e os procedimentos seguidos para obter os resultados; o Capítulo 5 apresenta os resultados obtidos a partir do processo de avaliação dos modelos; e, por fim, o Capítulo 6 apresenta as conclusões do trabalho, a contribuição e as limitações e os trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que sustentam a presente pesquisa, com o objetivo de contextualizar e aprofundar os principais conceitos e abordagens relacionados ao NER. Inicialmente, na Seção 2.1, são discutidas as definições centrais de Entidades Nomeadas (EN), contemplando sua evolução conceitual, os esquemas de anotação comumente utilizados, os desafios específicos impostos por domínios especializados, como o jurídico, e a importância da construção de corpora anotados de qualidade.

Em seguida, na Seção 2.2, são explorados os principais modelos aplicados ao NER, desde abordagens estatísticas clássicas até as mais recentes arquiteturas baseadas em *deep learning* e *transformers*. Ao reunir essas dimensões, conceitual, técnica e aplicada, o capítulo fornece o embasamento necessário para compreender os métodos adotados e os experimentos desenvolvidos ao longo deste trabalho.

2.1 Definição de Entidades Nomeadas

O NER estabelece-se como uma subárea crucial do PLN, dedicada à tarefa de identificar e classificar instâncias de entidades predefinidas em textos. Sua função é fundamental para a extração de informação, pois habilita a identificação automatizada de conceitos de relevância em grandes volumes de dados não estruturados [20, 21]. A conceituação seminal da área remonta a Grishman & Sundheim (1996) definiram o NER primordialmente como o processo de reconhecer nomes próprios, categorizados em tipos como “pessoas”, “locais” e “organizações” [22]. Foi nesse contexto que o NER se consolidou como uma ferramenta indispensável para a extração automática de informações de fontes textuais, como diários oficiais e páginas da web.

Do ponto de vista formal, a tarefa de NER pode ser descrita como a geração de uma sequência de tuplas do tipo $\langle I_i, I_f, t \rangle$ a partir de uma sentença de entrada $s = \{w_1, w_2, ..., w_N\}$. Nesta representação, I_i e I_f denotam os índices inicial e final dos *tokens*

que compõem a entidade, enquanto t indica seu tipo semântico [6]. Essa transformação de texto livre em dados estruturados é o que possibilita uma classificação precisa das entidades reconhecidas, permitindo uma interpretação semântica mais acurada por parte dos sistemas computacionais. A Figura 2.1 oferece uma ilustração prática desse processo, exemplificando como um modelo genérico de NER identifica três entidades distintas em uma sentença e as representa por meio de tuplas.

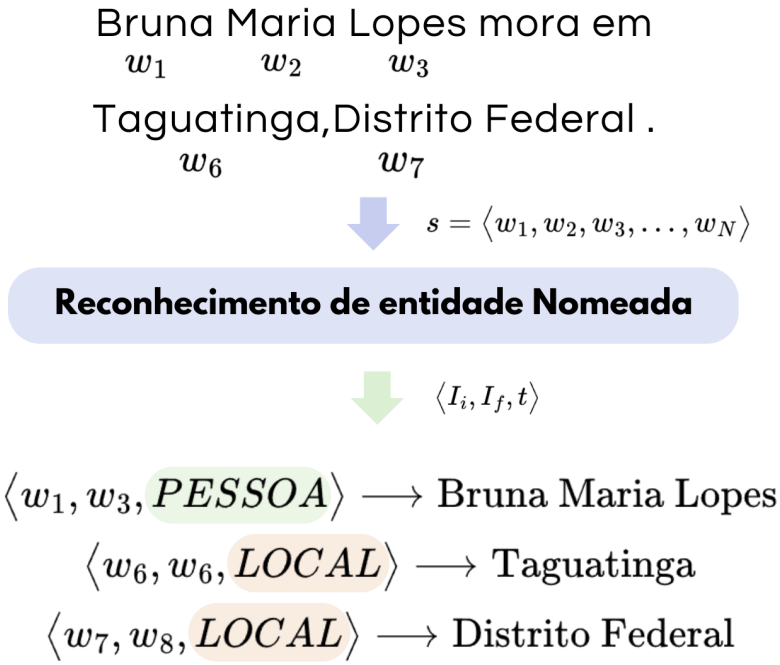

Figura 2.1: Exemplo de um modelo genérico de NER

A operacionalização eficaz do NER por modelos computacionais depende fundamentalmente do uso de esquemas de anotação padronizados. Tais esquemas impõem uma estrutura aos dados de treinamento, reformulando o problema de NER como uma tarefa de classificação de sequências, o que facilita a identificação precisa dos limites das entidades. O formato *Inside, Outside, Beginning* (IOB), proposto por Ramshaw & Marcus (1995), é um dos mais canônicos, utilizando a etiqueta "B" (*Beginning*) para marcar o início de uma entidade, "I" (*Inside*) para os *tokens* subsequentes da mesma entidade, e "O" (*Outside*) para palavras que não pertencem a nenhuma entidade [23]. Esse padrão cria uma representação sequencial que os modelos podem aprender e replicar. Existem variações importantes desse esquema, como IOB2, *Beginning, Inside, Outside, End, Single* (BIOES) e *Beginning, Inside, Last, Outside, Unit* (BILOU) [24]. A escolha entre eles não é trivial, pois cada variação oferece um nível distinto de granularidade na marcação, o que pode ser determinante para capturar as nuances das entidades. O esquema BIOES, por exemplo, aprimora o IOB ao introduzir uma etiqueta para entidades de um único *token* "S" (*Single*) e outra para o *token* final de uma entidade multi-palavra "E"

(*End*). Da mesma forma, o BILOU utiliza a etiqueta "L" (*Last*) para a última palavra de uma entidade e "U" (*Unit*) para entidades únicas. O impacto da escolha do esquema de anotação no desempenho do modelo é particularmente pronunciado em domínios especializados, como o jurídico. Nesses contextos, onde a terminologia é precisa e a estrutura das entidades é complexa, como nomes de leis, artigos ou partes processuais, um esquema de anotação mais granular pode capacitar o modelo a aprender padrões mais refinados e, consequentemente, alcançar maior acurácia.

Com a evolução da pesquisa, a definição original de NER expandiu-se para abranger um espectro mais amplo de categorias, incluindo datas, valores monetários e outras entidades numéricas, o que ampliou sua aplicabilidade para domínios como o financeiro e o jurídico [25]. Nadeau & Sekine (2007) argumentam que as EN não se restringem a nomes próprios, mas englobam quaisquer expressões que funcionem como designadores rígidos, incluindo as entidades temporais e numéricas que não estavam no escopo inicial do NER [26]. Embora essa ampliação conceitual tenha tornado o NER uma ferramenta mais poderosa e abrangente, ela introduziu, como contraponto, desafios significativos, notadamente a dificuldade em estabelecer uma taxonomia universalmente aceita. Conforme destaca Marrero *et al.* (2013), a própria definição de uma EN é dependente do domínio e dos objetivos da aplicação, tornando inviável a criação de uma categorização única e fixa [27].

Essa dependência contextual reflete-se diretamente na prática de desenvolver taxonomias específicas para cada domínio de aplicação. No setor biomédico, por exemplo, categorias como "doenças", "genes" e "medicamentos" são cruciais para a pesquisa e diagnóstico [28]. No meio jurídico, termos como "leis", "jurisprudência" e "partes processuais" assumem papel central para a análise de processos e documentos legais [7]. De forma análoga, no mercado financeiro, a identificação de "empresas", "produtos financeiros" e "valores monetários" é vital para a análise de risco e de mercado, enquanto na análise de inteligência, o reconhecimento de "eventos geopolíticos" e "organizações" permite o monitoramento de cenários complexos [26, 29]. Cada um desses cenários exige não apenas um conjunto de entidades customizado, mas também modelos capazes de compreender as sutilezas linguísticas daquele domínio. A busca por modelos com tal capacidade impulsionou diretamente a evolução das próprias arquiteturas computacionais, em uma trajetória que vai dos primeiros modelos estatísticos às complexas redes neurais profundas. A seção a seguir detalha esta progressão histórica e tecnológica.

2.2 Modelos para Reconhecimento de Entidades Nomeadas

Conforme introduzido, a história do NER é intrinsecamente ligada à evolução de seus modelos. Essa trajetória pode ser segmentada em distintos paradigmas tecnológicos, nos quais cada nova geração buscou superar as limitações das abordagens anteriores, em uma busca contínua por maior acurácia e capacidade de generalização [30].

Abordagens Clássicas

As abordagens iniciais para NER basearam-se em modelos estatísticos. Os esforços pioneiros se concentraram em modelos gerativos baseados em *Hidden Markov Models* (HMM), cuja aplicação à tarefa foi notabilizada por Freitag & McCallum (1999) [31]. Embora os HMM já fossem uma tecnologia estabelecida em PLN desde a década de 1970, foi nos anos 1990 que se consolidaram como ferramentas viáveis para NER [32]. No entanto, os HMM apresentavam limitações significativas, especialmente em sua dificuldade de modelar dependências de longo alcance e na incorporação de características textuais complexas e sobrepostas [33].

Para superar algumas dessas restrições, o *Maximum-Entropy Markov Model* (MEMM) foi introduzido por McCallum *et al.* (2000) como uma evolução dos HMM [34]. Apesar dos avanços, os modelos baseados em MEMM ainda enfrentavam o "problema do viés de rótulo" (*label bias problem*), que afetava seu desempenho ao favorecer estados com menos transições de saída. A solução para essa limitação veio com a introdução do CRF por Lafferty *et al.* (2001) [8]. O CRF é um método estatístico discriminativo que modela diretamente a probabilidade condicional de uma sequência de rótulos dada uma sequência de observações. Ao contrário dos MEMM, o CRF utiliza um modelo exponencial único para a probabilidade conjunta de toda a sequência de rótulos, resolvendo o problema do viés de rótulo e oferecendo maior flexibilidade para capturar dependências complexas entre os dados. Essa capacidade de incluir características arbitrárias e sobrepostas nas observações tornou o CRF especialmente valioso para o PLN, com estudos relatando resultados promissores em tarefas de NER [35, 16, 36].

Naquela época, modelos de NER de alto desempenho frequentemente utilizavam dicionários externos e características manualmente selecionadas, como demonstrado pelo modelo vencedor da conferência CoNLL-03 [37]. O CRF ainda é um dos pilares deste estudo devido à sua eficácia em domínios com dados limitados, como textos jurídicos, onde a estrutura sintática é crítica [16, 19, 38, 17]. Matematicamente, um CRF de cadeia linear modela a probabilidade condicional $p(\mathbf{y}|\mathbf{x})$ da sequência de rótulos $\mathbf{y} = (y_1, \ldots, y_n)$

dada a sequência de observações $\mathbf{x} = (x_1, \ldots, x_n)$. Essa probabilidade é expressa por uma função exponencial normalizada:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, \mathbf{x}, i)\right), \tag{2.1}$$

onde f_k são funções de características (que podem depender do rótulo atual, do anterior e de todo o contexto observacional), λ_k são os pesos aprendidos, e $Z(\mathbf{x})$ é a função de partição que garante a normalização. Essa formulação permite capturar dependências complexas, enquanto a normalização global resolve o viés de rótulo. Devido à convexidade da função de verossimilhança logarítmica, o treinamento dos parâmetros por métodos iterativos como L-BFGS garante a convergência para um ótimo global.

A era do aprendizado profundo marcou uma nova fronteira para o NER. Collobert et al. (2011) introduziram uma abordagem inovadora ao utilizarem Redes Neurais Convolucionais (CNN) sobre *word embeddings* para diversas tarefas de PLN. Nesta arquitetura, a sentença é tratada como uma matriz de entrada onde cada linha representa o vetor de uma palavra (embedding), permitindo que filtros convolucionais capturem a semântica local e relações de vizinhança entre os termos [39]. Diferente das abordagens baseadas em janelas simples, a CNN processa a sentença através de uma camada de convolução que utiliza filtros deslizantes para extrair características hierárquicas. Como ilustrado na Figura 2.2, o processo inicia-se com uma matriz $S \in \mathbb{R}^{n \times d}$, onde n é o número de palavras e d a dimensão do vetor. Filtros de tamanho $h \times d$ percorrem a matriz verticalmente, onde h representa o tamanho da janela de contexto (n-grama) [40].

A operação fundamental da camada convolucional consiste em aplicar um filtro w a uma janela de palavras para produzir uma nova característica. Matematicamente, uma característica c_i é gerada a partir de uma janela de palavras $x_{i:i+h-1}$ pela operação $c_i = f(w \cdot x_{i:i+h-1} + b)$, onde $b \in \mathbb{R}$ é um termo de viés e f é uma função de ativação não linear, como a unidade linear retificada (ReLU) ou a tangente hiperbólica (tanh). Este cálculo permite que a rede identifique padrões específicos, como prefixos, sufixos ou sequências de letras maiúsculas que indicam uma entidade nomeada.

Para lidar com o comprimento variável das sentenças e garantir que as características mais importantes sejam preservadas, utiliza-se a camada de *Max-over-time Pooling*. Esta camada extrai o valor máximo de cada mapa de características, definido pela fórmula $\hat{c} = \max\{c_1, c_2, \ldots, c_{n-h+1}\}$. Ao selecionar apenas o sinal mais forte de cada filtro, a rede torna-se invariante à posição da palavra na frase, facilitando a identificação de uma entidade independentemente de onde ela apareça. Por fim, esse vetor resultante é processado por camadas densas para a classificação final, embora a arquitetura ainda apresente limitações na captura de dependências de longuíssimo prazo devido à janela fixa do filtro.

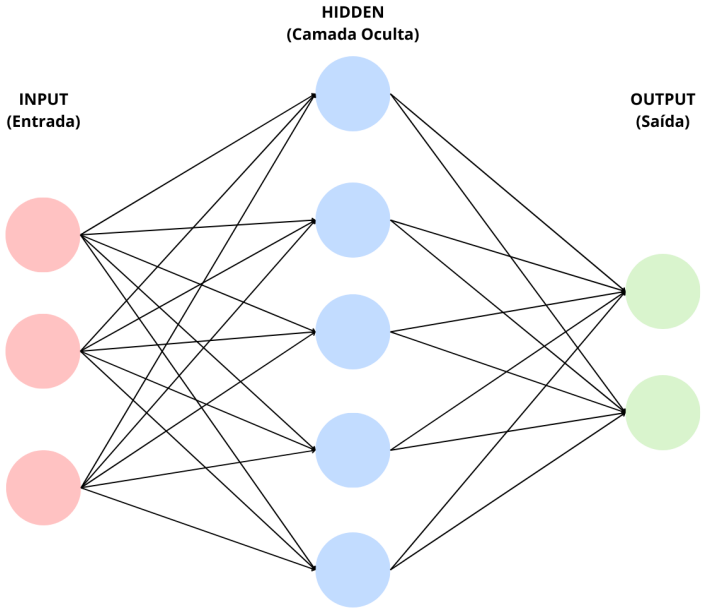

Figura 2.2: Uma ilustração simplificada de uma rede CNN.

Para resolver essa limitação, as LSTM surgiram como uma solução poderosa. A LSTM, uma arquitetura especializada de Redes Neurais Recorrentes (RNN), foi desenvolvida por Hochreiter & Schmidhuber (1997) para superar os desafios de aprendizado de dependências de longo prazo [9]. Em RNN tradicionais, a propagação do erro pode resultar na explosão ou dissipação do gradiente, dificultando o aprendizado de relações distantes. Em contraste, as células de memória da LSTM (com seus portões de entrada, esquecimento e saída), ilustradas na Figura 2.3, garantem um fluxo de erro mais constante, permitindo a retenção de informações relevantes por longos períodos. Seus componentes principais, o portão de entrada (*Input Gate*), o portão de esquecimento (*Forget Gate*) e o portão de saída (*Output Gate*), são essenciais para regular o fluxo de informação e manter as dependências temporais.

A combinação dessas arquiteturas impulsionou ainda mais o desempenho no NER. De acordo com Huang *et al.* (2015), o desenvolvimento de modelos baseados em LSTM tem-se mostrado eficaz para tarefas de marcação de sequência no PLN [10]. O modelo BiLSTM-CRF, uma combinação de LSTM bidirecionais e CRF, foi projetado para capturar características passadas e futuras de uma sequência, além de considerar informações no nível de sentença [10]. Isso significa que, para cada palavra, o modelo tem acesso não apenas ao que veio antes (através da LSTM *forward*), mas também ao que vem depois dela (através da LSTM *backward*), fornecendo um contexto muito mais rico e completo. A integração com o CRF na camada de saída manteve a vantagem de considerar as relações globais e as dependências entre os rótulos adjacentes, garantindo a coerência das

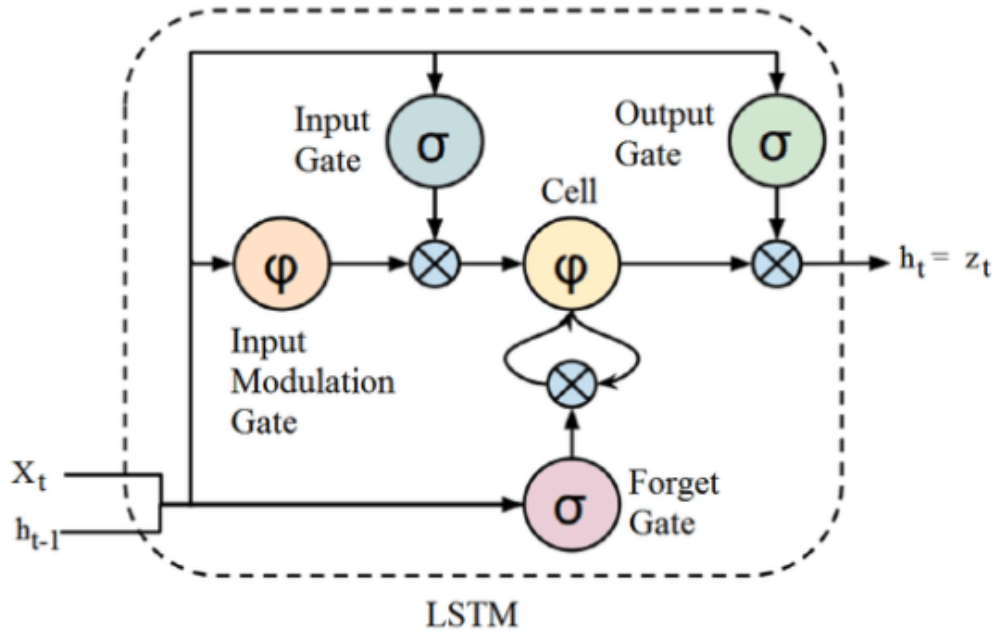

Figura 2.3: Arquitetura de uma célula LSTM

Fonte: Hochreiter e Schmidhuber [9]

previsões e superando significativamente o desempenho dos modelos anteriores [10]. Problemas como etiquetagem de POS, *chunking* e NER fazem parte da tarefa de marcação de sequência e são essenciais para várias aplicações em PLN, como análise de texto, motores de busca e sistemas de recomendação. A Figura 2.4 mostra a estrutura do modelo BiLSTM-CRF, demonstrando como ele combina informações passadas e futuras com uma camada CRF para aumentar a precisão na previsão de etiquetas.

Nos últimos anos, modelos de redes neurais, como redes convolucionais e recorrentes, têm substituído métodos convencionais de marcação de sequência, como HMM e MEMM. Em comparação com abordagens anteriores, esses modelos baseados em *deep learning* demonstram maior robustez e capacidade de generalização. Um desafio, contudo, foi a dificuldade desses modelos em capturar dependências de longo prazo na sequência, uma limitação que o LSTM ajudou a superar. Com suas células de memória especializadas, o LSTM se mostra ideal para tarefas que envolvem longas dependências temporais, como o sequenciamento de palavras em uma sentença [10]. Nesse contexto, foi avaliada a robustez de modelos de marcação de sequência em relação à dependência de características artificiais criadas na engenharia de características. Os resultados indicam que modelos tradicionais, como o CRF, apresentam quedas significativas de desempenho na ausência

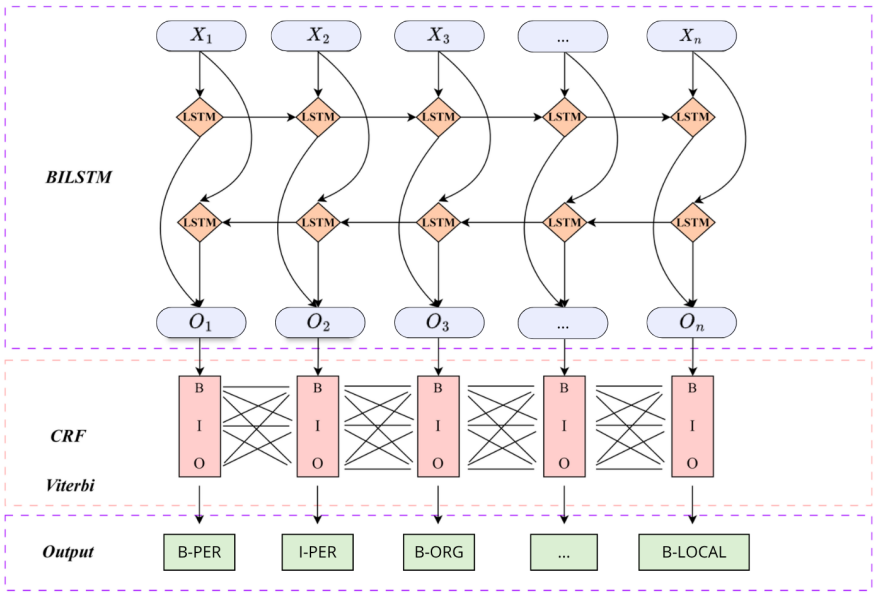

Figura 2.4: Modelo BiLSTM-CRF

Fonte: Chen *et al.* 2025 [41]

dessas características. Em contraste, modelos baseados em LSTM mantêm um desempenho satisfatório, sugerindo uma menor dependência de características artificiais e uma maior capacidade de generalização [10]. A Figura 2.5 apresenta uma variação do modelo BiLSTM-CRF, que incorpora características de máxima entropia, conferindo ainda mais robustez ao modelo.

