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

ROSIMEIRE PEREIRA DA COSTA

Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo

Goiânia
2023

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

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

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

1. Identificação do material bibliográfico

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

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

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

2. Nome completo do autor

Rosimeire Pereira da Costa

3. Título do trabalho

Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo

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

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

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

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

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Coordenador de Curso**, em 19/05/2023, às 12:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rosimeire Pereira Da Costa**, **Discente**, em 19/05/2023, às 15:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Referência: Processo nº 23070.017602/2023-72 SEI nº 3763728

ROSIMEIRE PEREIRA DA COSTA

Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo

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

Área de concentração: Ciência da Computação.

Orientadora: Profa. Dra. Nádia Félix Felipe da Silva

Coorientadora: Profa. Dra. Ellen Polliana Ramos Souza

Goiânia
2023

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

Costa, Rosimeire Pereira da
Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo [manuscrito] / Rosimeire Pereira da Costa. - 2023.
lxx, 70 f.

Orientador: Profa. Dra. Nádia Félix Felipe da Silva; co orientadora Dra. Ellen Polliana Ramos Souza.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2023.
Bibliografia. Apêndice.
Inclui siglas, lista de figuras, lista de tabelas.

1. Reconhecimento de Entidades Nomeadas. 2. Processamento de Linguagem Natural. 3. Textos informais. 4. BERT. 5. Legislativos. I. Félix Felipe da Silva, Nádia, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **03/2023** da sessão de Defesa de Dissertação de **Rosimeire Pereira da Costa**, que confere o título de Mestra em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos dezenove dias do mês de abril de dois mil e vinte e três, a partir das 09 horas, via webconferência, realizou-se a sessão pública de Defesa de Dissertação **"Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo"**. Os trabalhos foram instalados pela Orientadora, Professora Doutora Nádia Félix Felipe da Silva (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professora Doutora Ellen Polliana Ramos Souza (UAST/UFRPE), coorientadora; Professor Doutor Sérgio Francisco da Silva (IBiotec-UFCat), membro titular externo; Professora Doutora Deborah Silva Alves Fernandes (INF/UFG), membra titular interna. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pela Professora Doutora Nádia Félix Felipe da Silva, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dezenove dias do mês de abril de dois mil e vinte e três.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 19/04/2023, às 10:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **ELLEN POLLIANA RAMOS SOUZA**, **Usuário Externo**, em 19/04/2023, às 10:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Coordenador de Curso**, em 19/04/2023, às 10:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rosimeire Pereira Da Costa**, **Discente**, em 19/04/2023, às 13:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sérgio Francisco Da Silva**, **Professor do Magistério Superior**, em 19/04/2023, às 15:10, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3651278** e o código CRC **D3D58CD6**.

Referência: Processo nº 23070.017602/2023-72 SEI nº 3651278

Rosimeire Pereira da Costa

Graduou–se em Estatística na UFG – Universidade Federal de Goiás. Durante sua graduação, foi monitora e pesquisadora CNPq no Instituto de Matemática e Estatística da UFG. Durante o Mestrado, também na UFG, realizou o estágio de docência e concentrou sua pesquisa no campo do Reconhecimento de Entidades Nomeadas em Textos Informais. Atualmente desenvolve soluções de Processamento de Linguagem Natural em produtos bancários.

Dedico este trabalho a todos aqueles que acreditaram em mim, apoiaram meu crescimento e me inspiraram a perseguir meus sonhos. À minha família, expresso minha profunda gratidão pelo amor incondicional e suporte contínuo que sempre me deram. Aos meus amigos, dedico meu reconhecimento pelo incentivo e pelos momentos de descontração que tornaram essa jornada mais leve. Também dedico aos meus mentores e professores, cuja sabedoria e orientação moldaram meu caminho acadêmico.

Agradecimentos

Em primeiro lugar, gostaria de expressar minha profunda gratidão a Deus pela oportunidade de realizar este trabalho. Sou extremamente grata à minha família, em especial ao meu pai, Pedro Rita Xavier da Costa, e à minha madrasta, Jaqueline Lopes, pelo constante incentivo e por acreditarem em mim.

Gostaria de agradecer imensamente à minha orientadora, Nádia Félix, e à minha coorientadora, Ellen Polliana, por terem aceitado conduzir este trabalho e por terem desempenhado um papel brilhante nele. Agradeço também aos professores do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Goiás (PPGCC-UFG), em especial a Hugo Alexandre, Hebert Coelho, Humberto Longo e Celso Camilo, pelo conhecimento que me foi proporcionado.

Não posso deixar de expressar minha gratidão aos meus amigos, especialmente às minhas amigas Laís Franco e Samela Spíndola, que tornaram minha vida acadêmica mais agradável. Agradeço também ao Renan Ofugi, por seu constante apoio ao longo deste trabalho.

Por último, mas certamente não menos importante, expresso minha sincera gratidão à Universidade Federal de Goiás e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela valiosa bolsa de pesquisa que me foi oferecida para o desenvolvimento deste trabalho. Também gostaria de estender meus agradecimentos ao Projeto Ulysses, no qual tive a oportunidade de participar, e em especial, agradeço a Hidelberg Oliveira Albuquerque por suas discussões enriquecedoras e pela colaboração ao longo do processo.

Nós morremos. Esse pode ser o sentido da vida. Mas nós fazemos a linguagem. Essa pode ser a medida das nossas vidas.

Toni,
Morrison.

Resumo

Pereira da Costa, Rosimeire. **Reconhecimento de Entidades Nomeadas em Textos Informais no Domínio Legislativo**. Goiânia, 2023. 70p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

A tarefa de Reconhecimento de Entidades Nomeadas (REN) em Processamento de Linguagem Natural (PLN) é considerada desafiadora para idiomas complexos, como o português, especialmente quando aplicada em um contexto de linguagem informal e textos curtos, exigindo a manipulação de um léxico específico do domínio em questão. Neste estudo, foi expandido o *corpus* UlyssesNER-Br para a tarefa de REN, utilizando comentários em português do Brasil sobre projetos de leis. Além disso, o conjunto anotado foi enriquecido com um *corpus* formal, a fim de avaliar se a combinação de textos formais e informais de um mesmo domínio poderia melhorar o desempenho de modelos de REN. Foram utilizadas quatro arquiteturas de modelos de REN para execução dos experimentos: *Conditional Random Fields* (CRF), *Bidirectional* LSTM-CRF (BiLSTM-CRF), e o *fine-tuning* do modelo de linguagem BERT e RoBERTa para a tarefa de REN. Os resultados mostraram que o uso de textos formais auxiliou na identificação de entidades em textos informais, em todas arquiteturas de modelos de REN utilizadas no estudo. O modelo que obteve melhor desempenho foi o *fine-tuning* do BERT, com um F1-score de 78,65%. Esses resultados indicam que a combinação de textos formais e informais pode ser uma estratégia eficaz para melhorar o desempenho de modelos de REN em português do Brasil, e que o BERT é uma opção promissora para a tarefa.