Com a popularização dos BiLSTM-CRF, o foco migrou para a melhoria da representação de palavras e para a otimização da velocidade, culminando na primeira geração de *embeddings* contextualizados profundos e arquiteturas mais rápidas. O modelo Iterated Dilated Convolutional Neural Network (IDCNN), proposto por Strubell et al. (2017), surge como uma alternativa eficiente às arquiteturas recorrentes, como as LSTMs bidirecionais, para tarefas de rotulação de sequências, como o NER. Diferentemente das redes neurais convolucionais tradicionais, cujo campo receptivo cresce apenas linearmente com o número de camadas – exigindo um empilhamento excessivo para cobrir sequências longas, o cerne do IDCNN reside no uso de convoluções dilatadas (*dilated convolutions*) combinadas com iteração de blocos e compartilhamento de parâmetros [1].

As convoluções dilatadas introduzem espaçamentos (*dilation width* δ) nos filtros, permitindo que o kernel salte sobre tokens e amplie o campo receptivo sem aumentar o

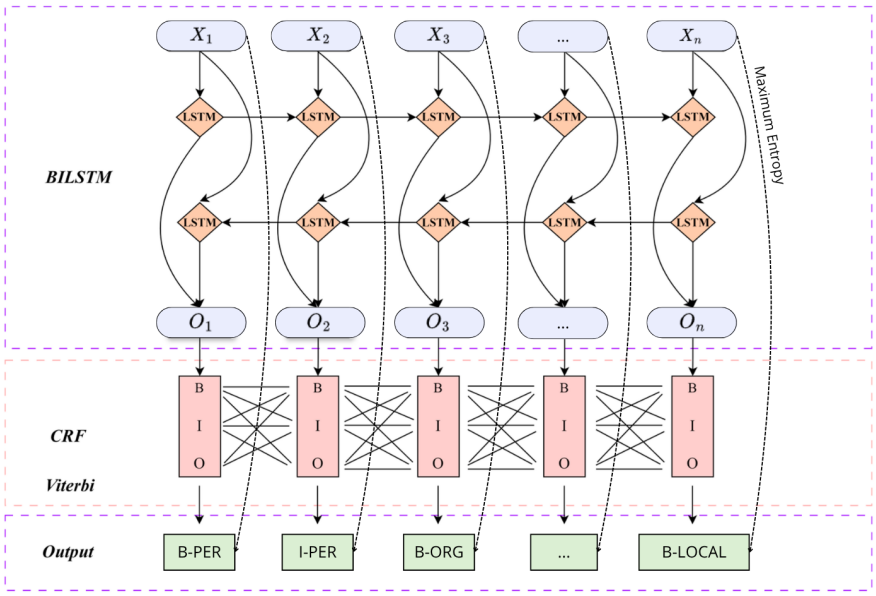

Figura 2.5: Modelo BiLSTM-CRF com características de máxima entropia

Fonte: Adaptado de Chen *et al.* 2025 [41]

número de parâmetros. Formalmente, a operação de convolução dilatada é definida como:

$$c_t = \bigoplus_{k=0}^{r} x_{t \pm k\delta},$$

onde $\oplus$ denota concatenação vetorial e r é o raio do filtro. Ao empilhar camadas com taxas de dilatação que crescem exponencialmente (por exemplo, $\delta = 1, 2, 4, 8, \ldots$), o campo receptivo da rede se expande exponencialmente. Conforme ilustrado na Figura 2.6, com apenas quatro camadas de largura 3 (δ máximo 4), o campo receptivo atinge 31 tokens, mais que o comprimento médio de sentenças no Penn TreeBank (23 tokens), permitindo modelar dependências globais sem perda de resolução por token. Diferentemente das LSTMs, cujo processamento sequencial exige tempo $O(N)$ mesmo com paralelização, as convoluções dilatadas possibilitam a paralelização completa do processamento em GPU's, resultando em ganhos significativos de velocidade: o IDCNN é até 8 vezes mais rápido que uma BiLSTM-CRF em sequências longas (documentos inteiros) e até 14 vezes mais rápido quando combinado com decodificação gulosa, mantendo precisão competitiva. A arquitetura é composta por blocos repetidos de convoluções dilatadas com compartilhamento de parâmetros, refinando progressivamente as representações por meio de um treinamento auxiliar que incentiva predições precisas após cada bloco, o que também alivia o problema do gradiente vanishing [1].

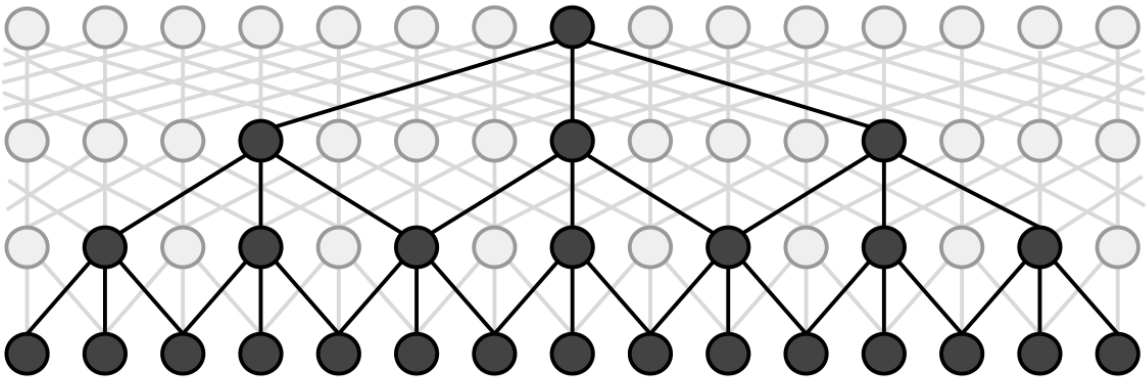

Figura 2.6: Bloco de convoluções dilatadas com largura de filtro 3 e dilatação máxima 4. Os neurônios que contribuem para um neurônio destacado na última camada estão realçados, evidenciando a expansão do campo receptivo. Fonte: Strubell et al. (2017) [1].

Modelos Baseados em *Transformers*

A partir desse marco, o campo de NER evolui com melhorias constantes, incluindo o uso de *embeddings* pré-treinados e variações dos modelos BiLSTM-CRF [10]. Exemplos notáveis são os modelos ELMo [42], *transformer* [43] e BERT [44]. Embora introduzidos em datas próximas, esses modelos se distinguem principalmente pela forma como constroem *embeddings* contextuais a partir dos níveis de uma rede neural.

Apresentado por Vaswani *et al.* (2017), o modelo *transformer* revolucionou a tradução de sequência ao abandonar o uso de redes recorrentes e convolucionais, que anteriormente dominavam tarefas como modelagem de linguagem e tradução automática [43]. O *transformer* utiliza mecanismos de atenção, permitindo que o modelo capture eficientemente as dependências globais de forma paralelizada. Essa abordagem não apenas simplifica a arquitetura, mas também reduz drasticamente o tempo de treinamento, melhorando o desempenho em diversas tarefas e estabelecendo novos padrões para a tradução automática. A arquitetura do *transformer* é baseada no modelo *encoder-decoder*, conforme ilustrado na Figura 2.7. Tanto o codificador quanto o decodificador possuem seis camadas idênticas, compostas por redes neurais totalmente conectadas e subcamadas de atenção multi-cabeças. A normalização da camada ajuda a estabilizar o treinamento, envolvendo cada subcamada com uma conexão residual. Os mecanismos de atenção conectam diretamente todas as posições de entrada e saída a um número constante de operações sequenciais, facilitando o aprendizado das dependências de longo alcance entre as palavras das sequências [43].

Devlin *et al.* (2019) apresentaram o BERT (*Bidirectional Encoder Representations from Transformers*), um modelo de representação de linguagem que revolucionou o PLN [44]. Ao contrário dos modelos unidirecionais anteriores, o BERT é pré-treinado para

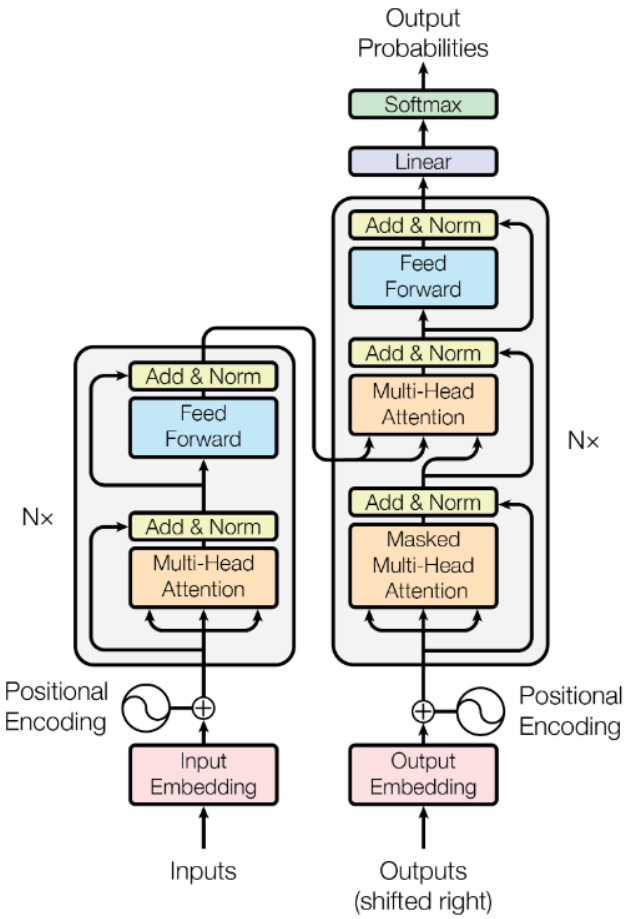

Figura 2.7: Arquitetura Transformer

Fonte: Vaswani *et al.* 2017 [43]

capturar o contexto de palavras em todas as camadas do modelo simultaneamente. Esse processo é possibilitado pelo método conhecido como *Masked Language Model* (MLM), que permite ao modelo aprender bidirecionalmente. Assim, o BERT utiliza uma arquitetura de transformador bidirecional durante o pré-treinamento e ajuste fino, como ilustrado na Figura 2.8. Além do MLM, o BERT também incorpora a tarefa de *Next Sentence Prediction* (NSP), que possibilita ao modelo entender as relações entre pares de sentenças. Essa técnica é essencial para diversas tarefas de PLN, como inferência de linguagem natural e sistemas de perguntas e respostas. O pré-treinamento do BERT é realizado em grandes volumes de texto, como a *Wikipedia*, tornando o modelo altamente eficaz e aplicável a uma ampla gama de tarefas [44]. Os experimentos realizados por Devlin *et al.* (2018) demonstraram a eficácia do BERT em várias tarefas de PLN. Por exemplo, o modelo aumentou a precisão do *benchmark* GLUE para 80,5%, e obteve uma pontuação *F1-Score* de 93,2% no conjunto de dados SQuAD 1.1. Esses resultados foram alcançados

apenas ajustando o modelo [44]. A Figura 2.9 ilustra as representações de entrada e como o BERT organiza os dados, o que é fundamental para o sucesso do ajuste fino.

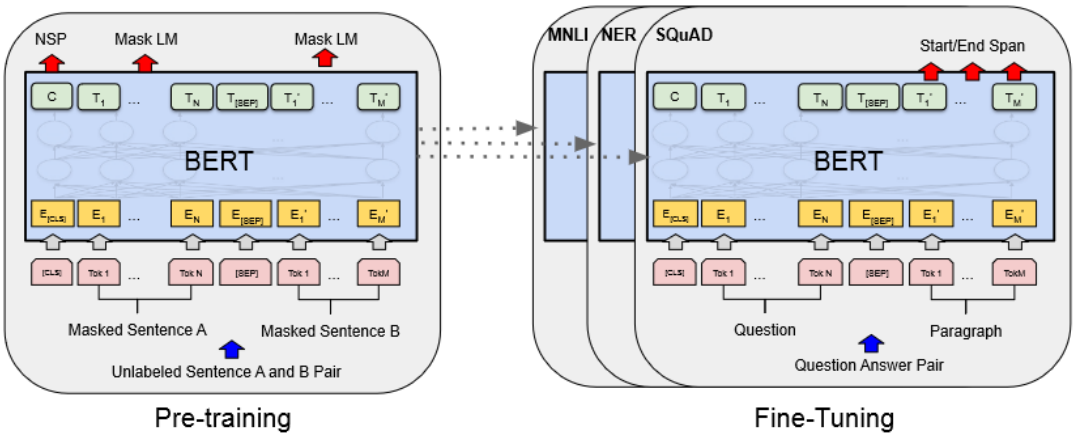

Figura 2.8: Pré-Treinamento e Ajuste fino de uma rede BERT

Fonte: Devlin *et al.* (2018) [44]

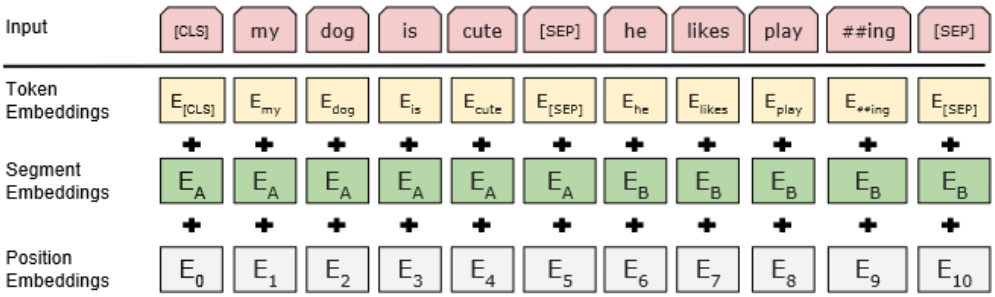

Figura 2.9: Representação de entrada da arquitetura BERT. Os *embeddings* de entrada são a soma dos *embeddings* de token, dos *embeddings* de segmentação e dos *embeddings* de posição.

Fonte: Devlin *et al.* 2018 [44]

O modelo de linguagem *Robustly Optimized BERT Approach* (RoBERTa), desenvolvido pela *Facebook AI Research* (FAIR) em 2019, é uma variação aprimorada do BERT [45]. A principal diferença entre o RoBERTa e o BERT está no volume e na diversidade de dados usados durante o treinamento, requerendo um tempo de treinamento maior e métodos de pré-processamento e treinamento mais sofisticados. Durante esse processo, o RoBERTa emprega mascaramento dinâmico, facilitando a compreensão das conexões entre as palavras e seu contexto. O RoBERTa obteve o melhor desempenho em diversas

tarefas de PLN, como análise de sentimentos, NER e resposta a questões. Seu sucesso consolidou o modelo como uma ferramenta essencial para diversas aplicações *downstream*, tanto na indústria quanto na academia [45]. Além disso, o RoBERTa superou consistentemente modelos pré-treinados anteriores, como o $\text{BERT}_{\text{LARGE}}$ [44] e o $\text{XLNet}_{\text{LARGE}}$ [46], em todas as nove tarefas do *benchmark* GLUE, consolidando-o como uma evolução em relação aos modelos anteriores de pré-treinamento de linguagem [45].

Outro avanço significativo veio com o modelo *Efficiently Learning an Encoder that Classifies Token Replacements Accurately* (ELECTRA), proposto por Clark *et al.* (2020) [2]. O ELECTRA introduz uma abordagem de pré-treinamento mais eficiente do que o MLM tradicional, ao reformular a tarefa de aprendizado como um problema de detecção de tokens substituídos. A arquitetura do ELECTRA é composta por dois *transformers*: um gerador G e um discriminador D. O gerador é tipicamente um modelo pequeno treinado com a tarefa de MLM: ele recebe uma sequência de tokens com algumas posições mascaradas e aprende a prever os tokens originais. Em vez de utilizar o token especial `[MASK]` nas posições corrompidas, o gerador substitui esses tokens por amostras plausíveis, gerando assim uma sequência corrompida $\mathbf{x}^{\text{corrupt}}$. O discriminador, por sua vez, recebe essa sequência e deve classificar, para cada posição t, se o token é original (proveniente do texto real) ou foi substituído pelo gerador. Essa tarefa é denominada *replaced token detection* e está ilustrada na Figura 2.10 [2].

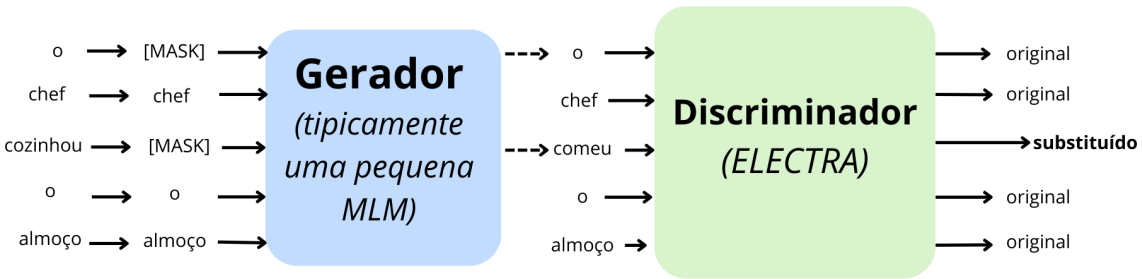

Figura 2.10: Visão geral do replaced token detection. O gerador substitui alguns tokens por amostras; o discriminador aprende a distinguir tokens originais dos substituídos. Adaptado de Clark *et al.* (2020). [2]

Formalmente, dada uma sequência original $\mathbf{x} = [x_1, \ldots, x_n]$, seleciona-se um conjunto de posições $\mathbf{m}$ a serem mascaradas. O gerador G produz, para cada posição $i \in \mathbf{m}$, uma distribuição de probabilidade $p_G(x_i \mid \mathbf{x}^{\text{masked}})$ sobre o vocabulário, onde $\mathbf{x}^{\text{masked}}$ é a sequência com os tokens originais substituídos por `[MASK]`. Amostra-se então um token $\hat{x}_i$ dessa distribuição para compor a sequência corrompida $\mathbf{x}^{\text{corrupt}}$. O discriminador D processa $\mathbf{x}^{\text{corrupt}}$ e produz uma probabilidade $D(\mathbf{x}^{\text{corrupt}}, t)$ de que o token na posição t seja original. O treinamento conjunto minimiza a soma da perda de MLM do gerador e da perda de classificação binária do discriminador [2]:

$$\mathcal{L} = \mathcal{L}_{\text{MLM}}(G) + \lambda \mathcal{L}_{\text{Disc}}(D) \tag{2.2}$$

onde $\mathcal{L}_{\text{Disc}}$ é a entropia cruzada binária que compara as predições do discriminador com os rótulos que indicam se cada token foi substituído ou não. Após o pré-treinamento, o gerador é descartado e apenas o discriminador é ajustado nas tarefas downstream. Diferentemente do MLM, que calcula perda apenas sobre os tokens mascarados (geralmente 15% da sequência), o discriminador do ELECTRA é treinado para classificar todos os tokens da sequência. Isso faz com que o modelo receba um sinal de aprendizado muito mais denso, tornando o pré-treinamento mais eficiente em termos computacionais. Experimentos demonstram que o ELECTRA atinge desempenho comparável ou superior a modelos como RoBERTa e XLNet utilizando apenas um quarto do custo computacional de treinamento [2]. Essa eficiência é particularmente notável em cenários com recursos limitados: por exemplo, um modelo ELECTRA-Small treinado em uma única GPU por quatro dias supera o GPT (que usou 30 vezes mais computação) no benchmark GLUE. Esses resultados evidenciam que a tarefa discriminativa de distinguir dados reais de amostras negativas desafiadoras é mais eficiente em termos de amostras e parâmetros do que as abordagens gerativas tradicionais para aprendizado de representações de linguagem.

O XLNet, proposto por Yang *et al.* (2019), é um modelo de linguagem autoregressivo generalizado que combina as vantagens de abordagens autoregressivas e de autoencodagem [46]. Diferentemente do BERT, que utiliza máscaras para reconstruir tokens corrompidos, o XLNet emprega uma modelagem de linguagem por permutação, maximizando a verossimilhança esperada sobre todas as ordens possíveis de fatorização da sequência. Essa abordagem permite captar contextos bidirecionais sem introduzir símbolos artificiais como *[MASK]*, eliminando a discrepância entre pré-treinamento e ajuste fino. Além disso, o XLNet integra técnicas avançadas do *Transformer-XL* [47], como: mecanismo de recorrência de segmentos, que permite reutilizar estados ocultos de segmentos anteriores, essencial para modelar textos longos; e codificação posicional relativa, que substitui codificações absolutas por relações posicionais entre *tokens*, melhorando a generalização. Em experimentos comparativos, o XLNet superou o BERT em 20 tarefas de PLN, incluindo

compreensão de leitura (SQuAD), inferência textual (MNLI) e classificação de documentos (ClueWeb09-B), muitas vezes por margens significativas. Por exemplo, no conjunto de dados RACE, o XLNet alcançou 85,4% de acurácia, contra 83,2% do RoBERTa e 72,0% do BERT [46]. Essa superioridade é atribuída à capacidade de modelar dependências de longo alcance e contextos bidirecionais densos.

O modelo de linguagem ALBERT, proposto por Lan *et al.* (2019), foi desenvolvido para reduzir o consumo de memória e acelerar o treinamento de modelos BERT, mantendo ou até melhorando o desempenho [48]. O ALBERT introduz duas técnicas principais de redução de parâmetros: (i) fatoração dos embeddings, que decompõe a matriz de embeddings de vocabulário em duas matrizes menores, separando a dimensão do vocabulário da dimensão oculta; e (ii) compartilhamento de parâmetros entre camadas, onde todas as camadas do transformer compartilham os mesmos pesos, reduzindo drasticamente o número total de parâmetros. Além disso, o ALBERT substitui a tarefa de NSP por uma tarefa de *sentence-order prediction* (SOP), que foca na coerência entre segmentos.

Souza *et al.* (2020) apresentaram o *Pretrained BERT Models for Brazilian Portuguese* (BERTimbau), uma versão adaptada da arquitetura BERT, projetada especificamente para o português brasileiro [12]. O objetivo principal do BERTimbau é aprimorar o desempenho em tarefas de PLN em português, aproveitando um vasto conjunto de dados que inclui textos da internet, artigos de jornais e obras literárias, todos voltados para a realidade linguística do Brasil. O BERTimbau mantém a estrutura do BERT base, com 12 camadas, 768 unidades ocultas e 12 cabeças de atenção, totalizando cerca de 110 milhões de parâmetros na versão base. Durante o treinamento, o modelo utiliza técnicas como o MLM e o NSP, métodos que permitem ao BERTimbau capturar a semântica e o contexto das palavras em português, melhorando seu desempenho em diversas tarefas de PLN [12].

O modelo ALBERTina, desenvolvido por Rodrigues *et al.* (2023), adapta a arquitetura ALBERT para o processamento avançado do português europeu e brasileiro [49]. Essa versão é leve e eficiente, tornando-se uma excelente opção para tarefas de PLN em português. Para garantir uma interpretação semântica precisa, o ALBERTina foi treinado com um extenso conjunto de dados, incluindo o *brWaC*, um *corpus* de 2,7 bilhões de tokens em português brasileiro, e os conjuntos de dados *Open Super-large Crawled ALMAnaCH corpus* (OSCAR) [50, 51] e DCEP [52], que abrangem amplamente o português europeu. Essa diversidade de dados permite ao modelo capturar nuances e variações regionais, proporcionando uma compreensão abrangente do português em diferentes contextos. Utilizando camadas de autoatenção, o modelo realiza a codificação contextual das palavras, interpretando-as conforme o contexto em que ocorrem. Além disso, o ALBERTina emprega métodos de compactação, como a divisão de *embeddings* e a parametrização

conjunta, que otimizam o uso de memória sem comprometer o desempenho.

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta uma revisão sistemática da literatura com o objetivo de mapear o estado da arte no PLN aplicado ao domínio jurídico e legislativo brasileiro. A seleção das obras fundamentais seguiu critérios rigorosos baseados em três pilares metodológicos: a relevância cronológica e o pioneirismo no desenvolvimento de *datasets* anotados para o português brasileiro; a representatividade de diferentes subdomínios (como decisões de tribunais superiores, processos de instâncias estaduais e documentos legislativos); e a evolução arquitetural dos modelos empregados, abrangendo desde abordagens estatísticas clássicas e redes neurais recorrentes até as mais recentes arquiteturas baseadas em transformadores e modelos de linguagem massivamente pré-treinados. Dessa forma, constrói-se um panorama crítico que fundamenta as escolhas de design e as hipóteses investigadas nesta pesquisa.

A crescente demanda por soluções de PLN aplicadas ao domínio jurídico brasileiro impulsionou o desenvolvimento de modelos de linguagem especializados, capazes de capturar as nuances terminológicas, estilísticas e semânticas que caracterizam os textos legais. A linguagem jurídica possui particularidades que a distinguem da linguagem geral: vocabulário técnico específico, estrutura formal, frequente citação de leis e jurisprudências, e construções sintáticas complexas [13]. Essas características tornam os modelos de propósito geral, ainda que robustos, para tarefas no domínio jurídico, justificando a criação de modelos especializados por meio de pré-treinamento ou ajuste fino em *corpora* jurídicos de grande escala [13, 53].

Luz de Araújo *et al.* (2018) constituem uma contribuição seminal com a criação do LeNER-BR, o primeiro conjunto de dados dedicado ao NER em português brasileiro. Este marco metodológico compreende 70 documentos jurídicos meticulosamente anotados, incluindo 66 decisões proferidas por tribunais de elevada instância como o STF e STJ, além de quatro documentos legislativos emblemáticos. O esquema de anotação organiza as entidades em categorias gerais (“ORGANIZACAO”, “PESSOA”, “TEMPO”, “LOCAL”)

e específicas do domínio jurídico ("LEGISLACAO", "JURISPRUDENCIA"). Para a implementação, os autores utilizaram a arquitetura BiLSTM-CRF com *word embeddings* baseados em representações *GloVe* pré-treinadas [7].

Os resultados demonstraram a eficácia da abordagem, com o modelo alcançando um *F1-Score* médio de $92,53\%$ no LeNER-BR. O desempenho foi particularmente superior nas entidades "LEGISLACAO" $(97,04\%)$ e "JURISPRUDENCIA" $(88,82\%)$. Além disso, a validação inicial no *datasets* Paramopama mostrou superioridade sobre o modelo ParamopamaWNN, com ganhos expressivos de $2,48\%$ e $4,58\%$ em *F1-Score* em diferentes conjuntos de teste, consolidando o modelo como uma referência para a época.

Entretanto, Luz de Araújo *et al.* (2018) apontam como limitação a performance inferior na entidade "LOCAL" $(60,54\%)$, atribuída à baixa frequência de ocorrências e a ambiguidades contextuais que geram erros de rotulagem. Adicionalmente, os pesquisadores ressaltam que o estudo se restringiu exclusivamente à arquitetura BiLSTM-CRF, não explorando o potencial de modelos baseados em transformadores, que poderiam capturar melhor o contexto e elevar o desempenho em entidades mais raras [7].

Polo *et al.* (2021) apresentam uma contribuição fundamental com o desenvolvimento do LegalNLP, uma iniciativa brasileira que reúne métodos, conjuntos de recursos e ferramentas de PLN aplicados ao domínio jurídico, incluindo versões especializadas de modelos baseados na arquitetura BERT para classificação de textos e reconhecimento de entidades [54]. O treinamento utilizou quatro grandes conjuntos textuais totalizando 25,77 GB de dados, provenientes de múltiplos tribunais, com destaque para o TJSP. A equipe desenvolveu uma gama abrangente de modelos, incluindo Phraser, Word2Vec, Doc2Vec, FastText e o BERTikal — um modelo BERT-base *cased* especializado em textos legais [54]. A importância do projeto reside na sistematização e disponibilização de modelos adaptados ao português jurídico, contribuindo significativamente para a reprodutibilidade e o avanço das pesquisas na área.

A avaliação foi realizada em tarefas de classificação de status processual utilizando o conjunto *Brazilian Legal Proceedings*. Os experimentos revelaram que a combinação de CatBoost com embeddings do BERTikal ou Doc2Vec alcançou acurácia de 0,86 e F1 macro de 0,82. Tais resultados mostraram-se competitivos quando comparados a *benchmarks* consolidados como o Word2Vec do NILC/USP e o BERTimbau, validando a biblioteca como uma ferramenta robusta para o domínio.

Como limitações, Polo *et al.* (2021) reconhecem que os ganhos de desempenho em relação a modelos genéricos foram modestos, indicando que a linguagem jurídica pode não diferir tão drasticamente da geral em tarefas de classificação de alto nível. Além disso, os autores observam que a avaliação foi restrita à classificação, sem explorar tarefas como NER ou extração de relações. Destaca-se ainda que o estudo não comparou o BERTikal

com outras arquiteturas de transformadores como RoBERTa ou DeBERTa [54].

Albuquerque *et al.* (2022) avançaram na investigação com o UlyssesNER-Br, um *datasets* focado em documentos legislativos da Câmara dos Deputados. A base compreende 150 projetos de lei e 800 consultas legislativas, estruturando entidades em sete categorias baseadas no padrão HAREM, divididas entre classes gerais e específicas do domínio ("FUNDAMENTO" e "PRODUTODELEI"). Metodologicamente, foram testados os modelos HMM, CRF e BiLSTM-CRF [16].

Os resultados indicaram a superioridade do modelo clássico CRF, que atingiu um *F1-Score* médio de $80,8\%$, enquanto o BiLSTM-CRF obteve $76,89\%$. A categoria "FUNDAMENTO" obteve $81,83 \pm 5,87$, um resultado relevante para o domínio legislativo, embora inferior em termos absolutos a estudos realizados em outros tipos de documentos jurídicos.

Quanto às limitações, Albuquerque *et al.* (2022) salientam que a performance geral ficou aquém do observado no LeNER-BR, indicando a necessidade de refinar o esquema de anotação e melhorar a qualidade do *datasets*. Outro ponto crítico é que a investigação limitou-se a modelos tradicionais e neurais baseados em BiLSTM, deixando de lado arquiteturas baseadas em transformadores, como o BERT, que poderiam lidar melhor com a complexidade do texto legislativo [16].

Costa *et al.* (2022) expandiram o escopo do UlyssesNER-Br ao criar o C-*datasets*, que integra textos informais de cidadãos ao domínio formal. A pesquisa investigou o impacto da mistura de domínios no desempenho de sistemas de NER, utilizando um processo rigoroso de anotação com três equipes independentes para garantir a confiabilidade. Foram avaliados comparativamente modelos CRF, BiLSTM-CRF e BERT (*fine-tuning* do BERTimbau) [19].

Os experimentos demonstraram a superioridade do BERT, que alcançou um *F1-Score* de $73,90\%$ na análise por categorias. O estudo provou que a união estratégica de dados formais e informais melhora significativamente a generalização dos modelos, especialmente para textos escritos por leigos, onde o ganho de predição foi substancial após o treinamento conjunto.

Contudo, Costa *et al.* (2022) destacam que o *F1-Score* geral ainda é considerado baixo, refletindo a alta complexidade e ruído inerentes aos textos informais. Além disso, os autores apontam que a investigação se restringiu ao BERTimbau, não explorando modelos jurídicos especializados ou transformadores mais recentes que poderiam oferecer uma compreensão semântica mais profunda desses textos desafiadores [19].

Brito *et al.* (2023) elevaram o padrão de qualidade do setor com o CDJUR-BR, uma coleção padrão-ouro com 1.216 documentos do TJCE. O esquema de anotação é altamente granular, com sete categorias principais e 21 subcategorias específicas, anotadas por juízes

e promotores. A pesquisa delineou cinco cenários experimentais para testar a robustez do *dataset* frente a modelos como *spaCy*, BiLSTM-CRF e BERT [15].

O modelo BERT obteve o melhor desempenho, com *F1-Score* macro de $0,58$, destacando-se em entidades específicas como "PES-AUTORID-POLICIAL" $(0,90)$. Um resultado crucial foi a capacidade de generalização: o modelo treinado no CDJUR-BR e testado no LeNER-BR obteve *F1* de $0,68$, superando o teste inverso $(0,56)$ e provando a robustez do novo *datasets*.

Em contrapartida, Brito *et al.* (2023) observam que o coeficiente Kappa global de 0,69 sugere ambiguidades residuais que exigem revisões futuras no esquema de anotação. Além disso, algumas entidades apresentaram desempenho insatisfatório devido à sua complexidade. Os autores também ressaltam como limitação a não utilização de modelos mais recentes e específicos, como o LegalBert-pt, que poderiam ter potencializado os resultados [15].

Silveira *et al.* (2023) desenvolveram o LegalBERT-pt, um modelo de linguagem projetado especificamente para o português brasileiro e para o processamento de documentos jurídicos, como petições, sentenças e decisões judiciais. O modelo é apresentado em duas variações estruturais que seguem a arquitetura BERT-base (com 12 camadas, 768 unidades ocultas, 12 cabeças de atenção e aproximadamente 110 milhões de parâmetros): a primeira, denominada LegalBERT-pt SC (*from scratch*), foi treinada do zero exclusivamente com textos jurídicos; a segunda, LegalBERT-pt FP (*fine-tuning from pretrained*), parte do BERTimbau [12] e realiza uma adaptação progressiva ao domínio legal por meio de ajuste fino. O treinamento utilizou um *corpus* extenso composto por aproximadamente 1,5 milhão de documentos provenientes de diversas instâncias judiciais. O vocabulário foi construído com 36.345 unidades sublexicais via técnicas de *Byte-Pair Encoding* (BPE) e *SentencePiece*, adaptadas ao léxico jurídico brasileiro [13].

A variante LegalBert-pt FP demonstrou superioridade consistente em avaliações intrínsecas e extrínsecas. Experimentos realizados em tarefas de NER utilizando os *corpora* LeNER-BR e CDJUR-BR demonstraram que o LegalBERT-pt FP supera significativamente o BERTimbau e o LegalBERT-pt SC. No LeNER-BR, o modelo superou o BERTimbau-Base em $1,75\%$ no *F1* macro; no CDJUR-BR, o incremento foi de $3,78\%$, evidenciando a robustez da estratégia de adaptação progressiva para o domínio jurídico.

Todavia, Silveira *et al.* (2023) admitem que os ganhos de performance ainda são limitados em entidades muito complexas e que a adaptação de domínio pode não cobrir todas as nuances de subdomínios jurídicos hiperespecíficos. Adicionalmente, os autores apontam como limitação o foco exclusivo na família BERT, sem realizar comparações diretas com arquiteturas como RoBERTa ou DeBERTa, que poderiam oferecer diferentes vantagens de representação [13].

Nunes *et al.* (2024) propuseram uma abordagem inovadora para lidar com a escassez de dados rotulados, combinando autoaprendizagem e amostragem ativa no domínio legislativo. Utilizando o UlyssesNER-Br e dados não rotulados da Câmara dos Deputados, os autores integraram o BERTimbau em um processo iterativo onde o modelo incorpora sentenças com alta confiança preditiva ao seu próprio conjunto de treinamento [55].

A metodologia provou ser altamente eficaz, alcançando um *F1-Score* médio de $86,70 \pm 2,28\%$, superando significativamente o BERTimbau padrão. Ganhos notáveis foram observados em categorias críticas como "LOCAL" e "ORGANIZACAO", além da categoria "EVENTO", que saltou de uma performance nula para $58,10 \pm 34,16\%$, validando a estratégia de autoaprendizagem para domínios com poucos dados.

Como limitações, Nunes *et al.* (2024) alertam que a técnica depende criticamente da definição de limiares de confiança para evitar a propagação de erros. Além disso, a entidade "EVENTO" ainda apresenta alta variabilidade e desempenho abaixo do ideal. Por fim, os autores destacam que o estudo utilizou apenas o BERTimbau como base, sem investigar se transformadores maiores ou específicos do domínio ponderiam otimizar ainda mais o processo de autoaprendizagem [55].

Garcia *et al.* (2024) apresentam uma contribuição de grande escala para o processamento de linguagem natural jurídico com o desenvolvimento do RoBERTaLexPT. O modelo baseia-se na arquitetura RoBERTa e foi pré-treinado especificamente para o domínio jurídico em português utilizando o *LegalPT*, o maior *dataset* jurídico em português disponível, com aproximadamente 125,1 GiB (cerca de 1,5 milhão de documentos). Um diferencial metodológico relevante da obra é a ênfase dada ao tratamento de duplicação de documentos — um desafio frequente em *corpora* jurídicos —, aplicando técnicas rigorosas de deduplicação via algoritmos *MinHash* e *Locality Sensitive Hashing* (LSH) para garantir maior diversidade e representatividade, filtrando altas taxas de redundância documental (50,63%) identificadas nos dados judiciais. Além disso, os autores agregaram o *PortuLex benchmark*, um conjunto de referência para avaliação composto por tarefas como NER, classificação de textos e similaridade semântica [53].

Em termos de desempenho, Garcia *et al.* (2024) demonstraram que o $\text{RoBERTaLexPT}_{\text{base}}$ superou modelos significativamente maiores, como o $\text{BERTimbau}_{\text{large}}$ e o $\text{Albertina-PT-BR}_{\text{large}}$, além de predecessores específicos do domínio como o BERTikal e o JurisBERT, comprovando que o pré-treinamento em dados jurídicos, combinado com dados gerais, produz modelos mais eficazes para tarefas legais. A variante estendida $\text{RoBERTaLexPT+}_{\text{base}}$ alcançou 85,46% no *PortuLex*, mantendo-se competitiva mesmo em *benchmarks* genéricos como ASSIN2 e HAREM.

Entretanto, os autores observaram uma limitação importante na escalabilidade: o aumento do modelo para a versão *large* não resultou em ganhos substanciais de performance,

o que sugere que o *datasets* jurídico atual pode ainda ser insuficiente para alimentar modelos de maior capacidade paramétrica. Além disso, o foco exclusivo na família RoBERTa deixou em aberto o potencial de outras arquiteturas, como T5 ou ELECTRA, e o *benchmark PortuLex* permanece restrito a tarefas de classificação e NER, não contemplando aplicações como sumarização ou resposta a perguntas [53].

A Tabela 3.1 sintetiza cronologicamente as principais contribuições mapeadas nesta revisão da literatura, destacando os conjuntos de dados, arquiteturas e métricas alcançadas pelos respectivos autores.

Tabela 3.1: Principais trabalhos em NER jurídico para o português brasileiro.

Artigo	**Dataset / Corpus**	**Modelos Utilizados**	**Ent.**	**Resultados (F1-Score)**
Luz de Araújo *et al.* (2018) [7]	LeNER-BR (70 docs)	BiLSTM-CRF + GloVe	6	Média: 92,53%
Polo *et al.* (2021) [54]	LegalNLP (25,77 GB)	Phraser, Word2Vec, Doc2Vec, FastText, BERTikal	–	Classificação: Word2Vec+CNN (0,80), BERTikal+CatBoost (0,82)
Albuquerque *et al.* (2022) [16]	UlyssesNER-Br (950 docs)	HMM, CRF, BiLSTM-CRF	7	CRF: 80,8%; BiLSTM-CRF: 76,89%
Costa *et al.* (2022) [19]	C-datasets (expansão)	CRF, BiLSTM-CRF, BERTimbau	7	BERT: 73,90% (categorias)
Brito *et al.* (2023) [15]	CDJUR-BR (1.216 docs)	spaCy, BiLSTM-CRF, BERT	21	BERT: macro 0,58
Silveira *et al.* (2023) [13]	1,5M docs jurídicos	LegalBERT-pt (SC e FP)	–	LeNER-BR: +1,75%; CDJUR-BR: +3,78% (F1 macro vs BERTimbau)
Nunes *et al.* (2024) [55]	UlyssesNER-Br	BERTimbau + Autoaprendizagem	7	Média: 86,70%
Garcia *et al.* (2024) [53]	LegalPT (125,1 GiB)	RoBERTa (base, large) + MinHash-LSH	4 tar.	PortuLex: $\text{RoBERTaLexPT}_{\text{base}}$ 85,41%

A evolução dos modelos para NER mapeada nesta literatura reflete uma trajetória de crescente sofisticação técnica: partindo de modelos estatísticos baseados em características manuais (como HMM e CRF) para arquiteturas neurais capazes de aprender representações automaticamente (CNN, LSTM, BiLSTM-CRF), culminando nos modelos de linguagem pré-treinados baseados em *transformers* (BERT, RoBERTa) e suas respectivas adaptações para domínios específicos e para o idioma português (BERTimbau, LegalBERT-pt, RoBERTaLexPT). Cada nova geração trouxe ganhos significativos em acurácia, capacidade de generalização e eficiência computacional.