Palavras–chave

Reconhecimento de Entidades Nomeadas, Processamento de Linguagem Natural, Textos informais, *BERT*, Português, Legislativos

Abstract

Pereira da Costa, Rosimeire. **Recognition of Named Entities in Informal Texts in the Legislative Domain**. Goiânia, 2023. 70p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The task of Named Entity Recognition (NER) in Natural Language Processing (NLP) is considered challenging for complex languages such as Portuguese, especially when applied in an informal language context and short texts, requiring the manipulation of a specific domain lexicon. In this study, the UlyssesNER-Br corpus was expanded for the NER task, using Brazilian Portuguese comments on bills. In addition, the annotated set was enriched with a formal corpus to evaluate whether the combination of formal and informal texts from the same domain could improve the performance of NER models. Four NER model architectures were used to perform the experiments: Conditional Random Fields (CRF), Bidirectional LSTM-CRF (BiLSTM-CRF), and fine-tuning of the BERT and RoBERTa language models for the NER task. The results showed that the use of formal texts aided in the identification of entities in informal texts for all NER model architectures used in the study. The model that performed best was the fine-tuning of BERT, with an F1-score of 78.65%. These results indicate that the combination of formal and informal texts can be an effective strategy to improve the performance of NER models in Brazilian Portuguese, and that BERT is a promising option for the task.

Keywords

Named Entity Recognition, Natural Language Processing, Informal Texts, BERT, Portuguese, Legislative

Lista de Abreviaturas e Siglas

BERT - *Bidirectional Encoder Representations from Transformers*
BiLSTM - *Bidirectional Long Short-Term Memory*
CNN - *Convolutional Neural Networks*
CoNLL - *Computational Natural Language Learning*
CRF - *Conditional Random Fields*
ELMo - *Embeddings from Language Model*
EN - Entidade Nomeada
HMMs - *Hidden Markov Models*
LSTM - *Long Short-Term Memory*
MSM - *Making Sense of Microposts*
MUC - *Message Understanding Conference*
NILC - Núcleo Interinstitucional de Linguística Computacional
PLN - Processamento de Linguagem Natural
REN - Reconhecimento de Entidades Nomeadas
RNN - *Recurrent Neural Networks*
RoBERTa - *Robustly Optimized* BERT
UT - *User-generated text*
W-NUT - *Workshop on Noisy User-generated Text*

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

1.1 Motivação

O Reconhecimento de Entidades Nomeadas (REN) origina-se da tarefa de Extração de Informações (EI), que transforma dados não estruturados em informações estruturadas. Neste cenário, os dados não estruturados são textos escritos em linguagem natural, e a finalidade da tarefa de EI é extrair informações importantes em um formato bem definido. Logo, a subtarefa REN encontra e classifica a Entidade Nomeada (EN) em uma classe predefinida [58].

O termo EN, foi citado pela primeira vez em 1996, na VI *Message Understanding Conference* (MUC-6) [45], que teve como foco tarefas de Extração de Informação (EI). O REN visa identificar menções a designadores rígidos em textos pertencentes a tipos semânticos predefinidos [80].

A definição formal para o REN pode ser apresentada da seguinte forma: dada uma sequência de *tokens* $s = \{w_1, w_2, \ldots, w_N\}$, um modelo REN deve gerar uma lista de tuplas $< I_{inicio}, I_{fim}, t >$, em que cada uma representa uma EN em s, onde $I_{inicio} \in [1, N]$ e $I_{fim} \in [1, N]$ são os índices que indicam o início e o fim da EN, e t o tipo da entidade presente em um conjunto predefinido de categorias [86].

Conforme Konkol [58], desde a sua introdução, o REN provou ser uma etapa de pré-processamento útil para muitas tarefas de Processamento de Linguagem Natural (PLN), por exemplo: na tradução automática, a EN é traduzida de forma diferente das outras palavras; na sumarização, o segmento contendo uma EN pode ser mais relevante; no agrupamento de documentos, os documentos que contêm as mesmas ENs são provavelmente sobre o mesmo assunto; entre outras.

Desde o MUC-6, essa tarefa tem sido explorada por vários pesquisadores [11, 15, 56], e é assunto de estudo em vários eventos (MUC-6 [45], CoNLL03 [109], IREX [33], e TREC Entity Track [91]), sendo um deles o *Workshop on Noisy User-generated Text* (W-NUT), que em 2015, inseriu uma tarefa compartilhada de REN em textos do *Twitter* [8], uma vez que grande parte dos sistemas de REN foram desenvolvidos para textos de notícias e não possuem um desempenho satisfatório em gêneros mais informais [93].

No W-NUT, os autores utilizam o termo *user-generated text* (UT) ([81]) para se referirem à conteúdo textual gerado pelo usuário. Neste trabalho os termos *user-generated text* (UT) e textos informais são usados indistintamente.

De acordo com Derczynski et al. [35], REN é descrito por alguns pesquisadores como uma tarefa resolvida devido a altas pontuações relatadas em conjuntos de dados bem conhecidos, mas, na verdade, os sistemas que atingem essas pontuações tendem a falhar em entidades mais raras ou inéditas, fazendo com que a maior parte de sua pontuação de desempenho seja superior com entidades bem formadas e não inéditas [35]. Esse fenômeno é explicado pela distribuição Zipfiana presente na maioria das expressões linguísticas, conforme evidenciado em estudos anteriores [120, 78].

A distribuição Zipfiana é caracterizada por um pequeno número de observações muito frequentes e uma cauda longa de observações menos frequentes. Como resultado desse viés e da forma como muitas abordagens de PLN são desenvolvidas, é comum que muitos sistemas de PLN deem mais importância às observações de alta frequência em detrimento das observações de baixa frequência [36].