No contexto do português brasileiro e do domínio jurídico, a disponibilidade de modelos especializados abre novas possibilidades para o desenvolvimento de sistemas de NER robustos. A combinação de modelos pré-treinados em português com ajuste fino em *corpora* anotados de referência (como o LeNER-BR e o CDJUR-BR) tem se mostrado a estratégia mais promissora do estado da arte. O presente trabalho insere-se exatamente nessa linha de investigação, explorando o impacto e os limites desses modelos de linguagem avançados frente às peculiaridades linguísticas do ecossistema jurídico nacional.

Capítulo 4

Proposta de Pesquisa

Este capítulo apresenta o detalhamento metodológico que norteiam o desenvolvimento dos modelos de NER para o domínio jurídico brasileiro. O texto está estruturado de forma a conduzir o leitor desde a concepção do projeto até os detalhes técnicos de sua execução. Inicialmente, a Seção 4.1 contextualiza a proposta de pesquisa, fundamentada na lacuna de estudos em português e no potencial das abordagens híbridas.

A Seção 4.2 descreve o fluxo metodológico completo, enquanto as Seções 4.2.1, 4.2.2 e 4.2.3 detalham, respectivamente, os *dataset* utilizados (LeNER-BR, UlyssesNER-BR e CDJUR-BR), as etapas de limpeza e tokenização dos dados, e a arquitetura dos 15 modelos avaliados. Por fim, as Seções 4.2.4 e 4.3 apresentam o protocolo de avaliação estatística adotado e as especificações da infraestrutura computacional e hiperparâmetros que garantem a reprodutibilidade dos experimentos.

4.1 Projeto

Conforme demonstrado por Souza et al. (2020), avanços recentes em abordagens híbridas para o NER na língua portuguesa têm alcançado resultados promissores em conjuntos de dados de domínio geral, especialmente aqueles que combinam abordagens baseadas em *transformers* com abordagens clássicas [56]. Tais desenvolvimentos são particularmente relevantes no contexto atual do PLN, onde a necessidade de sistemas robustos e eficientes para a identificação de entidades em textos é crescente. Motivada por esses progressos e pela lacuna de estudos aprofundados nessa área para o domínio jurídico, este trabalho propõe estender e aprofundar a aplicação de abordagens híbridas especificamente para o NER em português brasileiro [56].

O objetivo é aplicar essas técnicas a conjuntos de dados jurídicos públicos em português, notadamente LeNER-BR, CDJUR-BR e UlyssesNER-BR, a fim de avaliar sua eficácia e potencial de melhoria no desempenho do NER. O NER apresenta desafios úni-

cos que o diferenciam do NER de domínio geral, incluindo a complexidade da terminologia jurídica, a variabilidade estrutural dos documentos legais e a sensibilidade do contexto. A integração de diferentes paradigmas metodológicos, como as abordagens híbridas, surge como uma estratégia promissora para superar essas dificuldades. A combinação da capacidade das abordagens clássicas em capturar padrões sequenciais e a riqueza das representações contextuais fornecidas pelas abordagens baseadas em *transformers* pode resultar em um desempenho superior na identificação precisa de entidades legais. O arranjo experimental desta pesquisa explora uma variedade de combinações de modelos, integrando tanto abordagens clássicas quanto aquelas baseadas em *transformers*, com o intuito de otimizar o desempenho do NER. Uma visão geral das abordagens exploradas é apresentada na Figura 4.1, delineando o fluxo de combinação de diferentes abordagens [57].

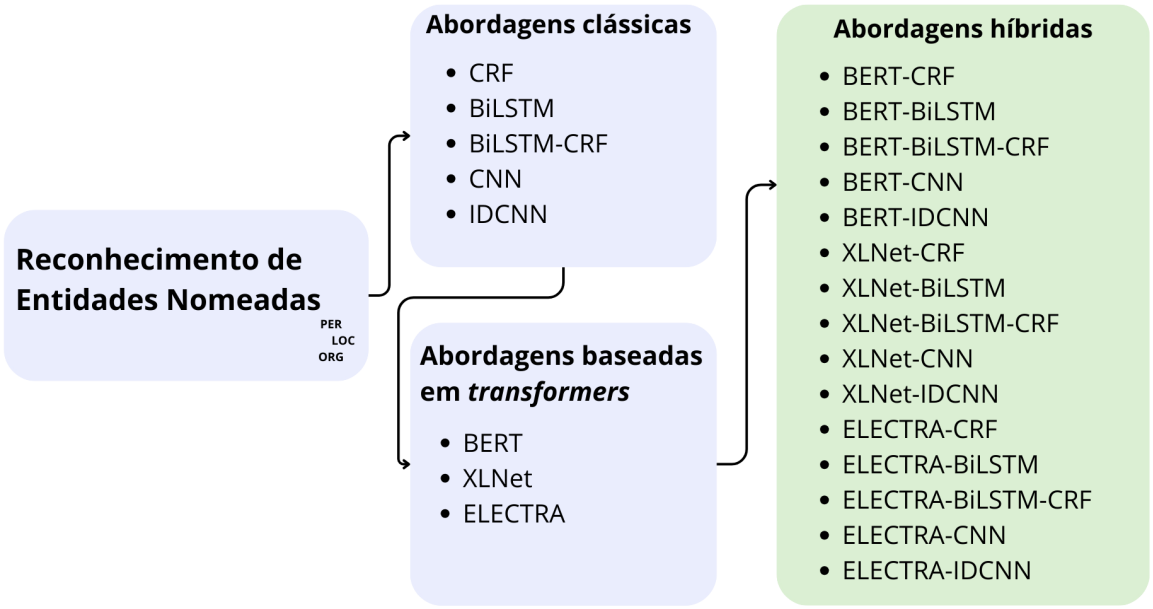

Figura 4.1: Abordagens exploradas para NER

A seleção dos modelos foi embasada na sua comprovada eficácia em pesquisas prévias sobre NER e NER para a língua portuguesa. As abordagens clássicas, como CRF e BiLSTM, e suas combinações (BiLSTM-CRF), foram escolhidas devido à sua robusta capacidade de capturar padrões sequenciais e dependências de longo alcance em dados textuais [58, 59, 60]. Essas técnicas são particularmente eficazes na modelagem de relações contextuais locais, o que é crucial para a identificação de limites de entidades em sequências de texto. Adicionalmente, este trabalho incorpora as CNN e as IDCNN como abordagens clássicas complementares.

As CNNs são reconhecidas por sua eficiência na extração de características locais e padrões morfológicos, como informações de prefixos e sufixos em nível de caractere,

enquanto as IDCNNs, por meio de convoluções dilatadas, expandem o campo receptivo sem perder resolução, permitindo capturar dependências de maior alcance com custo computacional reduzido [1]. Essas características tornam tais modelos particularmente úteis para complementar as representações contextuais obtidas por outras arquiteturas.

Por outro lado, os modelos baseados em *transformers*, como BERT e XLNet, foram selecionados pela sua capacidade de gerar representações contextuais ricas e profundas, que codificam informações semânticas e sintáticas complexas de forma bidirecional. Além desses, incluímos também o modelo ELECTRA, que se destaca por seu pré-treinamento eficiente por meio de uma tarefa de detecção de tokens substituídos, proporcionando representações de alta qualidade com menor custo computacional. A capacidade desses modelos de pré-treinamento em grandes *dataset* e de *fine-tuning* em tarefas específicas de PLN tem demonstrado resultados de estado da arte em diversas aplicações, incluindo NER [2].

A combinação dessas duas classes de abordagens clássicas e *transformers* em modelos híbridos visa explorar o melhor de ambos os mundos: a sensibilidade sequencial dos modelos clássicos complementada pela compreensão contextual abrangente dos *transformers*. Essa sinergia é esperada para mitigar as limitações individuais de cada abordagem e, consequentemente, aprimorar a precisão e a robustez do NER em português brasileiro. Para isso, todas as combinações possíveis entre os três *transformers* (BERT, XLNet, ELECTRA) e as cinco arquiteturas clássicas (CRF, BiLSTM, BiLSTM-CRF, CNN, IDCNN) serão exploradas, totalizando quinze modelos híbridos.

A escolha por essa variedade de combinações fundamenta-se na hipótese de que diferentes arquiteturas clássicas podem capturar aspectos complementares das sequências textuais, como dependências temporais (BiLSTM), restrições de rótulos (CRF) ou padrões locais (CNN/IDCNN), enquanto os *transformers* fornecem *embeddings* contextuais robustos. Essa diversidade de combinações visa identificar quais sinergias são mais eficazes para o domínio jurídico. A seleção focou em técnicas já validadas em estudos relacionados, garantindo a reprodutibilidade e a aplicabilidade dos resultados no domínio jurídico [30, 29, 61, 62].

A seleção dos conjuntos de dados, LeNER-BR [7], CDJUR-BR [15] e UlyssesNER-BR [16], foi embasada por sua disponibilidade pública e pela diversidade de seu conteúdo, que abrange diferentes tipos de textos jurídicos [55]. Essa diversidade é fundamental para avaliar o potencial de generalização dos modelos em subdomínios jurídicos variados, desde petições e sentenças até acórdãos e ementas. A disponibilidade pública desses recursos é um fator crucial para a reprodutibilidade e a transparência da pesquisa, permitindo que outros pesquisadores validem e expandam os resultados obtidos. A inclusão de múltiplos conjuntos de dados visa também testar a robustez dos modelos em face da heterogeneidade

de estilos, vocabulário e estruturas presentes em diferentes conjuntos de dados jurídicos [55].

A estratégia de avaliação foi meticulosamente embasada para abordar os desafios inerentes ao NER em português, notadamente o desequilíbrio entre as classes de entidades e a variabilidade do domínio. Para garantir uma avaliação abrangente e significativa do desempenho dos modelos, as métricas de *Precision*, *Recall* e *F1-Score* foram escolhidas [15]. Para mitigar o impacto do desequilíbrio de classes, as métricas foram calculadas e, quando apropriado, *Micro-average* ou *Macro-average* foram aplicados entre os tipos de entidades. O *Micro-average* é útil para ter uma visão geral do desempenho, tratando todas as predições como uma única unidade, enquanto o *Macro-average* dá igual peso a cada classe, independentemente do seu tamanho, sendo mais informativo em casos de classes desbalanceadas [63].

Para garantir resultados robustos e representativos, empregou-se a validação cruzada estratificada *k-fold*. Essa técnica divide o conjunto de dados em k subconjuntos (*folds*), mantendo a proporção das classes de entidades em cada *fold*, e repete o processo de treinamento e teste k vezes, usando um *fold* diferente para teste em cada iteração. Isso ajuda a reduzir o viés e a variância da avaliação, proporcionando uma estimativa mais confiável do desempenho do modelo em dados não vistos. Adicionalmente, para permitir comparações estatisticamente fundamentadas entre os modelos, o teste de Friedman [64] foi aplicado, seguido por uma análise *post-hoc* de Nemenyi.

O teste de Friedman é um teste não-paramétrico que avalia se há diferenças significativas entre os resultados de múltiplos modelos em múltiplos conjuntos de dados (ou *folds* na validação cruzada). Se o teste de Friedman indicar uma diferença significativa, a análise *post-hoc* de Nemenyi é utilizada para identificar quais pares de modelos diferem significativamente entre si. Essas análises estatísticas são importantes para validar as conclusões sobre a superioridade de uma abordagem sobre a outra, garantindo que as observações não sejam meramente devido ao acaso. Em conjunto, essas escolhas visam assegurar avaliações justas, confiáveis e interpretáveis, alinhadas com os objetivos de aprimoramento do NER em português brasileiro.

4.2 Metodologia

Esta seção detalha os métodos e processos empregados para atingir os objetivos propostos neste estudo, com uma ênfase particular no desenvolvimento de um modelo robusto de NER especificamente para o português brasileiro. Para fornecer uma visão clara e organizada de toda a pesquisa, o processo é sintetizado no fluxograma apresentado na Figura 4.2, que ilustra cada etapa, desde a fase inicial de obtenção dos dados até a avaliação final

dos modelos. A metodologia se inicia com a etapa de Bases de Dados, onde as fontes de informação mais relevantes para o estudo são definidas e coletadas. Em seguida, procede-se ao Pré-processamento, uma fase que engloba a normalização, limpeza e preparação minuciosa dos textos jurídicos. Essa preparação inclui a conversão dos rótulos de entidades para o formato IOB, uma convenção amplamente reconhecida e adotada em tarefas de NER devido à sua eficácia na representação de entidades em sequências textuais e na garantia da consistência dos dados.

Conforme o delineamento, a etapa de NER é então executada. Esta fase explora diversas estratégias, sendo dividida na aplicação de abordagens clássicas (como CRF, BiLSTM, CNN e IDCNN), baseadas em *transformers* (como BERT, XLNet e ELECTRA) e híbridas, que combinam abordagens baseadas em *transformers* com clássicas. Essas estratégias são exploradas para capitalizar tanto a capacidade de otimização sequencial quanto a rica contextualização semântica inerente aos textos jurídicos, buscando identificar entidades com maior precisão e abrangência.

Finalmente, a etapa de Avaliação é fundamental para validar a eficácia e o desempenho dos modelos desenvolvidos. Para tal, são empregadas métricas padrão do campo de NER, como *Precision*, *Recall* e *F1-Score*. Adicionalmente, são realizados testes de significância estatística para comparar o desempenho dos diferentes modelos e, subsequentemente, procede-se à análise aprofundada das entidades reconhecidas por cada modelo. Tais métricas e análises são cruciais para garantir a confiabilidade, robustez e reprodutibilidade dos experimentos conduzidos. Em suma, este capítulo, alinhado ao fluxograma apresentado, oferece uma abordagem sistemática e coesa, conectando intrinsecamente as diversas etapas da pesquisa ao objetivo central de aprimorar o reconhecimento de entidades em textos jurídicos, notoriamente complexos.

4.2.1 Bases de Dados

Neste estudo, a identificação de EN no contexto jurídico brasileiro é realizada utilizando três bases de dados anotadas, reconhecidas por sua representatividade e abrangência: o LeNER-BR [7], o UlyssesNER-BR [16] e o CDJUR-BR [15]. A escolha desses conjuntos de dados se baseia na sua importância e extensão, fornecendo um fundamento sólido para as análises e experimentos propostos. Vale ressaltar que, até a presente data, esses são os únicos conjuntos de dados abertos disponíveis na literatura especificamente direcionados ao NER em português brasileiro. Cada base explora diferentes aspectos da linguagem e documentação legal, contribuindo para uma avaliação abrangente. O primeiro conjunto de dados, LeNER-BR, foi construído a partir da coleta de 66 documentos legais provenientes de diversos tribunais brasileiros, incluindo o STF, o STJ, o TJMG e o TCU. Adicionalmente, foram incluídos quatro textos legislativos, como a Lei Maria da Penha,

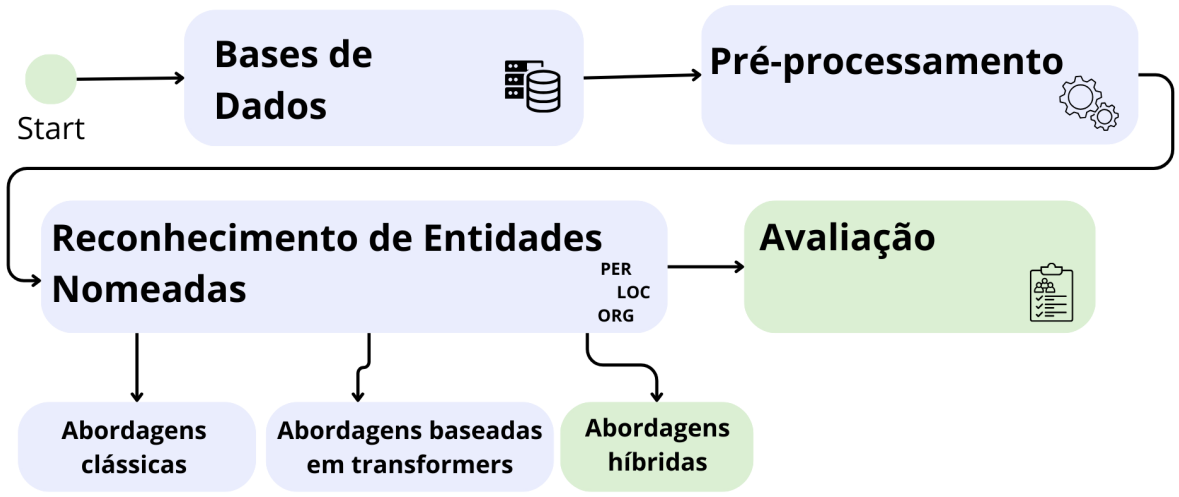

Figura 4.2: Fluxograma da metodologia proposta

totalizando 70 documentos. Para o processamento inicial, utilizou-se a biblioteca `NLTK` para segmentar os textos em frases e tokenizá-las. O resultado final de cada documento foi armazenado em arquivos estruturados, onde cada palavra ocupa uma linha e as frases são separadas por linhas em branco.

A anotação manual das EN foi realizada com o auxílio da ferramenta WebAnno [65]. As categorias de entidades definidas foram: "ORGANIZACAO" (para identificar instituições e órgãos públicos ou privados); "PESSOA" (para nomes próprios de indivíduos); "TEMPO" (para expressões temporais); "LOCAL" (para referências geográficas); "LEGISLACAO" (para leis e normativas); e "JURISPRUDENCIA" (para decisões judiciais). As duas últimas classes, "LEGISLACAO" e "JURISPRUDENCIA", correspondem às categorias "Ato de Lei" e "Decisão" da *Legal Knowledge Interchange Format Ontology* [66], permitindo um alinhamento com modelos ontológicos voltados ao direito. A divisão dos dados seguiu um critério de amostragem aleatória, distribuindo 50 documentos para o conjunto de treinamento, 10 para o conjunto de desenvolvimento e 10 para o teste. O número total de *tokens* no LeNER-BR (318.073 *tokens*) é comparável a outros *dataset* amplamente utilizados em tarefas de reconhecimento de entidades nomeadas, como os conjuntos Paramopama [67] (310.000 *tokens*) e *CONLL-2003 English* [68] (301.418 *tokens*). A Tabela 4.1 apresenta a distribuição das entidades anotadas no LeNER-Br.

Complementando o foco judicial do LeNER-BR, o conjunto de dados UlyssesNER-Br [16] aborda o domínio legislativo. Este conjunto de dados especializado em documentos legislativos brasileiros foi desenvolvido para auxiliar tarefas de NER em português brasi-

Tabela 4.1: Distribuição de Entidades no LeNER-BR.

Entidades	**#Anotações**	%
jurisprudencia	5.370	12,06
legislacao	18.317	41,15
local	1.793	4,03
organizacao	9.646	21,67
pessoa	6.241	14,02
tempo	3.146	7,07
Total	**44.513**	**100**

leiro. Construído a partir de 154 projetos de lei e 800 consultas legislativas da Câmara dos Deputados do Brasil, o *datasets* organiza 7 classes semânticas principais. As sete categorias de entidades no UlyssesNER-Br são: "DATA" (datas específicas mencionadas em textos legislativos); "EVENTO" (eventos de relevância legislativa ou social, como eleições e sessões parlamentares); "FUNDAMENTO" (entidades vinculadas a normas, processos legislativos e instrumentos jurídicos); "LOCAL" (referências geográficas ou virtuais); "ORGANIZACAO" (entidades institucionais envolvidas no processo legislativo ou na sociedade); "PESSOA" (indivíduos ou grupos relacionados ao contexto legislativo); e "PRODUTODELEI" (entidades criadas ou regulamentadas por legislação). A Tabela 4.2 apresenta a distribuição das entidades anotadas no UlyssesNER-Br.

Tabela 4.2: Distribuição de Entidades no UlyssesNER-BR.

Entidades	**#Anotações**	%
data	1.000	6,89
evento	10	0,07
fundamento	6.700	46,18
local	1.500	10,33
organizacao	2.000	13,80
pessoa	1.600	11,02
produtodelei	1.700	11,71
Total	**14.510**	**100**

Expandindo o escopo para o fluxo de trabalho judicial mais amplo, o conjunto de dados CDJUR-BR [15] foi construído a partir de 1.216 documentos legais, como petições, queixas, inquéritos, decisões e sentenças, provenientes do Tribunal de Justiça do Ceará (TJCE). Diferentemente dos conjuntos de dados anteriores, o CDJUR-BR apresenta anotações detalhadas em seis categorias de entidades principais, organizadas em 21 entidades específicas. Estas são agrupadas em três eixos temáticos interligados: Componentes Normativos, Participantes do Processo e Cenários Espaciais e Decisórios. Juristas manualmente anotaram os dados, com validação por meio de adjudicação e confiabilidade

medida usando o Kappa de Cohen. A hierarquia completa das entidades e suas distribuições está apresentada na Tabela 4.3. Essa diversidade de dados suporta uma avaliação robusta do modelo, já que os modelos de NER são treinados em versões pré-processadas desses conjuntos de dados variados.

Tabela 4.3: Distribuição de Entidades no CDJUR-BR.

Categoria	**#Anotações**	**%**	**Entidades**	**#Anotações**	**%**
Pessoa	24.844	55,80	advogado	735	1,65
			autor	1.259	2,83
			identificador Policial	2.012	4,52
			juiz	576	1,29
			outros	6.003	13,48
			promotor	363	0,82
			réu	8.773	19,70
			testemunha	2.967	6,66
			vítima	2.156	4,84
Evidência	3.318	7,45	evidência	3.318	7,45
Pena	407	0,91	pena	407	0,91
Endereço	2.065	4,64	autor	132	0,30
			crime	466	1,05
			outros	355	0,80
			réu	693	1,56
			testemunha	295	0,66
			vítima	124	0,28
Sentença	172	0,39	sentença	172	0,39
Norma	13.720	30,81	secundária	5.767	12,95
			jurisprudência	1.823	4,09
			principal	6.130	13,77
Total	**44.526**	**100**	**Total**	**44.526**	**100**

As categorias e exemplos de anotação no CDJUR-BR são organizados conforme diferentes tipos de informação jurídica. A entidade "NORMA" refere-se a componentes normativos como legislação, jurisprudência e doutrina, e inclui as subentidades "NOR-PRINCIPAL", que representa normas que regulam o objeto principal do processo; "NOR-ACESSORIA", voltada a normas de contextualização; e "NOR-JURISPRUDÊNCIA", que abrange decisões representativas da interpretação dominante dos tribunais. A entidade "PESSOAS" abrange as menções completas a pessoas físicas ou jurídicas envolvidas no processo, sendo subdividida em "PES-AUTOR" (autor ou demandante), "PES-REU" (réu ou demandado), "PES-VITIMA" (vítima), "PES-JUIZ" (juiz ou desembargador), "PES-ADVOG" (advogado), "PES-TESTEMUNHA" (testemunha), "PES-AUTORID-

POLICIAL" (autoridade policial), "PES-PROMOTOR-MP" (representante do Ministério Público) e "PES-OUTROS" (demais pessoas mencionadas nos autos).

A entidade "ENDEREÇO" trata da identificação completa de locais relevantes no processo, incluindo logradouro, número, complemento, bairro, cidade, estado e CEP, excluindo endereços presentes em cabeçalhos e rodapés; suas subentidades são "END-AUTOR" (endereço do autor), "END-REU" (endereço do réu), "END-DELITO" (local da ocorrência do delito), "END-TESTEMUNHA" (endereço da testemunha), "END-VÍTIMA" (endereço da vítima) e "END-OUTROS" (outros endereços mencionados). A entidade "SENTENÇA" corresponde ao trecho que expressa a definição sobre o mérito da causa, ou seja, a decisão propriamente dita, sem incluir sua justificativa. A entidade "PENA" refere-se ao trecho do texto que descreve a pena aplicada, também sem considerar sua fundamentação. Por fim, a entidade "PROVA" abrange as menções a diferentes tipos de provas presentes no processo, tais como testemunhais, documentais, periciais, circunstanciais e materiais.

4.2.2 Pré-processamento

O pré-processamento dos dados é uma etapa fundamental em tarefas de NER, especialmente em domínios sensíveis como o jurídico. Um pré-processamento bem executado garante que os dados estejam adequadamente organizados, normalizados e estruturados, o que é essencial para o treinamento eficaz e a performance otimizada dos modelos. A tokenização constitui o primeiro passo essencial no *pipeline* de pré-processamento. Esta fase envolve a segmentação do texto em unidades linguísticas menores, conhecidas como *tokens*, que podem ser palavras ou subpalavras. A escolha específica do tipo de tokenização é fortemente influenciada pelo modelo subjacente a ser utilizado. Por exemplo, modelos baseados em *transformers*, como BERT, XLNet e ELECTRA, não operam diretamente sobre palavras completas, mas sim sobre subpalavras, utilizando algoritmos especializados como WordPiece [69, 70] e SentencePiece [71].

A aplicação desses métodos para o tratamento de palavras fora do vocabulário *out-of-vocabulary* (OOV), que são frequentemente encontradas em textos jurídicos devido à presença de jargões técnicos complexos, termos compostos e neologismos. Por exemplo, uma palavra como "desembargadora" pode ser segmentada em *tokens* menores como *"des"*, *"##embar"*, e *"##gadora"*. Embora este processo aumente a capacidade de generalização do modelo, ele exige uma atenção meticulosa ao alinhamento preciso entre os *tokens* resultantes e seus respectivos rótulos de entidade, uma consideração particularmente importante ao se trabalhar com o formato IOB.

A normalização é outra etapa vital, cujo objetivo principal é reduzir a variabilidade inerente aos dados textuais, facilitando assim o processo de aprendizado dos modelos.

As principais operações de normalização aplicadas neste estudo foram: *Lowercasing*, que consiste em converter todos os caracteres para minúsculas, o que geralmente contribui para a redução da complexidade do vocabulário. Contudo, em modelos *cased* como o BERT-*cased*, essa etapa foi desabilitada, pois o modelo foi projetado para distinguir e se beneficiar da diferença entre letras maiúsculas e minúsculas. Além disso, foi realizada a remoção de pontuação irrelevante (caracteres como "!", "@" são eliminados, uma vez que não contribuem significativamente para a semântica jurídica do texto), a padronização de abreviações (siglas como "TJSP" ou "STJ" foram mantidas, mas normalizadas em sua grafia, evitando variações como "T.J.S.P." para garantir consistência), e a preservação de nomes próprios e expressões jurídicas, termos técnicos e nomes próprios foram mantidos em sua forma original para evitar a perda de seu significado semântico específico e essencial para o domínio jurídico.

A anotação dos dados segue rigorosamente o formato IOB, uma das convenções mais difundidas e eficazes em tarefas de NER. Neste esquema, cada *token* do texto é rotulado de acordo com sua posição em relação a uma entidade: "B" (*Begin*) para o primeiro *token* de uma entidade, "I" (*Inside*) para os *tokens* subsequentes que compõem a mesma entidade, e "O" (*Outside*) para *tokens* que não fazem parte de nenhuma entidade.

Após a tokenização e anotação, os *tokens* e seus rótulos correspondentes são convertidos em índices numéricos, que são as representações de entrada para os modelos. Além disso, aplicam-se máscaras de atenção [43] para diferenciar os *tokens* relevantes do conteúdo textual dos *tokens* adicionados via *padding* (preenchimento), que são inseridos para uniformizar o comprimento das sequências de entrada dos modelos. Por fim, em vez de uma divisão fixa dos dados em conjuntos de treino, validação e teste, adotou-se a estratégia de validação cruzada *k-fold* com $k = 5$. Essa abordagem permite uma avaliação mais robusta do desempenho dos modelos, utilizando diferentes subconjuntos dos dados para treino e teste de forma alternada, ao mesmo tempo em que assegura uma distribuição equilibrada das entidades em cada uma das dobras. Com os dados devidamente pré-processados e no formato adequado, o foco passa para a descrição dos modelos de NER, que serão avaliados neste estudo.

4.2.3 Legal Entity Recognition

Nesta pesquisa, foi realizada uma avaliação abrangente de 15 modelos distintos de NER. Esses modelos foram classificados em três categorias principais para melhor compreensão de suas arquiteturas e capacidades: Abordagens clássicas, baseadas em *transformers* e híbridas. As abordagens clássicas, mas ainda relevantes, para a tarefa de NER. Elas dependem fortemente de engenharia de *features* e são eficazes para capturar relações sequenciais. As abordagens clássicas avaliadas incluem:

- **CRF**: Este é um modelo probabilístico discriminativo que se destaca na modelagem de sequências. Ele aprende a sequência ótima de rótulos ao considerar as dependências entre *tokens* e seus vizinhos, tornando-o particularmente eficaz para tarefas de rotulagem sequencial como o NER [8].
- **BiLSTM**: Uma arquitetura de rede neural recorrente que é capaz de capturar informações contextuais de ambas as direções no texto (passado e futuro). Isso permite que a BiLSTM compreenda melhor o contexto de cada palavra, essencial para a identificação de entidades [72].
- **BiLSTM-CRF**: Este modelo combina os pontos fortes da BiLSTM na representação contextual e na captura de dependências de longo alcance com a capacidade do CRF de modelar as interdependências entre os rótulos de saída. Essa combinação resulta em um desempenho superior em tarefas de rotulagem sequencial [10].
- **CNN**: As redes neurais convolucionais são particularmente eficazes na extração de características locais e padrões morfológicos, como informações de prefixos e sufixos em nível de caractere. Essa capacidade é valiosa para complementar as representações contextuais em tarefas de NER [40].
- **IDCNN**: Esta arquitetura estende as CNNs tradicionais por meio de convoluções dilatadas, que expandem o campo receptivo sem perder resolução, permitindo capturar dependências de maior alcance com custo computacional reduzido, sendo particularmente útil para processar sequências textuais longas [1].

As abordagens baseadas em *transformers* utilizam arquiteturas de rede neural pré-treinadas em grandes volumes de texto, gerando *embeddings* contextuais robustos. Elas são caracterizadas por sua capacidade de processar sequências em paralelo e capturar dependências de longo alcance de forma eficiente. Para os experimentos em português, a variante BERTimbau [12] foi utilizada para os modelos baseados em BERT. As abordagens baseadas *transformers* avaliadas incluem:

- **BERT**: Este é o modelo *transformer* fundamental, pré-treinado com um mecanismo de atenção bidirecional. Serve como uma base poderosa para diversas tarefas de PLN, incluindo NER, devido à sua capacidade de gerar representações contextuais profundas para cada palavra [44].
- **XLNet**: Este modelo *transformer* difere do BERT por sua abordagem autoregressiva com permutação dos *tokens* de entrada. Essa técnica permite que o XLNet capture relações bidirecionais de uma maneira alternativa e mais completa, muitas vezes complementando ou superando o BERT em certas tarefas ao considerar todos os pares de *tokens* em um contexto [46].

- **ELECTRA**: Este modelo se destaca por seu pré-treinamento eficiente por meio de uma tarefa de detecção de tokens substituídos, proporcionando representações de alta qualidade com menor custo computacional em comparação com outros *transformers* [2].

As abordagens híbridas combinam a capacidade de representação contextual das abordagens baseadas em *transformers* com abordagens clássicas, que usam como camadas adicionais de modelagem sequencial ou de dependência de rótulos. Essa combinação visa tirar o melhor proveito de ambas as abordagens, otimizando o desempenho em tarefas complexas como o NER. Para isso, todas as combinações possíveis entre os três *transformers* (BERT, XLNet, ELECTRA) e as cinco arquiteturas clássicas (CRF, BiLSTM, BiLSTM-CRF, CNN, IDCNN) foram exploradas, totalizando quinze modelos híbridos. As abordagens híbridas avaliadas incluem:

- **BERT-CRF**: Esta configuração aprimora o BERT ao adicionar uma camada CRF diretamente sobre suas saídas. O CRF é capaz de capturar as transições e dependências entre os rótulos de forma mais precisa, o que é vital para garantir a coerência das sequências de entidades identificadas [56].
- **BERT-BiLSTM**: Neste modelo, uma camada BiLSTM é inserida entre a saída do BERT e a camada de classificação final. Essa adição visa complementar o BERT na captura de dependências sequenciais complexas, fornecendo uma camada adicional de modelagem temporal.
- **BERT-BiLSTM-CRF**: Esta é a combinação mais completa e robusta, integrando as três camadas: BERT para representação contextual profunda, BiLSTM para modelagem sequencial bidirecional e CRF para garantia da coerência e otimização das transições de rótulos. Esse modelo explora ao máximo o contexto, a sequência e as dependências de rótulos, buscando o desempenho ideal.
- **BERT-CNN**: Nesta configuração, uma camada CNN é adicionada após as representações do BERT para extrair características locais e padrões morfológicos adicionais, complementando a representação contextual com informações de vizinhança imediata.
- **BERT-IDCNN**: Este modelo substitui a CNN convencional pela IDCNN, permitindo capturar dependências de maior alcance com eficiência computacional, expandindo o campo receptivo sem perda de resolução.
- **XLNet-CRF**: O modelo XLNet-CRF utiliza os *embeddings* gerados pelo XLNet como entrada para uma camada CRF. Similar ao BERT-CRF, essa abordagem com-

bina o poder contextual dos *embeddings* do XLNet com a capacidade de modelagem de rótulos sequenciais do CRF, garantindo predições mais consistentes.

- **XLNet-BiLSTM**: Neste modelo, os *embeddings* do XLNet são processados por uma camada BiLSTM. Essa adição pode ajudar a capturar padrões sequenciais adicionais que não são diretamente modelados pelos mecanismos de atenção dos *transformers*, otimizando a compreensão de longas dependências.
- **XLNet-BiLSTM-CRF**: Esta combinação representa um modelo híbrido avançado e complexo. O XLNet fornece *embeddings* contextuais ricos, o BiLSTM modela sequências de forma bidirecional, e a camada CRF garante a coerência e a otimização da sequência de rótulos. Esse modelo é particularmente eficaz em domínios desafiadores como o jurídico, onde a estrutura sintática e semântica pode ser intrincada e exigir uma compreensão aprofundada.
- **XLNet-CNN**: Esta abordagem adiciona uma camada convolucional após o XLNet para capturar características locais e padrões em nível de n-grama, complementando a representação contextual com informações morfológicas e de vizinhança.
- **XLNet-IDCNN**: Ao combinar XLNet com IDCNN, este modelo busca integrar representações contextuais profundas com a capacidade de capturar dependências de longo alcance por meio de convoluções dilatadas, otimizando o processamento de sequências extensas.
- **ELECTRA-CRF**: Este modelo combina as representações contextuais eficientes do ELECTRA com a modelagem de transições de rótulos do CRF, visando um equilíbrio entre qualidade representacional e coerência sequencial.
- **ELECTRA-BiLSTM**: Nesta configuração, uma camada BiLSTM é adicionada após o ELECTRA para reforçar a captura de dependências temporais bidirecionais, complementando as representações do transformer.
- **ELECTRA-BiLSTM-CRF**: Esta é a combinação mais completa envolvendo o ELECTRA, integrando representação contextual eficiente, modelagem sequencial bidirecional e otimização de rótulos, buscando maximizar o desempenho com custo computacional reduzido.
- **ELECTRA-CNN**: A adição de uma camada CNN ao ELECTRA permite a extração de padrões locais e morfológicos, complementando as representações contextuais com informações de vizinhança imediata.
- **ELECTRA-IDCNN**: Este modelo combina a eficiência do ELECTRA com a capacidade da IDCNN de capturar dependências de longo alcance por meio de con-

voluções dilatadas, oferecendo uma alternativa computacionalmente eficiente para processar textos jurídicos extensos.

4.2.4 Avaliação

Para assegurar uma avaliação abrangente e fidedigna do desempenho de cada modelo, foi adotada uma abordagem multifacetada, combinando métricas padrão da área com testes estatísticos rigorosos e técnicas de validação cruzada. As métricas de avaliação empregadas são as métricas padrão para tarefas de rotulagem de sequência: *Precision*, *Recall* e *F1-Score* [73]. Essas métricas foram calculadas no nível da entidade, proporcionando uma análise tanto por entidades, média macro, que oferece uma média não ponderada entre todas as entidades, quanto de forma agregada, média micro, que fornece uma média ponderada considerando o suporte, ou seja, o número de ocorrências de cada entidade. Esta abordagem *dual* é crucial para oferecer uma perspectiva equilibrada sobre o desempenho do modelo, considerando tanto entidades frequentes quanto raras. A biblioteca `seqeval` foi utilizada para o cálculo dessas métricas, pois é especificamente projetada para lidar com os resultados de tarefas de NER no formato IOB.

As métricas são formalmente definidas como:

- *Precision*: É a proporção de instâncias positivas que foram corretamente identificadas pelo modelo dentre todas as instâncias que ele previu como positivas. Matematicamente, é definida como $P = \frac{TP}{TP+FP}$, onde TP (*True Positives*) são as entidades corretamente identificadas e FP (*False Positives*) são as entidades incorretamente identificadas.

- *Recall*: Representa a proporção de instâncias positivas que foram corretamente identificadas pelo modelo dentre todas as instâncias positivas reais existentes no conjunto de dados. É definida como $R = \frac{TP}{TP+FN}$, onde FN (*False Negatives*) são as entidades que o modelo não conseguiu identificar.

- *F1-Score*: É a média harmônica da *Precision* e da *Recall*, oferecendo uma medida única e equilibrada da acurácia do modelo. É particularmente útil quando há um desequilíbrio entre as entidades. A fórmula é $F1 = 2 \times \frac{Precision \times \text{Recall}}{\text{Precision}+\text{Recall}}$.

Para aprimorar a precisão na avaliação do modelo e estabelecer a significância estatística das diferenças de desempenho entre os modelos, foram aplicados testes não-paramétricos. Especificamente, o Teste de Friedman [74] foi empregado para avaliar se existiam diferenças significativas no desempenho geral entre os múltiplos modelos testados. Caso o Teste de Friedman indicasse uma diferença significativa, ou seja, rejeitasse a hipótese nula de que todos os modelos têm desempenho equivalente, o Teste de Nemenyi,

um teste *post hoc*, seria aplicado. O Teste de Nemenyi é utilizado para identificar quais pares específicos de modelos apresentavam diferenças estatisticamente significativas no desempenho. A aplicação desses testes é crucial em situações que envolvem múltiplas comparações, pois ajudam a controlar o aumento da taxa de erro Tipo I, o risco de identificar falsos positivos de significância. Todas as avaliações estatísticas foram realizadas considerando um nível de significância de 5% ($\alpha = 0.05$). Adicionalmente, para garantir a robustez e a generalização dos resultados, utilizou-se a validação cruzada estratificada com 5 *folds*. Esta técnica garante que a distribuição das classes de rótulos seja preservada de forma consistente em cada *fold* (subdivisão do conjunto de dados), o que é essencial para reduzir o viés de desempenho que poderia surgir de divisões de dados arbitrárias ou desequilibradas. A combinação dessas métricas de avaliação, testes estatísticos rigorosos e a validação cruzada fornece uma estrutura de avaliação completa e confiável para o desempenho dos modelos de NER propostos [64].

4.3 Implementação

Os experimentos foram conduzidos integralmente na linguagem Python, fazendo uso extensivo de bibliotecas amplamente reconhecidas e adotadas na comunidade de PLN e *deep learning*. As principais bibliotecas utilizadas incluíram: `transformers` para o acesso e manipulação de modelos pré-treinados e abordagens baseadas em *transformers*; `seqeval` para o cálculo eficiente e preciso das métricas de avaliação de tarefas de rotulagem sequencial, como o NER; `PyTorch` como o *framework* de *deep learning* subjacente para a construção e treinamento dos modelos de rede neural; e `pytorch-crf` para a implementação de CRF e também foi usado na camada de saída das abordagens híbridas, que permitem modelar as dependências entre os rótulos de saída. As configurações de treinamento para as abordagens clássicas, baseadas em *transformers* e híbridas são detalhadas na Tabela 4.4.