Portanto, este trabalho tem como motivação a investigação do desempenho de sistemas de REN em um *corpus* do domínio legislativo, composto por comentários de cidadãos em uma plataforma de mídia social aberta ao público. Essa pesquisa tem o potencial de contribuir para o desenvolvimento de ferramentas que auxiliem no acompanhamento de discussões legislativas nessas plataformas. Ao detectar e classificar entidades nomeadas nos comentários dos cidadãos, os sistemas REN podem facilitar a identificação de tópicos relevantes, atores importantes e perspectivas diversas relacionadas às questões legislativas. Essas informações podem fornecer *insights* valiosos para parlamentares, governos e outros interessados na tomada de decisões políticas embasadas.

1.2 Desafios de REN em UT

Li et al. [64] afirmam que um dos desafios na tarefa REN, seria o de desenvolver modelos que alcancem um melhor desempenho em textos informais, uma vez que o melhor resultado obtido no W-NUT-2017 [36] apresentou um F1-score um pouco acima de 40%. De acordo com o autor, alcançar uma melhor pontuação no texto informal é mais desafiador do que em um texto formal, devido ao seu tamanho (geralmente mais curto) e o ruído.

Ainda de acordo com Li et al. [64], um outro fator que dificulta os sistemas de REN a alcançarem um bom desempenho em textos informais, é o fato de que os textos gerados pelo usuário possam pertencer a um domínio específico, contendo nomes técnicos de uma área de conhecimento.

Nagarajan [81] elenca características da linguagem nas plataformas de mídias sociais nas quais podem prejudicar técnicas e algoritmos tradicionais:

- **Informal e não policiada**: A comunicação nas plataformas de mídia social é voltada para a comunicação interpessoal e é inerentemente menos formal. Uma grande parte da linguagem está consequentemente no domínio do idioma informal - uma mistura de abreviaturas, gírias e termos específicos de contexto sem regularidades e entregues com uma abordagem indiferente à gramática e ortografia. Isso, juntamente com a natureza não mediada da maioria das plataformas, fornece dados de qualidade variável.
- **Criatividade e Variabilidade**: A variabilidade na produção da linguagem é mais acentuada em mídias sociais do que em outros meios, como notícias ou artigos científicos. Essa diferença se deve, em grande parte, ao grande volume e diversidade de participantes presentes nessas plataformas. Além disso, uma parcela significativa de usuários de mídias sociais é composta por adolescentes que se envolvem em formas bastante criativas de expressão online. Esse uso criativo muitas vezes envolve o emprego de gírias, expressões bastante comuns, mas que requerem um tratamento especial na interpretação automática de suas partes.
- **Contexto compartilhado**: as conversas nas mídias sociais são tipicamente interações independentes entre pessoas com ideias semelhantes, onde um autor geralmente fala para um público conhecido que já tem um senso de contexto compartilhado. Consequentemente, o conteúdo gerado carece de contexto explícito, deixando espaço para ambiguidade em sua interpretação.
- **Protocolos Médios**: Cada plataforma difere no protocolo de expressão que impõe. Em alguns casos, o meio social não permite uma expressão elaborada. O *Twitter*, uma plataforma de *microblog* popular recente, limita a expressão do usuário a um conteúdo de 140[1] caracteres, consequentemente limitando a quantidade de informações contextuais disponíveis para os sistemas. Os fóruns de discussão e respostas a perguntas online, por outro lado, incentivam a discussão do usuário e aumentam as chances de discussões fora do tópico. A estrutura de uma conversa online em certas plataformas também é, na melhor das hipóteses, arbitrária. Embora os encadeamentos de conversas sejam indicadores de contexto valiosos, eles não são rastreáveis uniformemente em todas as plataformas.

Considerando os desafios apresentados na análise de textos informais, o objetivo deste trabalho é contribuir para o aumento da quantidade de *corpora* disponíveis para o

[1]Em 2010, o *Twitter* impunha um limite de 140 caracteres para as postagens, mas atualmente, no momento em que este trabalho está sendo escrito, esse limite foi expandido para 280 caracteres.

REN na língua portuguesa, com o intuito de viabilizar o desenvolvimento de sistemas de REN capazes de lidar com textos gerados pelos usuários em linguagem informal.

1.3 Justificativa

De acordo com Castro [21], em estudos existentes sobre REN aplicados ao português, como no trabalho desenvolvido por Collovini et al. [27], há evidências de que os modelos existentes para a língua portuguesa ainda possuem muita dificuldade em alcançar o estado da arte, o que pode ser justificado pelo baixo volume de informações textuais e ferramentas desenvolvidas para esse idioma quando comparados com o inglês.

Existem vários trabalhos de REN voltados para a área do direito [21, 72, 96, 101, 6, 7, 20, 38], porém, poucos destes são voltados para a língua portuguesa [21, 72, 4, 96].

No que se refere a disponibilização de *corpus* para REN em textos formais, identificou-se apenas dois trabalhos [72, 4] que tratam especificamente de textos legislativos na língua portuguesa.

Embora hajam estudos desenvolvidos recentemente, em tarefas de PLN, em *corpora* informais para língua portuguesa [87, 19, 23, 49, 47, 28], não foi identificado nenhum trabalho desenvolvido especificamente para a tarefa de REN.

Dada a carência de *corpora* disponibilizados para o REN na língua portuguesa somado à falta de trabalhos desenvolvidos para textos informais, esse estudo justifica-se pela importância da disponibilização de um *corpus* de texto informais no domínio legislativo para o REN, visando o aumento de produções acadêmicas nessa área de pesquisa, e consequentemente a melhoria dos desempenhos dos sistemas REN desenvolvidos para língua portuguesa.

1.4 Objetivos

O principal objetivo deste trabalho é o de disponibilizar um novo *corpus* para REN cuja a fonte são de textos gerados pelo usuário contendo termos do âmbito legislativo.

Os objetivos específicos são:

- Construir um *corpus* para a tarefa REN baseado em comentários sobre projetos de lei para a língua portuguesa - BR;
- Avaliar os benefícios de se treinar um modelo a partir de duas coleções de textos anotadas no mesmo domínio, porém com níveis de formalidade diferentes;
- Realizar uma análise experimental do *corpora* usando modelos de aprendizado profundo.

1.5 Hipótese de Pesquisa

Diferentemente dos textos de jornais, que são as fontes mais comuns para a construção de *corpora* para o REN [63], os documentos legais apresentam características específicas de domínio. No entanto, há uma carência de *corpora* disponíveis para a tarefa de REN em língua portuguesa, nesse domínio. Diante desse cenário, o intuito desse trabalho foi justamente aumentar a quantidade de *corpora* disponíveis no domínio legislativo, visando melhorar a precisão e eficácia do sistema de REN em português.

Além de aumentar a quantidade de *corpora* disponíveis no domínio legislativo para o REN, este trabalho propôs o desenvolvimento de um sistema capaz de lidar com textos gerados pelo usuário em linguagem informal no contexto legislativo. Nesse cenário, buscamos avaliar as hipóteses 1 e 2 para o contexto de Reconhecimento de Entidades Nomeadas no domínio legislativo:

Hipótese 1: A junção de textos formais e textos informais pode contribuir significativamente para melhorar o desempenho dos modelos de REN na detecção de entidades em textos informais no domínio legislativo.

Hipótese 2: Os modelos baseado na arquitetura *transformers* (atual estado da arte para tarefas de Processamento de Linguagem Natural) possuem um melhor desempenho para sistemas REN.

1.6 Estrutura do Trabalho

Os capítulos do projeto de pesquisa estão organizadas da seguinte forma:

- O Capítulo 2 trata da fundamentação teórica do trabalho, abordando os principais conceitos na área de REN;
- O Capítulo 3 trata dos trabalhos relacionados, abordando os principais *benchmarks* de avaliação de sistemas de REN; e descreve os principais trabalhos desenvolvidos na área de REN para textos informais;
- O Capítulo 4 aborda a metodologia do trabalho, os procedimentos que foram seguidos para se obter os resultados experimentais;
- No Capítulo 5 são apresentados os resultados experimentais a partir do processo de avaliação dos modelos.
- No Capítulo 6 é apresentado a conclusão e as contribuições do trabalho.

CAPÍTULO 2

Fundamentação Teórica

De acordo com Veneroso [111], tradicionalmente, a tarefa REN era resolvida com modelos generativos baseados em *Hidden Markov Models* (HMMs). O primeiro aparecimento de HMMs na área de PLN ocorreu em meados dos anos 70, no qual focou-se no problema de reconhecimento de fala. Mas no final dos anos 90, os HMMs também tiveram contribuições importantes no REN, como nos trabalhos [12, 40, 41].

O *maximum-entropy Markov model* [74] foi desenvolvido um pouco mais tarde, baseado nos HMMs. No entanto, devido ao problema de viés de rótulo, esse modelo foi substituído pelo *Conditional Random Fields* (CRF) [60, 75]. Nessa época, os sistemas de melhor desempenho quase sempre recorriam a dicionários geográficos externos e *features* escolhidas a dedo, como é o caso do modelo vencedor da *Conference on Computational Natural Language Learning* - CoNLL03 [39].

CRF é um algoritmo de aprendizado supervisionado comumente usado na tarefa de REN. De acordo com Laffert et al. [60], é um método de modelagem estatística que é frequentemente aplicado para classificação sequencial. Muitos trabalhos [64, 103, 10], relataram um desempenho satisfatório do CRF, para a tarefa REN.

Em 2011, Collobert et al. [26] introduziram as redes neurais para rotular tarefas de sequências, usando *Convolutional Neural Network* (CNN) sobre *word embeddings* com CRF na camada de saída para resolver tarefas de rotulagem, como *part-of-speech* (POS) *tagging*, *chunking*, rotulagem de função semântica e REN. Uma arquitetura semelhante proposta em 2015 por Huang et al. [53], uma *Long Short-Term Memory* - LSTM bidirecional com CRF (BiLSTM-CRF), alcançou melhores resultados que o modelo proposto por Collobert et al. [26].

Melhorias nos sistemas REN vem sendo feitas desde então, com a introdução de *embeddings* pré-treinados e variações da arquitetura BiLSTM-CRF. Desse grupo, pode-se citar *Embeddings from Language Model* - ELMo [92], *Bidirectional Encoder Representations from Transformers* - BERT [37] e Flair [3]. Todos eles foram introduzidos em datas próximas e diferem principalmente na forma de construir *embeddings* contextuais a partir dos estados internos de uma rede neural.

O modelo de linguagem BERT, apresentado por Devlin et al. [37], é um novo

método de pré-treinamento de representações de linguagem que obtém resultados de última geração em uma ampla gama de tarefas de PLN. O BERT foi projetado para pré-treinar representações bidirecionais profundas de texto não rotulado, condicionando conjuntamente o contexto esquerdo e direito em todas as camadas.

Como resultado, o modelo BERT pré-treinado pode ser ajustado com apenas uma camada de saída adicional para criar modelos de última geração para tarefas, como: REN, Resposta a Perguntas, e Inferência de Idioma, sem modificações substanciais na arquitetura específica da tarefa [24].

Mais recentemente, foi apresentado o RoBERTa, um melhoria do BERT, com alterações no procedimento de pré-treinamento no qual alcançou resultado de ponta em uma variedade de tarefas de PLN, inclusive REN [70, 67, 65].

Diversos trabalhos [53, 26, 22, 61, 73] utilizaram abordagens voltadas para arquiteturas baseadas em Redes Neurais Recorrentes Bidirecionais (BiLSTM) sendo que em muitos destes, foram adicionadas uma camada CRF para classificação. Alguns estudos que utilizam essa abordagem, focam em comparar diferentes *embeddings* textuais [92, 37, 3].

Como apresentado, modelos como CRF, BiLSTM-CRF, BERT e RoBERTa obtiveram resultados satisfatórios em vários trabalhos, e como foram implementados nesse trabalho, as Seções 2.1, 2.2, 2.3 e 2.4, serão dedicadas a descrever com maiores detalhes os modelos CRF, BiLSTM-CRF, BERT e RoBERTa, de forma respectiva.

2.1 *Conditional Random Fields*

A capacidade de prever múltiplas variáveis que dependem umas das outras é fundamental para muitas aplicações, em que deseja-se prever um vetor de saída $\mathbf{y} = \{y_0, y_1, ..., y_T\}$ de variáveis aleatórias dado um vetor de *features* observado $\mathbf{x}$. Um exemplo relativamente simples de PLN é a etiquetagem POS, na qual cada variável y_s é a *part-of-speech tag* na posição s, e a entrada $\mathbf{x}$ é dividida em vetores de *features* $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}$. Cada $\mathbf{x}_s$ contém várias informações sobre a palavra na posição s, como *features* ortográficas, prefixos e sufixos, entre outras [107].