A infraestrutura computacional utilizada para a execução dos experimentos consistiu em um sistema baseado em Linux, configurado com *hardware* de alto desempenho. Este sistema estava equipado com duas unidades de processamento gráfico (GPUs) NVIDIA Tesla $V100$ Super de $32GB$ cada, totalizando $64GB$ de memória de vídeo, além de $192GB$ de Memória RAM. O processamento central era realizado por um processador *Intel Xeon Gold* $5220R$ com 48 núcleos lógicos. Para garantir a reprodutibilidade de todos os experimentos e minimizar a variabilidade inerente aos processos de treinamento de modelos de NER, sementes aleatórias fixas foram aplicadas consistentemente em todas as etapas do processo. O código-fonte completo de todos os experimentos, incluindo os *scripts* para pré-processamento de dados, treinamento de modelos e avaliação de desempenho,

Tabela 4.4: Parâmetros de treinamento das abordagens clássicas, baseadas em *transformers* e híbridas

Modelo	**Épocas**	***Batch***	**Otimizador**	**Taxa de Aprendizado**	**BiLSTM Unidades**	***Dropout* BiLSTM**	**CRF Otimização**	**CRF Regularização**
CRF	–	–	–	–	–	–	Integrado	L2 ($\lambda = 0.1$)
BiLSTM	15	–	–	–	256	0.5	–	–
BiLSTM-CRF	15	–	–	–	256	0.5	Conjunta	L2 ($\lambda = 0.1$)
CNN	20	16	AdamW	1×10^{-3}	–	–	–	–
IDCNN	20	16	AdamW	1×10^{-3}	–	–	–	–
BERT	15	16	AdamW	2×10^{-5}–5×10^{-5}	–	–	–	–
XLNet	15	16	AdamW	2×10^{-5}–3×10^{-5}	–	–	–	–
BERT+CRF	15	16	AdamW	2×10^{-5}–5×10^{-5}	–	–	Conjunta	–
BERT+BiLSTM	15	16	AdamW	2×10^{-5}–5×10^{-5}	256	0.3	–	–
BERT+BiLSTM+CRF	15	16	AdamW	2×10^{-5}–5×10^{-5}	256	0.3	Conjunta	–
BERT+CNN	15	16	AdamW	5×10^{-5}	–	–	–	–
BERT+IDCNN	15	16	AdamW	5×10^{-5}	–	–	–	–
XLNet+CRF	15	16	AdamW	2×10^{-5}–3×10^{-5}	–	–	Conjunta	–
XLNet+BiLSTM	15	16	AdamW	2×10^{-5}–3×10^{-5}	256	0.3	–	–
XLNet+BiLSTM+CRF	15	16	AdamW	2×10^{-5}–3×10^{-5}	256	0.3	Conjunta	–
XLNet+CNN	15	16	AdamW	3×10^{-5}	–	–	–	–
XLNet+IDCNN	15	16	AdamW	3×10^{-5}	–	–	–	–
ELECTRA	10	16	AdamW	3×10^{-5}	–	–	–	–
ELECTRA+CRF	10	16	AdamW	3×10^{-5}	–	0.1	Conjunta	–
ELECTRA+BiLSTM	25	16	AdamW	3×10^{-5}	128	0.2	–	–
ELECTRA+BiLSTM+CRF	25	16	AdamW	3×10^{-5}	128	0.2	Sim	–
ELECTRA+CNN	10	16	AdamW	3×10^{-5}	–	0.1	–	–
ELECTRA+IDCNN	10	16	AdamW	3×10^{-5}	–	0.1	–	–

está disponível publicamente no GitHub[1], promovendo a transparência e a capacidade de verificação dos resultados.

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

Capítulo 5

Resultados

Este capítulo apresenta uma análise dos resultados obtidos com os modelos de LER, com foco nos conjuntos de dados LeNER-Br, UlyssesNER e CDJUR-BR. A discussão visa oferecer uma compreensão detalhada do desempenho de cada abordagem e do impacto das arquiteturas híbridas. Para isso, a análise é estruturada em duas partes principais: primeiramente, o desempenho geral dos modelos é comparado, validando as diferenças observadas com testes de significância estatística. Em seguida, realiza-se uma avaliação minuciosa por tipo de entidade para identificar as abordagens mais eficazes para categorias jurídicas específicas. Esta análise criteriosa é necessária para navegar na complexidade dos textos legais e na variabilidade das entidades entre domínios, justificando as escolhas de modelagem e as conclusões alcançadas.

5.1 Resultados Gerais e Análise de Significância Estatística

Os modelos investigados foram categorizados em dois grupos distintos para facilitar a análise e a comparação: (i) abordagens clássicas, que incluem CRF, BiLSTM, BiLSTM-CRF, CNN e IDCNN; e (ii) abordagens baseadas em *transformers* e suas integrações híbridas, abrangendo BERT, XLNet, ELECTRA, bem como suas combinações com camadas de BiLSTM, CRF, CNN e IDCNN. Esta distinção compara a capacidade dos *transformers* de interpretar o contexto de forma global e paralela com a natureza das abordagens clássicas, que modelam as dependências da linguagem de maneira sequencial e direcional, ou por meio da extração de características locais.

A Tabela 5.1 apresenta os valores médios do *F1-Score* obtidos por cada modelo nos três conjuntos de dados, acompanhados dos respectivos desvios padrão, que indicam a consistência do desempenho. Os melhores resultados para cada grupo de modelos, clássicos,

transformers puros e híbridos, estão destacados em *itálico* em suas respectivas categorias, evidenciando as arquiteturas com desempenho superior dentro de cada abordagem e facilitando a comparação entre diferentes paradigmas de modelagem.

Tabela 5.1: Média do F1-Score por modelo e conjunto de dados, com desvio padrão.

Modelo	**LeNER-BR**	**UlyssesNER-BR**	**CDJUR-BR**
CRF	*0.8900 ± 0.0063*	0.6806 ± 0.0159	*0.8460 ± 0.0049*
BiLSTM	0.7340 ± 0.0224	0.6060 ± 0.0049	0.4780 ± 0.0147
BiLSTM-CRF	0.8300 ± 0.0126	0.5820 ± 0.0194	0.7320 ± 0.0075
CNN	0.5672 ± 0.0169	0.8090 ± 0.0088	0.5485 ± 0.0023
IDCNN	0.4053 ± 0.0319	*0.8399 ± 0.0053*	0.7021 ± 0.0049
BERT	*0.9420 ± 0.0075*	*0.8000 ± 0.0261*	*0.8660 ± 0.0049*
XLNet	0.9280 ± 0.0040	0.6520 ± 0.0271	*0.8660 ± 0.0049*
ELECTRA	0.9120 ± 0.0040	0.5080 ± 0.0435	0.8300 ± 0.0000
BERT-CRF	0.9360 ± 0.0049	***0.8020 ± 0.0204***	0.8400 ± 0.0000
BERT-BiLSTM	***0.9440 ± 0.0049***	0.4300 ± 0.2161	***0.8800 ± 0.0000***
BERT-BiLSTM-CRF	***0.9440 ± 0.0049***	0.7680 ± 0.0172	0.8520 ± 0.0040
BERT-CNN	0.9340 ± 0.0049	0.7680 ± 0.0172	0.8388 ± 0.0036
BERT-IDCNN	0.9180 ± 0.0040	0.7680 ± 0.0172	0.8115 ± 0.0052
XLNet-CRF	0.8200 ± 0.0063	0.6620 ± 0.0366	0.8680 ± 0.0040
XLNet-BiLSTM	0.9280 ± 0.0075	0.4120 ± 0.2218	0.8640 ± 0.0049
XLNet-BiLSTM-CRF	0.9280 ± 0.0075	0.5840 ± 0.0625	0.8700 ± 0.0000
XLNet-CNN	0.9320 ± 0.0075	0.6740 ± 0.0294	0.8440 ± 0.0622
XLNet-IDCNN	0.9340 ± 0.0049	0.5260 ± 0.2648	0.8680 ± 0.0040
ELECTRA-CRF	0.9200 ± 0.0089	0.5460 ± 0.0332	0.8340 ± 0.0049
ELECTRA-BiLSTM	0.8960 ± 0.0080	0.6960 ± 0.0273	0.8000 ± 0.0049
ELECTRA-BiLSTM-CRF	0.9080 ± 0.0075	0.6960 ± 0.0287	0.8000 ± 0.0049
ELCTRA-CNN	0.9180 ± 0.0075	0.6200 ± 0.0341	0.8300 ± 0.0000
ELCTRA-IDCNN	0.9140 ± 0.0080	0.4820 ± 0.0668	0.8280 ± 0.0040

A análise dos resultados consolidados na Tabela 5.1 permite identificar padrões de desempenho distintos entre as arquiteturas clássicas e as baseadas em *transformers*, revelando como cada componente adicional influencia a precisão do reconhecimento de entidades mencionadas.

No contexto do conjunto de dados LeNER-BR, a superioridade das abordagens híbridas foi confirmada pelos modelos BERT-BiLSTM e BERT-BiLSTM-CRF, que atingiram o topo do desempenho com um *F1-Score* de 0.9440 ± 0.0049. Essa performance superior evidencia que a capacidade de representação contextual profunda do BERT, quando aliada à memória sequencial da BiLSTM, permite ao modelo capturar tanto a semântica global quanto as dependências locais essenciais para identificar corretamente nomes de tribunais, órgãos judiciais e citações legislativas complexas. É importante notar que o BERT puro com 0.9420 já apresentava desempenho competitivo, mas a adição da camada bidirecional proporcionou o ganho marginal necessário para alcançar a liderança isolada.

A investigação das camadas de extração de características locais, como a CNN e a IDCNN, trouxe novos *insights* sobre a estrutura do LeNER-BR. O modelo BERT-CNN atingiu 0.9340, superando o BERT-IDCNN com 0.9180. Tecnicamente, isso sugere que, para este conjunto de dados, as características morfológicas e contextuais mais relevantes para o *NER* jurídico estão contidas em janelas de texto contíguas e curtas, onde a CNN tradicional opera com eficácia. Por outro lado, a IDCNN, apesar de possuir um campo de visão expandido através de convoluções dilatadas para capturar dependências de maior alcance, não conseguiu traduzir essa capacidade em ganho métrico sobre a CNN comum ou a BiLSTM, indicando uma possível redundância informacional para este cenário específico, onde as informações estão relativamente próximas umas das outras.

No que tange aos modelos baseados em XLNet e ELECTRA no primeiro conjunto de dados, observa-se que, embora competitivos, eles raramente superaram as variantes do BERT. O XLNet puro alcançou 0.9280, enquanto o ELECTRA obteve 0.9120. Um ponto de interesse reside no XLNet-IDCNN com 0.9340 e no XLNet-CNN com 0.9320, que se aproximaram do desempenho do BERT-CNN. Este resultado indica que a arquitetura autorregressiva do XLNet, que modela o contexto através de permutações de palavras, beneficia-se da estrutura de convoluções para estabilizar a identificação de entidades em textos onde a ordem das palavras e as referências cruzadas são densas. Em contrapartida, as variações do ELECTRA mantiveram uma performance sólida acima da marca dos 0.90, com destaque para o ELECTRA-CRF atingindo 0.9200 e o ELETRA-CNN 0.9180, demonstrando que seu pré-treinamento eficiente focado em distinguir *tokens* reais de corrompidos é altamente eficaz para a língua portuguesa, mesmo com um custo computacional reduzido.

A análise do conjunto UlyssesNER-BR revela um cenário de maior instabilidade e dificuldade técnica, refletido em métricas de *F1-Score* consideravelmente menores quando comparadas aos demais *datasets*. O destaque absoluto foi o BERT-CRF, que atingiu 0.8020 ± 0.0204, consolidando a importância da camada de CRF na imposição de restrições estruturais sobre a saída do modelo. Em textos legislativos, onde a estrutura das entidades pode ser altamente padronizada mas o vocabulário é extremamente volátil e diversificado, a capacidade do CRF de modelar a probabilidade conjunta das etiquetas de uma sentença inteira evita erros de sequência, como etiquetas I-ORG sem um B-ORG precedente, que modelos puramente neurais costumam cometer. É notável que, neste conjunto, o ELECTRA-BiLSTM e o ELECTRA-BiLSTM-CRF tenham mantido uma consistência de 0.6960, superando versões equivalentes de modelos baseados em arquiteturas maiores e mais parametrizadas, como o próprio BERT em algumas configurações.

Um fenômeno estatístico foi observado nas combinações XLNet-BiLSTM atigindo o *F1-Score* de 0.4120 ± 0.2218 e BERT-BiLSTM com 0.4300 ± 0.2161 dentro do UlyssesNER-

BR. Ambos os modelos apresentaram desvios padrão alarmantemente altos, chegando a 0.22, o que denota uma falha grave de convergência durante o ajuste fino (*fine-tuning*). Essa instabilidade sugere que o acoplamento de camadas BiLSTM a certos *transformers* em textos parlamentares pode levar ao colapso do gradiente ou a um ajuste excessivo a características ruidosas e esparsas dos dados. Curiosamente, o modelo ELCTRA-IDCNN obteve 0.4820 ± 0.0668 também sofreu com essa queda de desempenho, enquanto o ELECTRA-CRF com 0.5460 ± 0.0332 manteve uma performance superior, reforçando a tese de que decodificadores probabilísticos estruturados como o CRF são mais resilientes a dados desbalanceados ou ruidosos por imporem restrições globais que estabilizam o aprendizado.

No conjunto CDJUR-BR, o panorama de resultados foi marcado por uma estabilidade incomum entre as melhores abordagens. O modelo BERT-BiLSTM alcançou o melhor resultado com 0.8800 ± 0.0000, apresentando um desvio padrão nulo, o que indica uma concordância perfeita entre as diferentes partições de teste e uma robustez exemplar para este domínio. Outro destaque significativo foi a família XLNet, onde o XLNet-CRF e o XLNet-IDCNN atingiram ambos 0.8680, superando ligeiramente o BERT puro, que teve 0.8660, e o XLNet original com 0.8660. Esse comportamento sugere que a arquitetura do XLNet é particularmente apta para lidar com documentos jurídicos padronizados, onde a permutação de contexto durante o treinamento ajuda a resolver ambiguidades terminológicas que modelos de mascaramento simples, como o BERT, podem não capturar com a mesma precisão em domínios específicos com vocabulário repetitivo, mas semanticamente denso.

As abordagens clássicas e os novos modelos baseados em convolução também apresentaram dados dignos de nota no CDJUR-BR. O CRF isolado obteve 0.8460 ± 0.0049, um valor superior ao de modelos modernos como o ELECTRA-CRF com 0.8340 ± 0.0049 e o BERT-IDCNN obteve 0.8115 ± 0.0052. Esse dado é fundamental para a análise, pois demonstra que, em bases de dados com padrões linguísticos rígidos e entidades bem definidas, a modelagem estatística baseada em funções de características manuais pode ser tão eficaz quanto redes neurais profundas, com a vantagem adicional de ser mais interpretável e menos custosa computacionalmente. Por outro lado, o desempenho do ELCTRA-CNN e do ELCTRA-IDCNN na faixa de 0.83, mostra que o ELECTRA atua como uma base robusta para aplicações em ambientes com maior restrição de hardware, mantendo a competitividade com o ELECTRA original com 0.8300 e oferecendo um equilíbrio interessante entre eficiência e precisão.

Em suma, a comparação entre as diversas configurações testadas evidencia que não existe uma arquitetura universalmente superior para todos os conjuntos de dados jurídicos brasileiros. Enquanto as combinações BERT-BiLSTM dominam o cenário de acórdãos

complexos LeNER-BR e documentos padronizados CDJUR-BR, o BERT-CRF mostra-se indispensável para a volatilidade dos textos legislativos UlyssesNER. A inclusão de camadas como CNN e IDCNN provou ser uma alternativa eficiente para a extração de padrões locais, especialmente quando acopladas ao XLNet ou BERT, mas seu desempenho é sensível à natureza do corpus. Esses resultados reforçam a necessidade premente de considerar a natureza intrínseca do *dataset* jurídico ao selecionar a arquitetura do modelo, equilibrando cuidadosamente a profundidade contextual dos *transformers* com a capacidade de estruturação oferecida por camadas probabilísticas como o CRF ou recorrentes como a BiLSTM.

Para uma validação estatística rigorosa dos resultados, foram aplicados os testes de Friedman & Nemenyi. A seleção dos modelos para essa análise incluiu os de melhor desempenho e seus componentes base: BERT-BiLSTM-CRF, BERT e BiLSTM-CRF. A escolha do BiLSTM-CRF como representante da abordagem clássica se justificou por seu desempenho consistentemente superior ao BiLSTM.

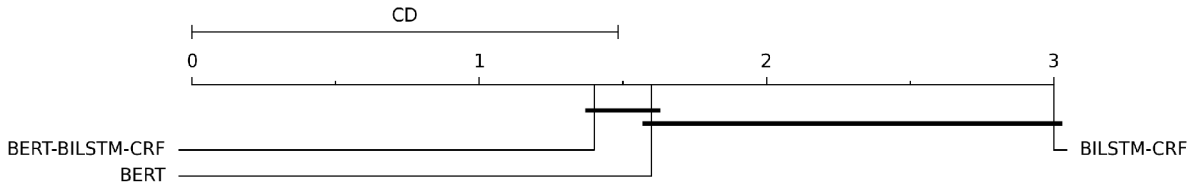

Figura 5.1: Diagrama de Diferença Crítica (Teste de Nemenyi) para o conjunto de dados LeNER-BR.

Conforme o diagrama de Diferença Crítica (CD) na Figura 5.1, o teste de Friedman indicou diferenças estatisticamente significativas entre os modelos ($\chi^2 = 8.40$, $p = 0.0147$), com uma CD de 1.4823. O modelo BERT-BiLSTM-CRF obteve o melhor ranqueamento médio (1.4), seguido de perto pelo BERT (1.6), enquanto o BiLSTM-CRF ficou em terceiro lugar (3.0). A análise *post-hoc* de Nemenyi confirmou que a diferença entre o BERT-BiLSTM-CRF e o BiLSTM-CRF é estatisticamente significativa com $p = 0.0307$.

Isso demonstra que a integração de *embeddings* contextuais poderosos do BERT com a capacidade de modelagem sequencial e de dependência de rótulos do BiLSTM-CRF gera um ganho de desempenho, validando a abordagem híbrida. Por outro lado, a não significância estatística entre BERT-BiLSTM-CRF e BERT sugere que, embora a abordagem híbrida seja marginalmente melhor, o BERT por si só já captura grande parte da complexidade do conjunto de dados. Para uma compreensão mais profunda das capacidades de cada modelo, foi realizada uma análise por tipo de entidade.

Como mostrado na Tabela 5.1, no conjunto de dados UlyssesNER-BR, que se distingue por sua natureza de textos legislativos, a análise dos resultados revela um cenário onde as abordagens híbridas e baseadas em *transformers* continuam a ser dominantes, embora

com algumas particularidades. A combinação BERT-CRF alcançou o melhor desempenho geral, com um *F1-Score* de 0.8020 ± 0.0204. Isso sugere que, para este conjunto de dados, a capacidade do CRF de modelar dependências entre rótulos adjacentes foi um complemento valioso aos *embeddings* contextuais do BERT, especialmente em textos com estruturas de frase menos formais que os documentos jurídicos.

O BERT manteve um desempenho muito competitivo 0.8000 ± 0.0261, confirmando sua robustez independentemente da arquitetura de *tagging* final. O BERT-BiLSTM-CRF também se mostrou forte com 0.7680 ± 0.0172. No entanto, o BERT-BiLSTM apresentou uma queda notável no desempenho, com um *F1-Score* de 0.4300 ± 0.2161 e um desvio padrão muito elevado, o que pode indicar instabilidade do modelo ou *overfitting* para este conjunto de dados específico. Essa inconsistência pode ser atribuída à sensibilidade da camada BiLSTM a padrões muito específicos do treinamento, que não se generalizam bem. Similarmente, os modelos baseados em XLNet continuaram a demonstrar um desempenho inferior, com o XLNet-BiLSTM obtendo o menor *F1-Score* de 0.4120 ± 0.2218 e a maior variabilidade, reforçando a percepção de que o XLNet pode não ser a melhor escolha para o português em domínios específicos sem um ajuste fino extensivo.

Entre as abordagens clássicas, o CRF com 0.6806 ± 0.0159 superou o BiLSTM e o BiLSTM-CRF, novamente indicando a importância das dependências de rótulos. Contudo, mesmo o melhor modelo clássico ficou consideravelmente aquém das abordagens baseadas em *transformers*, sublinhando a superioridade da representação contextual na compreensão semântica. A fim de verificar se a superioridade do BERT-CRF, o modelo de melhor desempenho, é estatisticamente relevante quando comparado aos seus modelos base BERT e CRF, realizou-se uma análise com os testes de Friedman & Nemenyi.

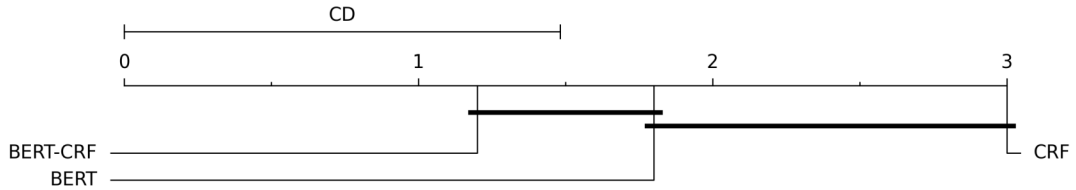

Figura 5.2: Diagrama de Diferença Crítica (Teste de Nemenyi) para o conjunto de dados UlyssesNER-BR.

A Figura 5.2 ilustra os resultados dos testes estatísticos de Friedman & Nemenyi para o conjunto de dados UlyssesNER. O teste de Friedman revelou diferenças significativas entre os modelos de topo ($\chi^2 = 8.40$, $p = 0.0150$), com uma CD de 1.4823. O BERT-CRF obteve o melhor ranqueamento médio (1.02), seguido de perto pelo BERT (1.8), e o CRF em terceiro (3.0). A análise *post-hoc* de Nemenyi confirmou uma diferença estatisticamente significativa entre o BERT-CRF e o CRF com $p = 0.0123$. Este resultado valida a hipótese de que a combinação de *embeddings* contextuais de alta qualidade com uma camada de

predição estruturada como o CRF é superior para tarefas de LER neste conjunto de dados, superando o desempenho de um modelo clássico autônomo.

Conforme indicado na Tabela 5.1, o melhor desempenho no conjunto de dados CDJUR-BR, caracterizado por sentenças criminais, foi alcançado pela abordagem híbrida BERT-BiLSTM, com um *F1-Score* de 0.8800 ± 0.0000. O desvio padrão zero é um indicativo de uma consistência perfeita entre as *folds* de validação, algo extremamente raro e que denota uma robustez excepcional para este modelo específico neste conjunto de dados.

Modelos como BERT e XLNet também apresentaram resultados fortes, ambos com 0.8660 ± 0.0049. É interessante notar que, ao contrário do LeNER-BR e UlyssesNER, a adição de uma camada CRF ao BERT, BERT-CRF com 0.8400 ± 0.0000 e o BERT-BiLSTM-CRF com 0.8520 ± 0.0040 não resultou em melhorias de desempenho em comparação com o BERT-BiLSTM. Isso pode sugerir que, para as entidades do CDJUR-BR, a modelagem sequencial bidirecional do BiLSTM já captura as dependências de rótulos de forma eficaz, e a camada CRF adiciona complexidade desnecessária ou redundância, ou até mesmo restrições que não são ideais para a granularidade das entidades nesse conjunto de dados.

Entre as abordagens clássicas, o CRF se destacou com 0.8460 ± 0.0049, superando o BiLSTM com 0.4780 ± 0.0147 e o BiLSTM-CRF com 0.7320 ± 0.0075. O desempenho notavelmente baixo do BiLSTM neste conjunto de dados com 0.4780 é um ponto a ser investigado, pois sugere que a capacidade do BiLSTM em modelar dependências de longo alcance pode não ser suficiente por si só para entidades complexas ou pouco frequentes, sem o auxílio de *embeddings* pré-treinados ricos em contexto ou de uma camada de predição estruturada como o CRF. Os modelos baseados em XLNet mostraram resultados comparáveis aos do BERT, com as versões híbridas apresentando melhorias modestas: XLNet-CRF com 0.8680 ± 0.0040, XLNet-BiLSTM-CRF com 0.8700 ± 0.0000. Para determinar se o BERT-BiLSTM supera significativamente seus modelos base, BERT e BiLSTM, uma análise comparativa foi conduzida usando os testes estatísticos de Friedman & Nemenyi.

Figura 5.3: Diagrama de Diferença Crítica (Teste de Nemenyi) para o conjunto de dados CDJUR-BR.

A Figura 5.3 exibe o diagrama de CD para o conjunto de dados CDJUR-BR. O teste

de Friedman indicou diferenças significativas entre os modelos avaliados ($\chi^2 = 10.00$, $p = 0.0067$), com uma CD de 1.4823. O BERT-BiLSTM obteve o melhor ranqueamento médio (1.0), seguido pelo BERT (2.0), e o BiLSTM em terceiro (3.0). A análise *post-hoc* de Nemenyi confirmou uma diferença estatisticamente significativa entre o BERT-BiLSTM e o BiLSTM com $p = 0.0045$.

Este resultado destaca de forma contundente a vantagem de combinar os *embeddings* contextuais de alta qualidade do BERT com as camadas recorrentes bidirecionais do BiLSTM, que são eficazes na captura de dependências sequenciais, especialmente em categorias de entidades com maior variabilidade estrutural. No entanto, as diferenças entre BERT-BiLSTM e BERT, bem como entre BERT e BiLSTM, não foram estatisticamente significativas com $p = 0.2538$, o que sugere que o BERT já oferece uma base muito forte, e a adição do BiLSTM, embora aprimore o desempenho geral, não o faz de maneira estatisticamente significativa em todas as comparações diretas para este conjunto de dados em particular, mas sim em relação à abordagem clássica.

5.2 Desempenho por tipo de entidade

Tabela 5.2: Média do *F1-Score* ± desvio padrão por entidade para o conjunto de dados LeNER-BR

Entidade	**BiLSTM-CRF**	**BERT**	**BERT-BiLSTM-CRF**
jurisprudencia	0.7580 ± 0.0504	**0.9100 ± 0.0110**	0.8880 ± 0.0204
legislacao	0.8600 ± 0.0089	**0.9540 ± 0.0120**	0.9420 ± 0.0075
local	0.7080 ± 0.0791	0.9240 ± 0.0326	**0.9280 ± 0.0147**
organizacao	0.8520 ± 0.0147	0.9240 ± 0.0049	**0.9320 ± 0.0040**
pessoa	0.8620 ± 0.0232	**0.9740 ± 0.0102**	0.9720 ± 0.0075
tempo	0.8140 ± 0.0258	0.9740 ± 0.0049	**0.9780 ± 0.0117**

A Tabela 5.2 detalha o *F1-Score* médio e o desvio padrão para cada uma das seis entidades no conjunto de dados LeNER-BR, comparando BiLSTM-CRF, BERT e BERT-BiLSTM-CRF. A análise por entidade reforça a superioridade das abordagens baseadas em *transformers* e híbridas. O BERT demonstrou excelência na identificação de "JURISPRUDENCIA" com *F1-Score* de 0.9100, "LEGISLACAO" com 0.9540 e "PESSOA" com 0.9740. Isso se deve à sua capacidade de processar o texto bidirecionalmente, capturando informações contextuais ricas que são cruciais para entidades que podem variar em estrutura ou aparecer em diversos contextos sintáticos. A alta pontuação em "PESSOA", por exemplo, indica que o BERT é muito eficaz em discernir nomes próprios e pronomes específicos de pessoas no domínio legal.

Por outro lado, o BERT-BiLSTM-CRF superou os demais em "LOCAL" com *F1-Score* de 0.9280, "ORGANIZACAO" com 0.9320 e "TEMPO" com 0.9780. A combinação da robustez contextual do BERT com a capacidade de modelagem sequencial do BiLSTM e a imposição de restrições de transição de rótulos do CRF parece ser particularmente benéfica para entidades que exigem uma compreensão mais profunda de padrões sequenciais e interdependências entre os *tokens*. Entidades como "LOCAL" e "ORGANIZACAO" podem apresentar variações significativas e a camada BiLSTM-CRF auxilia na identificação de limites e na manutenção da coerência dos rótulos. Para "TEMPO", a precisão é crucial, e a combinação de modelos provavelmente ajuda a discernir expressões temporais complexas.

Adicionalmente, a observação de desvios padrão mais baixos nas entidades "PESSOA" com ± 0.0102 e "tempo" com ± 0.0049 para BERT é um achado significativo. Isso indica que, além de um desempenho superior, esses modelos são mais estáveis e confiáveis em suas previsões. Essa consistência é um atributo valioso em aplicações do mundo real, especialmente no setor jurídico, onde a precisão e a minimização de erros são de suma importância para evitar interpretações equivocadas ou decisões inadequadas. Conclui-se que a abordagem híbrida não só oferece um desempenho médio superior, mas também uma maior robustez e confiabilidade na identificação de entidades legais no conjunto de dados LeNER-BR.

Tabela 5.3: Média do *F1-Score* (± desvio padrão) por entidade para o conjunto de dados UlyssesNER

Entidade	**CRF**	**BERT**	**BERT-CRF**
data	0.2320 ± 0.1993	0.6340 ± 0.0948	**0.6520 ± 0.1530**
evento	0.6540 ± 0.1437	**0.6580 ± 0.1485**	0.6540 ± 0.1048
fundamento	0.6540 ± 0.0524	**0.7500 ± 0.0434**	0.7400 ± 0.0385
local	0.6480 ± 0.1026	0.7840 ± 0.0641	**0.8380 ± 0.0371**
organização	0.6980 ± 0.0471	**0.8180 ± 0.0676**	0.8100 ± 0.0566
pessoa	0.6620 ± 0.0299	0.8260 ± 0.0224	**0.8400 ± 0.0400**
produtodelei	0.7380 ± 0.0312	**0.7840 ± 0.0162**	0.7740 ± 0.0206

A Tabela 5.3 apresenta os F1-Scores médios e desvios padrão por entidade, comparando CRF, BERT e BERT-CRF. A análise por entidade para o UlyssesNER revela padrões interessantes. O BERT-CRF demonstrou a maior eficácia na identificação de "DATA" com 0.6520 ± 0.1530, "LOCAL" com 0.8380 ± 0.0371 e "PESSOA" com 0.8400 ± 0.0400. A combinação da capacidade de contextualização do BERT com a modelagem de sequências do CRF é especialmente vantajosa para entidades que dependem de um forte contexto sequencial, como datas que podem ter múltiplos formatos, e para localizar e pessoas que podem ser ambíguas fora de seu contexto. Para "DATA", por exemplo, o CRF pode ajudar

a garantir que todos os componentes de uma data (dia, mês, ano) sejam corretamente identificados como parte da mesma entidade.

O BERT, por sua vez, destacou-se no reconhecimento de entidades semanticamente ricas, como "FUNDAMENTO" com 0.7500 ± 0.0434, "ORGANIZACAO" com 0.8180 ± 0.0676 e "PRODUTODELEI" com 0.7840 ± 0.0162. Essas entidades podem ser mais sensíveis à compreensão semântica profunda fornecida pelos *embeddings* do BERT, que podem diferenciar termos similares em diferentes contextos. A categoria "EVENTO" mostrou resultados muito próximos entre os modelos, sugerindo que sua identificação pode ser menos dependente de arquiteturas complexas ou que sua definição no conjunto de dados é mais direta.

A observação de desvios padrão mais baixos nas abordagens híbridas, como o BERT-CRF, indica uma maior estabilidade e confiabilidade nas previsões. Essa consistência é um atributo valorizado, não apenas no domínio legal, mas também na análise de textos literários, onde a interpretação precisa de entidades pode ser crucial para estudos acadêmicos e para a construção de sistemas de recuperação de informação mais robustos. Em suma, as descobertas no conjunto de dados UlyssesNER reforçam as vantagens das abordagens baseadas em *transformers*, especialmente quando complementadas por camadas de predição estruturada como o CRF, para o LER.

Tabela 5.4: Média do F1-Score (± desvio padrão) por Entidade para o conjunto de dados CDJUR-BR

Categoria	**Entidade**	**BiLSTM**	**BERT**	**BERT-BiLSTM**
Pessoa	advogado	0.0600 ± 0.0167	0.5880 ± 0.0286	**0.6280 ± 0.0255**
	autor	0.3740 ± 0.0422	0.7760 ± 0.0398	**0.7880 ± 0.0237**
	identificador-policial	0.6960 ± 0.0361	0.9200 ± 0.0063	**0.9300 ± 0.0075**
	juiz	0.1820 ± 0.0445	0.8020 ± 0.0279	**0.8260 ± 0.0248**
	outro	0.4100 ± 0.0253	0.8220 ± 0.0075	**0.8320 ± 0.0089**
	promotor	0.1460 ± 0.0484	0.6880 ± 0.0279	**0.7200 ± 0.0193**
	réu	0.6380 ± 0.0564	0.9200 ± 0.0126	**0.9240 ± 0.0089**
	testemunha	0.2260 ± 0.0689	0.8080 ± 0.0117	**0.8200 ± 0.0137**
	vítima	0.5520 ± 0.0421	0.9000 ± 0.0219	**0.9060 ± 0.0150**
Evidência	evidência	0.1740 ± 0.0287	0.6460 ± 0.0273	**0.6720 ± 0.0237**
Pena	pena	0.4100 ± 0.0938	0.8540 ± 0.0969	**0.8680 ± 0.0545**
Endereço	autor	0.0160 ± 0.0185	0.4880 ± 0.0880	**0.5560 ± 0.0699**
	crime	0.1900 ± 0.0482	0.7900 ± 0.0460	**0.8120 ± 0.0117**
	outro	0.1620 ± 0.0492	0.7300 ± 0.0518	**0.7360 ± 0.0253**
	réu	0.0900 ± 0.0690	0.7260 ± 0.0215	**0.7520 ± 0.0237**
	testemunha	0.0760 ± 0.0162	0.6680 ± 0.0643	**0.6840 ± 0.0591**
	vítima	0.0000 ± 0.0000	0.5020 ± 0.0778	**0.5700 ± 0.0643**
Sentença	sentença	0.0100 ± 0.0126	0.5780 ± 0.1089	**0.6200 ± 0.0703**
Norma	secundária	0.6700 ± 0.0167	**0.8980 ± 0.0098**	0.8940 ± 0.0117
	jurisprudência	0.8040 ± 0.0441	**0.9820 ± 0.0040**	0.9800 ± 0.0036
	principal	0.6820 ± 0.0440	**0.9020 ± 0.0040**	0.9000 ± 0.0024

A Tabela 5.4 apresenta o *F1-Score* médio e o desvio padrão por entidade, comparando BiLSTM, BERT e BERT-BiLSTM no conjunto de dados CDJUR-BR. A análise por entidade no conjunto de dados CDJUR-BR confirma de forma contundente a superioridade das abordagens híbridas e baseadas em *transformers*. O modelo BiLSTM demonstrou um desempenho alarmantemente baixo em várias categorias críticas, como "endereço-vítima" com 0.0000 ± 0.0000, "SENTENÇA" com 0.0100 ± 0.0126 e "ENDEREÇO-AUTOR" com 0.0160 ± 0.0185.

Esse resultado sugere que, sem a riqueza contextual dos *embeddings* de *transformers*, o BiLSTM tem dificuldade em lidar com a complexidade e a variabilidade das entidades no domínio jurídico, especialmente quando se trata de identificar categorias com ocorrências esparsas ou contextos muito específicos, como endereços associados a diferentes papéis processuais. Em contraste, o BERT já proporcionou melhorias massivas, superando 0.9000 de *F1-Score* em categorias como "PESSOA-IDENTIFICADOR-POLICIAL", "PESSOA-RÉU", "NORMA-JURISPRUDÊNCIA" e "NORMA-PRINCIPAL". Isso ressalta a capacidade do BERT de capturar a semântica e o contexto necessário para identificar entidades mesmo sem camadas adicionais de modelagem sequencial.

No entanto, o BERT-BiLSTM conseguiu os melhores resultados em grande parte das entidades, com ganhos notáveis em categorias particularmente desafiadoras. Por exemplo, em "EVIDÊNCIA" com *F1-Score* de 0.6720 ± 0.023 e "PENA" com 0.8680 ± 0.0545, a combinação híbrida se destacou, indicando que a camada BiLSTM adiciona uma capacidade de modelagem de dependências sequenciais que é crucial para essas entidades, que muitas vezes são expressas por frases mais longas ou estruturas complexas. As categorias de "ENDEREÇO", como "ENDEREÇO-CRIME" com 0.8120 ± 0.0117, "ENDEREÇO-TESTEMUNHA" com 0.6840 ± 0.0591 e "ENDEREÇO-VÍTIMA" com 0.5700 ± 0.0643, também viram melhorias significativas com o BERT-BiLSTM. A detecção de endereços no contexto jurídico pode ser complexa devido à sua variabilidade estrutural e à necessidade de associá-los corretamente aos papéis processuais, e a camada BiLSTM parece auxiliar na captura desses padrões.

Para entidades relacionadas a pessoas, como "PESSOA-AUTOR", "PESSOA-JUIZ", "PESSOA-OUTRO", "PESSOA-TESTEMUNHA" e "PESSOA-PROMOTOR", o BERT-BiLSTM também mostrou melhorias consistentes. Isso reflete uma modelagem mais apurada das relações sequenciais e contextuais que definem esses tipos de entidades. Além do desempenho superior, a menor variância observada em modelos como BERT e BERT-BiLSTM, "NORMA-JURISPRUDÊNCIA" com ± 0.0040 e ± 0.0036, respectivamente, é um indicativo de maior estabilidade e confiabilidade. Em um domínio tão sensível quanto o jurídico, onde a precisão é fundamental e erros podem ter consequências graves, a consistência nas previsões é um benefício substancial. Em conclusão, os resultados do

CDJUR-BR demonstram que as abordagens híbridas como o BERT-BiLSTM não apenas superam as abordagens clássicas de forma expressiva, mas também aprimoram o desempenho em comparação com abordagens puramente baseadas em *transformers* em categorias de entidades com variabilidade estrutural significativa ou papéis semânticos que requerem uma compreensão mais aprofundada da sequência.

Em uma visão geral, os resultados obtidos nos três conjuntos de dados reforçam a superioridade das abordagens baseadas híbridas para tarefas de LER. As abordagens híbridas, que combinam a capacidade de contextualização profunda dos *transformers* com camadas de modelagem sequencial BiLSTM e/ou de dependências de rótulos CRF, frequentemente resultaram em um desempenho ainda melhor, validado por testes estatísticos que demonstraram ganhos significativos em relação às abordagens clássicas.

Enquanto o CRF se mostrou uma abordagem clássica sólida em certos conjuntos de dados, sua performance foi consistentemente superada pelos modelos mais avançados, que se beneficiam da aprendizagem de representações densas e contextuais. Os modelos baseados em XLNet apresentaram um desempenho mais variável em comparação com o BERT, sugerindo que, para os conjuntos de dados de língua portuguesa e o domínio jurídico utilizados, o XLNet pode exigir um ajuste fino mais intensivo ou que o BERT possua características de pré-treinamento mais adequadas para as nuances do português legal. Essa variabilidade destaca a importância de testar diferentes arquiteturas de *transformers* e suas configurações para cada domínio e língua específica.

É reconhecer que, apesar de suas vantagens de desempenho, as abordagens baseadas em *transformers*, em particular as híbridas, impõem demandas computacionais significativamente maiores. Isso se traduz em tempos de treinamento mais longos, já que modelos com milhões ou bilhões de parâmetros exigem consideravelmente mais tempo para serem treinados e ajustados, mesmo com otimizações. Além disso, há a necessidade de hardware especializado, pois a execução desses modelos, tanto em treinamento quanto em inferência, frequentemente requer GPUs de alto desempenho e grande quantidade de memória RAM e VRAM, o que limita sua aplicabilidade em ambientes com infraestrutura restrita ou com custos computacionais controlados. Soma-se a isso a complexidade de implantação, já que o tamanho e a sofisticação desses modelos podem dificultar sua adoção em sistemas de produção que exigem baixa latência ou operam com recursos limitados.

Esses fatores são considerações práticas importantes que devem ser avaliadas ao escolher um modelo para aplicações em LER. Embora o desempenho seja um objetivo primário, a viabilidade computacional e econômica é igualmente relevante. Para cenários onde os recursos são limitados, a escolha pode recair sobre abordagens baseadas em *transformers* menores ou abordagens clássicas robustas, sacrificando uma fração do desempenho em prol da eficiência. No entanto, para aplicações onde a precisão máxima é mandatória,

como na automação de processos jurídicos críticos, o investimento em recursos para as abordagens híbridas de ponta é justificado pelos ganhos significativos e confiabilidade.

Capítulo 6

Conclusão

Neste trabalho, apresentamos uma análise comparativa aprofundada e inovadora para a tarefa de LER em português, avaliando de forma conjunta uma gama diversificada de modelos, desde abordagens clássicas até abordagens baseadas em *transformers* e híbridas, em três conjuntos de dados jurídicos heterogêneos: LeNER-BR (acórdãos), UlyssesNER-BR (textos legislativos) e CDJUR-BR (sentenças criminais). Diferentemente de estudos anteriores, que geralmente carecem de comparações abrangentes entre múltiplos modelos e *datasets*, nossa avaliação oferece uma visão detalhada da eficácia de cada abordagem em distintos contextos legais, contribuindo para o avanço do NER no domínio jurídico brasileiro.

Os resultados consolidados demonstram que não existe uma arquitetura universalmente superior; a escolha do modelo deve considerar a natureza intrínseca de cada *dataset*. As abordagens híbridas, especialmente aquelas que combinam BERT com camadas BiLSTM e/ou CRF, destacaram-se de forma consistente. No LeNER-BR, o BERT-BiLSTM e o BERT-BiLSTM-CRF alcançaram os melhores *F1-Scores* com 0,9440, evidenciando que a união da representação contextual profunda do BERT com a modelagem sequencial da BiLSTM é crucial para capturar as dependências locais em textos jurídicos complexos. No UlyssesNER-BR, o BERT-CRF obteve o melhor desempenho com 0,8020, indicando que a imposição de restrições globais de rótulos pelo CRF é essencial para lidar com a volatilidade vocabular e a estrutura padronizada dos textos legislativos. Já no CDJUR-BR, o BERT-BiLSTM liderou com 0,8800, mostrando que, para entidades com papéis processuais bem definidos, a modelagem sequencial bidirecional é suficiente, tornando a camada CRF redundante. A análise estatística por meio dos testes de Friedman e Nemenyi confirmou a significância das melhorias obtidas pelas abordagens híbridas em relação às clássicas, com valores de p inferiores a 0,05 nas comparações entre os modelos híbridos de topo e suas contrapartes clássicas.