Uma abordagem para o problema de previsão multivariada, é aprender um classificador independente por posição, que mapeie $\mathbf{x} \rightarrow y_s$ para cada s. A dificuldade, no entanto, é que as variáveis de saída têm dependências complexas, como por exemplo, em inglês, adjetivos geralmente não seguem substantivos. Outro obstáculo é que as variáveis de saída podem representar uma estrutura complexa. Uma maneira natural de representar a forma que as variáveis de saída dependem umas das outras é fornecida por modelos gráficos, como: *Bayesian networks, Markov random fields*, e outros, sendo

possível descrever como uma dada fatoração da densidade de probabilidade corresponde a um determinado conjunto de relações de independência condicional [107].

Um problema a ser abordado é a dependência das variáveis de entrada. Uma alternativa para lidar com modelos gráficos que visam modelar uma distribuição de probabilidade conjunta $p(\mathbf{y},\mathbf{x})$ sobre as entradas e saídas, é seguir uma abordagem discriminativa, semelhante à adotada em classificadores como a regressão logística. O objetivo é modelar diretamente a distribuição condicional $p(\mathbf{y}|\mathbf{x})$, que é essencial para a classificação. Essa é a abordagem adotada pelo CRF [107, 76].

O modelo CRF foi proposto por Lafferty et al. [60], sendo um modelo gráfico não direcionado, que é treinado para maximizar uma probabilidade condicional.

Um CRF de cadeia linear com parâmetros $\Delta = \{\lambda, ...\}$ define uma probabilidade condicional para uma sequência de estado (ou rótulo) $y = y_1 \dots y_T$, dada uma sequência de entrada $x = x_1 \dots x_T$, em que T é o comprimento da sequência. Tem-se então

$$P_\lambda(y|x) = \frac{1}{z_x} \exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{2-1}$$

em que z_x é a constante de normalização que faz a probabilidade de todas as sequências de estado somarem 1; $f_k(y_{t-1}, y_t, x, t)$ é uma função que geralmente assume valor binário; λ_k é um peso aprendido associado a f_k. Grandes valores positivos para λ_k indicam uma preferência por tal evento, enquanto grandes valores negativos tornam o evento improvável [62].

2.2 BiLSTM-CRF

Uma rede neural LSTM bidirecional empilha dois LSTMs regulares e o alimenta com observações em direções opostas. A primeira rede LSTM recebe estados para frente e a segunda LSTM recebe estados para trás. Os estados ocultos de ambas as redes podem ser concatenados em cada etapa de tempo para produzir rótulos de saída. Com esta arquitetura, as unidades LSTM podem usar informações de passos de tempo passados e futuros para decidir a etiqueta no tempo t [111].

No entanto, com uma rede LSTM bidirecional, os rótulos são decodificados individualmente em cada etapa de tempo, e para contornar esse problema, Huang et al. [53] propuseram uma rede LSTM bidirecional com uma camada CRF (BiLSTM-CRF) na saída. O principal benefício de adicionar uma camada de CRF no modelo de sequência neural é que os rótulos são decodificados em conjunto para uma frase inteira, em vez de serem previstos individualmente [111].

A arquitetura BiLSTM-CRF é ilustrada na Figura 2.1, onde a entrada da rede é uma sequência de *embeddings* (X_1, ...,X_n). A camada BiLSTM processa a sequência de

embeddings em ambas as direções (*forward* e *backward*), capturando o contexto passado e futuro de cada palavra. Por fim, a camada CRF aplica um algoritmo à saída da camada BiLSTM, atribuindo um rótulo a cada *token* da sequência com base nas probabilidades dos rótulos dados à sequência [53].

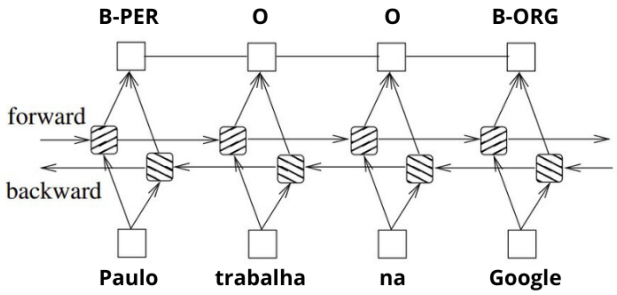


Figura 2.1: *Arquitetura BiLSTM-CRF.*
Fonte: Adaptado de Huang et al. [53].

2.3 BERT

Nos últimos anos, os modelos de texto mais avançados usam *transformers* para aprender como representar o texto. Os *transformers* são um tipo de rede neural que está encontrando cada vez mais seu uso em diversos ramos do aprendizado de máquina, na maioria das vezes em problemas quando as informações de entrada e saída são uma sequência [44].

Tais modelos usam uma combinação de redes neurais recorrentes e convolucionais. Redes recorrentes regulares têm um desempenho bastante ruim com contexto de longo alcance. No entanto, no texto natural, a representação do *token* pode ser influenciada pelo contexto por meio de diversas palavras e até frases do próprio *token*. Para levar em conta a influência de longo alcance, LSTMs são usados em conjunto com o mecanismo de atenção para melhorar a eficiência do aprendizado, levando em consideração a influência de *tokens* distantes [43].

No final de 2018, um grupo de cientistas do laboratório *Google AI Language* sob a liderança de J. Devlin apresentou um novo modelo linguístico denominado BERT [37]. Este modelo destina-se ao aprendizado preliminar profundo da representação de texto bidirecional para uso posterior em modelos de aprendizado de máquina. A vantagem deste modelo é sua facilidade de uso, que envolve adicionar apenas uma camada de saída à arquitetura neural existente para obter modelos de texto que superam a imprecisão de todos os existentes em diversos problemas de PLN [59].

Conforme Heck [52], o BERT é um modelo de linguagem baseado na arquitetura *transformer*, que utiliza apenas a camada do *encoder* para gerar um modelo da linguagem.

Durante o pré-treinamento, esse modelo é treinado com grandes quantidades de textos no idioma desejado. A entrada para o modelo é a soma dos *token embeddings* (a sequência de entrada segmentada em *tokens*), *segment embeddings* (indicam a que frase cada *token* pertence) e dos *position embeddings* (indicam a ordem dos *tokens* no texto), como apresentado na Figura 2.2.

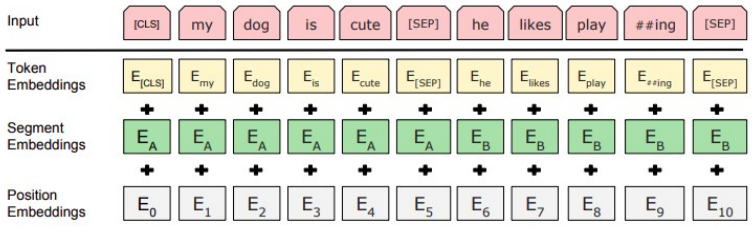


Figura 2.2: *Representação da entrada de uma arquitetura BERT.*
Fonte: Devlin et al., 2018 [37].