A avaliação por tipo de entidade revelou padrões complementares. Enquanto o BERT

puro se destacou em entidades semanticamente ricas (como “jurisprudência”, “fundamento” e normas), as configurações híbridas (BERT-BiLSTM-CRF ou BERT-CRF) foram superiores para categorias que exigem maior sensibilidade a limites e dependências sequenciais (como “local”, “organização”, “tempo” e endereços). Essa complementaridade reforça a importância de combinar diferentes paradigmas para lidar com a heterogeneidade das entidades jurídicas.

Para o setor público, considerando a necessidade de equilibrar desempenho e custo computacional em aplicações práticas, a recomendação varia conforme o contexto de uso. Para tribunais e órgãos que já dispõem de infraestrutura computacional com GPUs e precisam da máxima acurácia em tarefas críticas, como a classificação automática de processos ou a indexação de jurisprudência, o modelo BERT-BiLSTM surge como a escolha mais adequada. Ele oferece o melhor custo-benefício entre precisão e estabilidade, com *F1-Scores* superiores a 0,88 nos três conjuntos de dados e desvios padrão baixos, indicando confiabilidade nas predições. Para cenários com restrições orçamentárias ou de hardware, onde o uso de GPUs não é viável, o modelo CRF tradicional apresenta-se como uma alternativa surpreendentemente competitiva, especialmente para documentos com estrutura rígida, como sentenças criminais, onde alcançou 0,8460 de *F1-Score* no CDJUR-BR. Já para aplicações que exigem processamento em tempo real ou em dispositivos com recursos limitados, versões mais leves como o ELECTRA ou mesmo o BERT base podem ser consideradas, uma vez que oferecem um equilíbrio razoável entre desempenho (acima de 0,83–0,86) e eficiência computacional, podendo ser executadas em CPUs com otimizações adequadas.

6.1 Contribuições

As principais contribuições deste trabalho são: a primeira avaliação comparativa abrangente de modelos clássicos, baseados em *transformers* e híbridos para NER em português brasileiro no domínio jurídico, englobando todos os conjuntos de dados públicos disponíveis para o português jurídico, LeNER-BR, CDJUR-BR e UlyssesNER-BR; a validação estatística rigorosa das diferenças de desempenho, utilizando testes de Friedman e Nemenyi, o que confere robustez às conclusões sobre a superioridade das arquiteturas híbridas; a análise granular por tipo de entidade, que identifica quais categorias se beneficiam mais de cada abordagem, fornecendo *insights* práticos para o desenvolvimento de sistemas especializados; a disponibilização de um *benchmark* para futuras pesquisas em NER jurídica em português, facilitando a comparação com novos modelos e técnicas; a discussão sobre a viabilidade computacional das diferentes arquiteturas, destacando o equilíbrio necessário entre precisão e recursos em aplicações reais; e a publicação dos resultados no

XXIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2025) [75], onde o artigo intitulado "Evaluation of a Hybrid Approach to Legal Entity Recognition" demonstrou que as abordagens híbridas superam consistentemente tanto as abordagens clássicas quanto as baseadas exclusivamente em *transformers* na extração de entidades jurídicas, consolidando os achados desta pesquisa e disseminando-os para a comunidade científica.

6.2 Limitações e Trabalhos Futuros

Apesar dos avanços alcançados e da robustez dos resultados obtidos, este estudo apresenta limitações que devem ser consideradas e que, simultaneamente, abrem direções promissoras para investigações futuras. Todos os *transformers* avaliados, BERT, XLNet e ELECTRA, partiram de versões pré-treinadas em *datasets* multilingues ou especificamente em português, como o *BERTimbau*, mas nenhum foi treinado do zero exclusivamente com textos jurídicos brasileiros. Essa abordagem, embora prática e eficiente, pode limitar a captura de particularidades lexicais, sintáticas e semânticas do domínio jurídico, que possui vocabulário especializado, construções frasais recorrentes e entidades com comportamentos específicos que um modelo genérico pode não representar de forma ideal. Adicionalmente, as abordagens híbridas, embora precisas, exigem hardware especializado e longos tempos de treinamento, o que pode inviabilizar sua adoção em cenários com infraestrutura restrita ou necessidade de baixa latência em produção. Observou-se também instabilidade em algumas combinações, como XLNet-BiLSTM e BERT-BiLSTM no UlyssesNER, que apresentaram alta variabilidade, indicando sensibilidade a hiperparâmetros ou à natureza dos dados. Por fim, os resultados são específicos para o domínio jurídico, e a eficácia das arquiteturas pode variar significativamente em outras áreas, demandando novos experimentos para cada contexto.

Diante dessas limitações, trabalhos futuros podem explorar o treinamento de modelos do zero exclusivamente em *datasets* jurídicos brasileiros de grande escala, como acórdãos, leis, petições e doutrinas, gerando representações mais alinhadas com o vocabulário e as construções do domínio e potencialmente elevando o desempenho em NER e tarefas correlatas. A comparação direta entre modelos treinados do zero e versões pré-treinadas gerais constituiria uma contribuição significativa para a área. Por fim, a integração com conhecimento externo, como ontologias e vocabulários controlados do direito, pode melhorar o reconhecimento de entidades ambíguas ou raras com forte dependência de conhecimento de domínio. Essas direções, aliadas aos resultados consolidados neste trabalho e no artigo publicado no ENIAC 2025 [75], contribuirão para o amadurecimento da área de PLN jurídico em português, aproximando a pesquisa acadêmica das necessidades práticas do

setor e fomentando o desenvolvimento de ferramentas mais precisas, eficientes e acessíveis para a análise e gestão de informações legais.

Referências

[1] Strubell, Emma, Patrick Verga, David Belanger e Andrew McCallum: *Fast and accurate entity recognition with iterated dilated convolutions*, 2017. ix, 2, 3, 13, 14, 15, 31, 39

[2] Clark, Kevin, Minh-Thang Luong, Quoc V. Le e Christopher D. Manning: *ELECTRA: pre-training text encoders as discriminators rather than generators*. arXiv, 2020. ix, 2, 18, 19, 31, 40

[3] Siqueira, Dirceu Pereira, Frederico Mendes Junior e Marcel Ferreira Dos Santos: *Poder judiciário na era digital: o impacto das novas tecnologias de informação e de comunicação no exercício da jurisdição*. Consinter de Direito, 2023. 1

[4] Maynard, Diana, Kalina Bontcheva e Isabelle Augenstein: *Natural Language Processing for the Semantic Web*. Springer Cham, 2016. 1, 2

[5] Yadav, Vikas e Steven Bethard: *A survey on recent advances in named entity recognition from deep learning models*. arXiv, 2019. 1

[6] Li, Jing, Aixin Sun, Jianglei Han e Chenliang Li: *A Survey on Deep Learning for Named Entity Recognition*. arXiv, 2020. 1, 2, 7

[7] Luz De Araujo, Pedro Henrique, Teófilo E. De Campos, Renato R. R. De Oliveira, Matheus Stauffer, Samuel Couto e Paulo Bermejo: *LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text*. Em *Computational Processing of the Portuguese Language*. Springer International Publishing, 2018. 1, 2, 3, 4, 8, 23, 27, 31, 33

[8] Lafferty, J., A. McCallum e Fernando Pereira: *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Em *Proceedings of the 18th International Conference on Machine Learning*, 2001. 2, 9, 39

[9] Hochreiter, Sepp e Jürgen Schmidhuber: *Long Short-Term Memory*. Neural Computation, 1997. 2, 11, 12

[10] Huang, Zhiheng, Wei Xu e Kai Yu: *Bidirectional lstm-crf models for sequence tagging*. arXiv, 2015. 2, 11, 12, 13, 15, 39

[11] Abreu, Sandra Collovini de, Tiago Luis Bonamigo e Renata Vieira: *A review on Relation Extraction with an eye on Portuguese*. Brazilian Computer Society, 2013. 2, 3

[12] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*. Em *Intelligent Systems*. Springer International Publishing, 2020. 2, 3, 20, 25, 39

[13] Silveira, Raquel, Caio Ponte, Vitor Almeida, Vládia Pinheiro e Vasco Furtado: *LegalBert-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain*. Em *Intelligent Systems*. Springer Nature Switzerland, 2023. 2, 3, 22, 25, 27

[14] You, Yang, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer e Cho Jui Hsieh: *Large batch optimization for deep learning: Training bert in 76 minutes*, 2019. 2

[15] Brito, Maurício, Vladia Pinheiro, Vasco Furtado, João Araújo Monteiro Neto, Francisco das Chagas Jucá Bomfim, André Câmara Ferreira da Costa, Raquel Silveira e Nilsiton Aragão: *CDJUR-BR – A Golden Collection of Legal Document from Brazilian Justice with Fine-Grained Named Entities*. arXiv, 2023. 2, 3, 4, 25, 27, 31, 32, 33, 35

[16] Albuquerque, Hidelberg O., Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia F. F. Da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, Felipe Siqueira, João P. Tarrega, Joao V. Beinotti, Marcio Dias, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. De Carvalho e Adriano L. I. Oliveira: *UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition*. Em *Computational Processing of the Portuguese Language*. Springer, 2022. 3, 4, 9, 24, 27, 31, 33, 34

[17] Da Silva, F. X. B., G. M. C. Guimarães, R. M. Marcacini, A. L. Queiroz, V. R. P. Borges, T. P. Faleiros e L. P. F. Garcia: *Named Entity Recognition Approaches Applied to Legal Document Segmentation*. Em *X Symposium on Knowledge Discovery, Mining and Learning*. Sociedade Brasileira de Computação - SBC, 2022. 3, 9

[18] Andressa Vieira E Silva: *Uma revisão para o Reconhecimento de Entidades Nomeadas aplicado à língua portuguesa*. Linguamática, 2023. 3

[19] Costa, Rosimeire, Hidelberg Oliveira Albuquerque, Gabriel Silvestre, Nádia Félix F. Silva, Ellen Souza, Douglas Vitório, Augusto Nunes, Felipe Siqueira, João Pedro Tarrega, João Vitor Beinotti, Márcio de Souza Dias, Fabíola S. F. Pereira, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho e Adriano L. I. Oliveira: *Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-Generated Text*. Em *Progress in Artificial Intelligence*. Springer, 2022. 3, 4, 9, 24, 27

[20] Seyler, Dominic, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart e Gerhard Weikum: *A study of the importance of external knowledge in the named entity recognition task*. Em *56th Annual Meeting of the Association for Computational Linguistics*, 2018. 6

[21] Gorinski, Philip John, Honghan Wu, Claire Grover, Richard Tobin, Conn Talbot, Heather Whalley, Cathie Sudlow, William Whiteley e Beatrice Alex: *Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches*. arXiv, 2019. 6

[22] Grishman, Ralph e Beth Sundheim: *Design of the MUC-6 Evaluation*. Em *Proceedings of a Workshop held at Vienna*. Association for Computational Linguistics, 1996. 6

[23] Ramshaw, Lance e Mitch Marcus: *Text chunking using transformation-based learning*. arXiv, 1995. 7

[24] Ratinov, Lev e Dan Roth: *Design challenges and misconceptions in named entity recognition*. Em *Computational Natural Language Learning*, 2009. 7

[25] Hu, Zhentao, Wei Hou e Xianxing Liu: *Deep learning for named entity recognition: a survey*. Neural Comput. Appl., 2024. 8

[26] Nadeau, David e Satoshi Sekine: *A survey of named entity recognition and classification*. Lingvisticae Investigationes, 2007. 8

[27] Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato e Juan Miguel Gómez-Berbís: *Named Entity Recognition: Fallacies, challenges and opportunities*. Computer Standards & Interfaces, 2013. 8

[28] Samy, Doaa: *Reconocimiento y clasificación de entidades nombradas en textos legalesen español*. Procesamiento del Lenguaje Natural, 2021. 8

[29] Coelho, Gustavo, Alimed Celecia, Jefferson de Sousa, Melissa Lemos, Maria Lima, Ana Mangeth, Isabella Frajhof e Marco Casanova: *Information extraction in the legal domain: Traditional supervised learning vs. chatgpt*. Em *Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS*. SciTePress, 2024. 8, 31

[30] Keraghel, Imed, Stanislas Morbieu e Mohamed Nadif: *Recent advances in named entity recognition: A comprehensive survey and comparative study*, 2024. 9, 31

[31] Freitag, Dayne e Andrew McCallum: *Information extraction with hmms and shrinkage*. Association for the Advancement of Artificial Intelligence, 1999. 9

[32] Veneroso, João Mateus de Freitas: *Reconhecimento de Entidades Nomeadas na Web*, 2019. 9

[33] Freitag, Dayne e Andrew McCallum: *Information Extraction with HMM Structures Learned by Stochastic Optimization*. Em *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000. 9

[34] McCallum, Andrew, Dayne Freitag e Fernando Pereira: *Maximum Entropy Markov Models for Information Extraction and Segmentation*. Em *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000. 9

[35] Amaral, Daniela O. F. do e Renata Vieira: *Named Entity Recognition with Conditional Random Fields for the Portuguese Language*. Em *Brazilian Symposium in Information and Human Language Technology*, 2013. 9

[36] Au, Ting Wai Terence, Vasileios Lampos e Ingemar Cox: *E-NER — An Annotated Named Entity Recognition Corpus of Legal Text*. Em *Proceedings of the Natural Legal Language Processing Workshop*. Association for Computational Linguistics, 2022. 9

[37] McCallum, Andrew e Wei Li: *Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons*. Em *Conference on Natural Language Learning*, 2003. 9

[38] Guimarães, Gabriel M.C., Felipe X.B. Da Silva, Andrei L. Queiroz, Ricardo M. Marcacini, Thiago P. Faleiros, Vinicius R.P. Borges e Luís P.F. Garcia: *DODFMiner: An automated tool for Named Entity Recognition from Official Gazettes*. Neurocomputing, 2024. 9

[39] Collobert, Ronan, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu e Pavel Kuksa: *Natural Language Processing (Almost) from Scratch*. arXiv, 2011. 10

[40] Krichen, Moez: *Convolutional neural networks: A survey*. Computers, 12, 2023. 10, 39

[41] Chen, Sheng, Weijun Pan, Yidi Wang, Shenhao Chen e Xuan Wang: *Research on the Method of Air Traffic Control Instruction Keyword Extraction Based on the Roberta-Attention-BiLSTM-CRF Model*. Aerospace, 2025. 13, 14

[42] Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer: *Deep Contextualized Word Representations*. arXiv, 2018. 15

[43] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser e Illia Polosukhin: *Attention Is All You Need*. arXiv, 2017. 15, 16, 38

[44] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018. 15, 16, 17, 18, 39

[45] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer e Veselin Stoyanov: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019. 17, 18

[46] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov e Quoc V. Le: *Xlnet: Generalized autoregressive pretraining for language understanding*. arXiv, 2019. 18, 19, 20, 39

[47] Dai, Zihang, Zhilin Yang, Yiming Yang, William W. Cohen, Jaime Carbonell, Quoc V. Le e Ruslan Salakhutdinov: *Transformer-xl: Attentive language models beyond a fixed-length context*. arXiv, 2019. 19

[48] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma e Radu Soricut: *ALBERT: A lite BERT for self-supervised learning of language representations*. arXiv, 2020. 20

[49] Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso e Tomás Osório: *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, 2023. 20

[50] Ortiz Su'arez, Pedro Javier, Benoit Sagot e Laurent Romary: *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures*. Em *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut f"ur Deutsche Sprache, 2019. 20

[51] Ortiz Su'arez, Pedro Javier, Laurent Romary e Benoit Sagot: *A monolingual approach to contextualized word embeddings for mid-resource languages*. Em *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. 20

[52] Hajlaoui, Najeh, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger e Daniel Varga: *DCEP -digital corpus of the European parliament*. Em *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 2014. 20

[53] Garcia, Eduardo, Nádia Felix Felipe da Silva, Juliana Resplande Sant'Anna Gomes, Hidelberg Oliveira Albuquerque, Ellen Polliana Ramos Souza, Felipe Alves Siqueira, Eliomar Araujo de Lima e Andre Carlos Ponce de Leon Ferreira de Carvalho: *Robertalexpt: um modelo roberta jurídico pré-treinado com deduplicação para língua portuguesa*. LinguaMÁTICA, 2024. 22, 26, 27

[54] Polo, Felipe Maia, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J. Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia e Renato Vicente: *Legalnlp – natural language processing methods for the brazilian legal language*, 2021. 23, 24, 27

[55] Nunes, Rafael Oleques, Dennis Giovani Balreira, André Suslik Spritzer e Carla Maria Dal Sasso Freitas: *A Named Entity Recognition Approach for Portuguese Legislative Texts Using Self-Learning*. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Association for Computational Lingustics, 2024. 26, 27, 31, 32

[56] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *Portuguese named entity recognition using bert-crf*. arXiv, 2020. 29, 40

[57] Wang, Zhili, Yufan Wu, Pengbin Lei e Cheng Peng: *Named entity recognition method of brazilian legal text based on pre-training model*. Journal of Physics: Conference Series, 2020. 30

[58] Albuquerque, Hidelberg O., Ellen Souza, Carlos Gomes, Matheus Henrique de C. Pinto, Ricardo P. S. Filho, Rosimeire Costa, Vinícius Teixeira de M. Lopes, Nádia F. F. da Silva, André C. P. L. F. de Carvalho e Adriano L. I. Oliveira: *Named Entity Recognition: a Survey for the Portuguese Language*. Procesamiento del Lenguaje Natural, 2023. 30

[59] Razeira, R.M. e I.A. Rodello: *A LSTM Recurrent Neural Network Implementation for Classifying Entities on Brazilian Legal Documents*. Em *Computational Science and Its Applications*. Springer International Publishing, 2021. 30

[60] Castro, P.V. Quinta de, N. Félix Felipe da Silva e A. da Silva Soares: *Portuguese Named Entity Recognition Using LSTM-CRF*. Em *Computational Processing of the Portuguese Language*. Springer International Publishing, 2018. 30

[61] Trias, Fernando, Hongming Wang, Sylvain Jaume e Stratos Idreos: *Named entity recognition in historic legal text: A transformer and state machine ensemble method*. Em *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, 2021. 31

[62] Oliveira, Vitor Vasconcelos de, Gabriel da Silva Corvino Nogueira, Thiago de Paulo Faleiros e Ricardo Marcondes Marcacini: *Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents*. Artificial Intelligence and Law, 2024. 31

[63] Guimarães, Gabriel M. C., Felipe X. B. da Silva, Lucas A. B. Macedo, Victor H. F. Lisboa, Ricardo M. Marcacini, Andrei L. Queiroz, Vinicius R. P. Borges, Thiago P. Faleiros e Luis P. F. Garcia: *Legal Document Segmentation and Labeling Through Named Entity Recognition Approaches*. Journal of Information and Data Management, 2024. 32

[64] Demšar, Janez: *Statistical comparisons of classifiers over multiple data sets*. Machine Learning Research, 2006. 32, 43

[65] Castilho, Richard Eckart de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank e Chris Biemann: *A web-based tool for the integrated annotation of semantic and syntactic structures*. Em *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*. The COLING, 2016. 34

[66] Hoekstra, Rinke, Joost Breuker, Marcello Di Bello e Alexander Boer: *The lkif core ontology of basic legal concepts*. IJHPCA, 2007. 34

[67] Mendonça Jr, Carlos AE, Hendrik Macedo, Thiago Bispo, Flávio Santos, Nayara Silva e Luciano Barbosa: *Paramopama: a brazilian-portuguese corpus for named entity recognition*. XII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC, 2015. 34

[68] Tjong Kim Sang, Erik F. e Fien De Meulder: *Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition*. Em *Conference on Natural Language Learning*, 2003. 34

[69] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey *et al.*: *Google's neural machine translation system: Bridging the gap between human and machine translation*. Computing Research Repository, 2016. 37

[70] Schuster, Mike e Kaisuke Nakajima: *Japanese and korean voice search*. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. 37

[71] Kudo, Taku e John Richardson: *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. Em *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2018. 37

[72] Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami e Chris Dyer: *Neural architectures for named entity recognition*. Em *Human Language Technologies*. Association for Computational Lingustics, 2016. 39

[73] Jurafsky, Daniel e James H. Martin: *Speech and Language Processing*. Pearson, 2023. 42

[74] and, Milton Friedman: *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*. Journal of the American Statistical Association, 1937. 42

[75] Lopes, Fernando e Luís Garcia: *Evaluation of a hybrid approach to legal entity recognition*. Em *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*. SBC, 2025. 60