Uma rede BERT é pré-treinada em duas tarefas, *Masked Language Modeling* (MLM) e *Next Sentence Prediction* (NSP), para aprender o contexto de um idioma. No MLM, uma determinada proporção de *tokens* é escolhida aleatoriamente antes das sequências de texto serem usadas como entrada na rede de pré-treinamento. Destes, uma porcentagem é substituída pelo *token* `[MASK]`, outra porcentagem é substituída por *tokens* aleatórios e outra porcentagem não é alterada. O objetivo da rede é prever os *tokens* originais com base no contexto fornecido pelos demais. Durante o treinamento, o modelo é exposto a grandes quantidades de dados textuais e aprende a extrair padrões e relações semânticas entre as palavras [37].

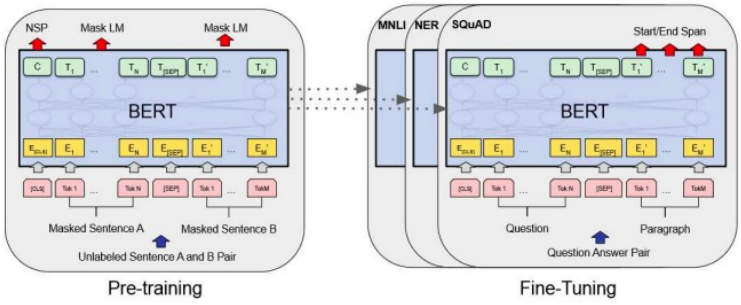


Figura 2.3: *Pré-treinamento e fine-tuning de uma rede BERT.*
Fonte: Devlin, 2018 [37].

Segundo Devlin et al. [37] depois do pré-treinamento, a etapa de *fine-tuning* é crucial para adaptar a rede pré-treinada à tarefa final que se deseja resolver. Por exemplo, no caso de REN, o *fine-tuning* envolve ajustar os pesos da rede para que ela seja capaz de identificar corretamente as ENs em um texto (Figura 2.3).

Em comparação com o pré-treinamento, o *fine-tuning* é relativamente barato em termos computacionais e pode ser concluído em questão de horas em TPUs ou GPUs.

Isso ocorre porque a etapa de pré-treinamento já extraiu as informações semânticas importantes do texto, e o *fine-tuning* consiste em ajustar os pesos da rede para uma tarefa específica. No geral, o pré-treinamento e o *fine-tuning* são etapas cruciais no uso de uma arquitetura BERT para PLN [37].

2.4 RoBERTa

Robustly optimized BERT *approach* (RoBERTa) é o modelo de linguagem desenvolvido pela *Facebook AI Research* (FAIR), em 2019. Esse modelo é semelhante ao BERT, pois também é uma arquitetura de rede neural baseada em *transformer*, mas é treinado em um conjunto de dados maior e mais diversificado, com tempo de treinamento mais longo e técnicas mais avançadas para pré-processamento e treinamento de dados. O RoBERTa também usa mascaramento dinâmico durante o treinamento, o que o ajuda a entender melhor as relações entre as palavras e seu contexto [70].

RoBERTa alcançou resultados de ponta em uma variedade de tarefas de PLN, incluindo análise de sentimento, REN e resposta a perguntas. Seu sucesso o levou a ser usado como modelo base em muitas aplicações *downstream* na indústria e na academia [70].

Além disso, RoBERTa$_{LARGE}$ também superou consistentemente modelos pré-treinados anteriores, como BERT$_{LARGE}$ e XLNet$_{LARGE}$, em todas as nove tarefas de desenvolvimento do GLUE, como é possível observar na Tabela 2.1. Os resultados são apresentados considerando uma única tarefa, no conjunto de desenvolvimento. É importante ressaltar que, apesar de utilizar o mesmo objetivo de pré-treinamento de modelagem de linguagem mascarada e arquitetura que o BERT$_{LARGE}$, o RoBERTa$_{LARGE}$ superou consistentemente tanto o BERT$_{LARGE}$ quanto o XLNet$_{LARGE}$ [70].

Tabela 2.1: *Resultados no GLUE. Todos os resultados são baseados em uma arquitetura de 24 camadas. Os resultados de BERT$_{LARGE}$ e XLNet$_{LARGE}$ são de Devlin et al. [37] e Yang et al. [119], respectivamente. Os resultados do RoBERTa$_{LARGE}$ no conjunto de desenvolvimento são uma média de cinco execuções.*

Modelo	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI
BERT$_{LARGE}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-
XLNet$_{LARGE}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-
RoBERTa$_{LARGE}$	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**

Fonte: Liu et al. [70].

2.5 Métricas de Avaliação

De acordo com Nadeau e Sekine [80], várias técnicas foram propostas para avaliar sistemas de REN. Jiang et al. [54] explicam que a avaliação serve para verificar a capacidade do modelo em encontrar os limites dos nomes e seus tipos corretos. Os sistemas de REN podem ser classificados com base em dois protocolos de pontuação:

a) *Correspondência exata* para limite e tipo: Essa medida avalia a capacidade do sistema em realizar uma detecção precisa de entidades nomeadas. O sistema recebe pontuação máxima apenas se conseguir identificar corretamente o limite (início e fim) de uma entidade nomeada e atribuir o tipo correto a ela.

b) *Correspondência parcial* para limite, contada somente quando o tipo detectado está correto: Essa medida atenua as falhas de correspondência exata quando as diferenças de limite são causadas por palavras sem importância nos nomes, como artigos e preposições. Nesse caso, mesmo que ocorra uma correspondência parcial nos limites, o sistema ainda pode obter pontuação, desde que o tipo da entidade nomeada seja detectado corretamente.

De acordo com Castro [21], uma vez adotado o critério de correspondência, o desempenho final do modelo é calculado usando a Medida F (F1-score), que é obtida a partir da Precisão (P) e do *Recall* (R) do modelo. As equações (2-2), (2-3) e (2-4) mostram como calcular o F1-score com base na contagem das entidades classificadas pelo modelo de REN.

$$\mathrm{P} = \frac{\mathrm{VP}}{\mathrm{VP+FP}} \tag{2-2}$$

$$\mathrm{R} = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \tag{2-3}$$

$$\text{F1-score} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{2-4}$$

em que VP (Verdadeiro Positivo) é a quantidade de entidades corretamente identificadas e classificadas pelo modelo, FP (Falso Positivo) é a quantidade de entidades erroneamente identificadas e classificadas pelo modelo, e FN (Falso Negativo) é a quantidade de entidades que deveriam ter sido identificadas pelo modelo, mas foram erroneamente deixadas de ser identificadas. No presente trabalho, é adotado o protocolo de avaliação baseado em correspondência exata.

2.6 Teste Wilcoxon

O teste Wilcoxon, também conhecido como teste Wilcoxon-Mann-Whitney ou teste Mann-Whitney U, é um teste estatístico não paramétrico usado para comparar duas

amostras independentes ou observações pareadas. É comumente empregado quando as suposições de normalidade e variâncias iguais exigidas por testes paramétricos, como o teste t de Student, são violadas ou quando os dados são ordinais ou classificados [115].

O teste é indicado quando as amostras são independentes, com observações pareadas, e se deseja determinar se existe uma diferença significativa na distribuição de uma variável contínua ou ordinal entre os grupos. O teste de Wilcoxon é robusto a violações de suposições de normalidade e fornece uma alternativa para testes paramétricos ao lidar com dados não normalmente distribuídos [115].

CAPÍTULO 3

Trabalhos Relacionados

Os *tweets*, mensagens curtas compartilhadas por meio da rede social *Twitter*, têm se tornado uma importante fonte de informação. Como resultado, a tarefa de reconhecimento de entidades nomeadas em *tweets* tem atraído crescente interesse de pesquisa [69].

Em 2011, os autores Ritter et al. [97] desenvolveram um sistema que utiliza um modelo CRF para segmentar as entidades nomeadas, seguido de uma abordagem de supervisão distante[1], baseada em *Labeled LDA* para classificar essas entidades. No mesmo ano, Liu et al. [68] combinaram um classificador baseado no algoritmo *K-Nearest Neighbors* (KNN) com um modelo baseado em CRF para extrair informações de *tweets* e adotaram o aprendizado semi-supervisionado para aproveitar *tweets* não rotulados.

Em 2011, também foi realizado o primeiro *workshop Making Sense of Microposts* (MSM) [99], com o objetivo principal de reunir pesquisadores que exploram novos métodos para analisar micropostagens. Em 2013, o *workshop* organizou um desafio de extração de entidades, no qual um conjunto de dados, composto por 4.341 micropostagens em inglês, contendo comentários sobre notícias e política, foi disponibilizado. O trabalho vencedor do desafio adotou uma abordagem combinando um CRF com *Support Vector Machine* (SVM) para a extração das entidades. Além disso, a *framework* AIDA foi utilizada para a classificação das entidades extraídas. Esse sistema atingiu o F1-score de 67.00% [18].

No ano de 2014, o *workshop* MSM compartilhou mais um desafio em micropostagens em inglês do *Twitter*. Neste desafio, os participantes foram desafiados a desenvolver abordagens que fossem capazes de identificar e vincular entidades nomeadas de forma automática. O trabalho vencedor deste desafio adotou uma abordagem baseada em regras para extração das entidades. Além disso, o trabalho também fez uso de bases de conhecimento existentes para auxiliar na vinculação das entidades extraídas, atingindo o F1-score de 70.06% [17].

[1] A intuição da abordagem de supervisão distante é usar um banco de dados de domínio específico, para fornecer um conjunto de treinamento de relações e pares de entidades que participam dessas relações [77].

Em 2015, o *workshop* MSM promoveu um desafio semelhante ao de 2014, no qual os participantes foram desafiados a reconhecer automaticamente entidades e seus tipos em micropostagens em inglês, além de vinculá-las aos recursos correspondentes do DBpedia 2014. Nesse desafio, foi utilizado um *corpus* aprimorado em relação ao disponibilizado em 2014. O time vencedor adotou uma abordagem combinando o algoritmo *Random Forest* com Regressão Logística para resolver a tarefa proposta, obtendo um F1-score de 80.67% [98].

No mesmo ano, também aconteceu a primeira edição do *Workshop on Noisy User-generated Text* (W-NUT) [8], em que foi apresentado duas tarefas compartilhadas de REN para mensagens do *Twitter*: uma para segmentação e classificação (Tarefa 1), e outra apenas para segmentação de entidades (Tarefa 2) [8]. Esta edição atraiu oito equipes e os vencedores de ambas as tarefas foram Yamada et al. [118], que obtiveram uma pontuação F1-score de 56.41% na Tarefa 1 e 70.63% na Tarefa 2. Eles usaram *Entity Linking*, um método para detectar menções de entidades no texto e resolvê-los para entradas correspondentes em bases de conhecimento como a *Wikipedia*.

Na segunda edição do W-NUT, em que foi disponibilizado o *corpus* W-NUT-16 [106], foi proposta uma tarefa de reconhecimento de entidades nomeadas com 10 equipes participantes. O melhor resultado obtido foi apresentado por Limsopatham e Collier [66], atingindo o F1-score para segmentação e classificação (Tarefa 1) de 52.41% com um modelo BiLSTM, que induz e potencializa automaticamente características ortográficas ao realizar REN. Apenas para segmentação (Tarefa 2), a mesma equipe venceu com uma pontuação F1-score de 65.89%.

No mesmo ano, o *workshop* MSM consolidou a tarefa apresentada em 2015, de reconhecimento automático de entidades e seus tipos, bem como a vinculação a recursos do DBpedia 2014. Foram fornecidas bases de dados aprimoradas, que foram utilizadas tanto em 2014 quanto em 2015, garantindo uma continuidade no avanço e desenvolvimento da área. A equipe vencedora obteve uma pontuação de desempenho geral de 54.86%, utilizando a adaptação de um sistema de desambiguação de entidade nomeada existente – KEA [16, 113].

Ainda em 2016, o *workshop* EVALITA apresentou o desafio de Reconhecimento e Vinculação de Entidade Nomeada no *Tweet* Italiano (NEEL-IT). Nesse desafio, a tarefa consistia em identificar corretamente as menções de entidades em um texto e vinculá-las às entidades nomeadas correspondentes na base de conhecimento DBpedia 2015. A equipe vencedora desenvolveu um sistema que explorou *embeddings* de palavras e utilizou uma rede BiLSTM para o reconhecimento e vinculação de entidades. Esse sistema alcançou um F1-score de 50.34%, demonstrando um desempenho razoável na tarefa proposta [9].

Em 2017, ocorreu a terceira edição do W-NUT, que forneceu o *corpus* W-NUT-

17 [35], com foco em entidades inéditas e raras no contexto de discussões emergentes. Este *corpus* inclui texto do *Twitter*[2], *Reddit*[3], *Youtube*[4] e *StackExchange*[5]. O melhor resultado foi apresentado por Aguilar et al. [2], com um F1-score, ao nível da entidade, de 41.86% (Tarefa 1). A arquitetura de rede neural multitarefa proposta por Aguilar et al. [2] aprende representações de *features* a partir de string de palavras e caracteres, juntamente com *tags Part-of-Speech* e informações de *gazetteers*. Essa rede neural atua como um extrator de *features* para alimentar um classificador CRF. Para a Tarefa 2, a mesma equipe venceu com uma pontuação F1-score de 40.24%.

No ano de 2020, aconteceu a sexta edição do W-NUT, disponibilizando o *corpus* para REN e Extração de Relação (ER) em *Wet-lab Protocols*. O time que venceu utilizou *Ensemble of Transformers* como abordagem para a tarefa REN, obtendo um F1-score de 77.99% [108].

Derczynski et al. [34], em 2016, apresentaram o *Broad Twitter Corpus* (BTC), com o objetivo de superar um dos principais desafios enfrentados no desenvolvimento e avaliação comparativa de REN em mídias sociais: a escassez de um *corpus* anotado considerável, diversificado e de alta qualidade. Agarwal e Nenkova [1] empregaram uma arquitetura BiLSTM-CRF em conjunto com *gazetteers* e segmentação de entidades no *corpus* BTC, obtendo um desempenho notável de 74.70%.

Mais recentemente, Wang et al. [114] mostraram que contextos externos em nível de documento podem melhorar significativamente o desempenho do modelo para os *corpora* W-NUT-16 e W-NUT-17. Eles selecionaram um conjunto de texto por meio de um mecanismo de busca usando a frase original como consulta, e as representações contextuais foram computadas com base na concatenação de cada frase do *corpora* e seus contextos externos. Os autores aplicaram essa representação contextual como entrada no modelo neural com uma camada CRF e obtiveram 58.98% e 60.45% de F1-score para W-NUT-16 e W-NUT-17, respectivamente.

Peng e Dredze [90] forneceram o WeiboNER, um *corpus* chinês para REN, que foi construído a partir de uma plataforma de mídia social chinesa. Os autores avaliaram três tipos de *embeddings* para texto informal e, com o modelo CRF, obtiveram um F1-score de 56.05%.

Xuan et al. [117] proporam um método para REN chinês, que pode extrair informações interativas entre a representação distribuída de caracteres e a representação de glifos por um mecanismo de fusão, denominado por *Fusion Glyph Network* (FGN). Este método pode capturar conhecimento interativo potencial entre contexto e glifo.

[2]https://twitter.com
[3]https://reddit.com
[4]https://youtube.com
[5]https://stackexchange.com

Experimentos mostraram que FGN com LSTM-CRF como *tagger* alcança um novo desempenho de última geração para o conjunto de dados WeiboNER, com um F1-score de 71.25%.

Nguyen et al. [84] disponibilizaram o primeiro modelo público de linguagem pré-treinada em larga escala para *tweets* em inglês, no qual nomearam de BERTweet. O BERTweet possui a mesma arquitetura do BERT_{base} [37], é treinado usando o procedimento de pré-treinamento do RoBERTa [70]. Resultados experimentais apresentaram um F1-score nos *corpora* WNUT-2016 e WNUT-2017 de 52.10% e 56.50%, respectivamente.

Em 2017 houve o evento CAp 2017 challenge [71], que propôs o desafio de REN em textos do *Twitter* na linguagem francesa. Oito equipes participaram, e a equipe vencedora [102] utilizou um modelo CRF com *features* morfossintáticas, distributivas e agrupamentos de palavras com base nas representações aprendidas.

Moon et al. [79] apresentaram uma nova tarefa chamada *Multimodal Named Entity Recognition* (MNER) para dados ruidosos gerados pelo usuário, como *tweets* ou legendas do Snapchat, que incluem texto curto com imagens que o acompanham. Os autores desenvolveram modelos REN baseados em caracteres BiLSTM de última geração com 1) uma rede de imagem profunda que incorpora contexto visual relevante para aumentar a informação textual e 2) um módulo genérico de atenção de modalidade que aprende a atenuar modalidades irrelevantes enquanto amplificam as mais informativas para extrair contextos. Esse modelo, nomeado por MNER, supera significativamente os modelos REN, aproveitando com sucesso os contextos visuais fornecidos.

O trabalho de Porcaro e Saggion [94] teve como principal objetivo propor um novo método para reconhecer entidades musicais em conteúdo do *Twitter* gerado por usuários que seguem um canal de rádio de música clássica. Eles utilizaram algoritmos como: SVM, uma arquitetura de rede neural recorrente, e um BiLSTM-CRF.

A Tabela 3.1 apresenta um resumo dos trabalhos relacionados [29]. É interessante notar que o W-NUT e o MSM foram os principais canais de divulgação de trabalhos relacionados ao REN em textos informais. Além disso, é perceptível que 10 dos 20 trabalhos utilizaram uma camada CRF em sua arquitetura, incluindo o trabalho que obteve o melhor desempenho para o *corpus* WeiboNER. Também é notável que a maioria dos *corpora* disponíveis é para o idioma inglês.

Trabalho	Ano	*Corpus*	Método	**F1-score**	Idioma
[97]	2011	Ritter et al.	CRF + LabeledLDA	66.00%	Inglês
[68]	2011	Liu et al.	KNN + CRF	80.20%	Inglês
[18]	2013	Basave et al.	CRF+SVM	67.00%	Inglês
[17]	2014	Basave et al.	baseado em regras	70.06%	Inglês
[98]	2015	Rizzo et al.	*Random Forest* + Regressão Logística	80.67%	Inglês
[118]	2015	W-NUT-2015	Entity Linking	56.41%	Inglês
[90]	2015	WeiboNER	CRF	56.05%	Chinês
[16]	2016	Basave et al.	KEA	54.86%	Inglês
[66]	2016	W-NUT-2016	BiLSTM	52.41%	Inglês
[34, 1]	2016	BTC	BiLSTM+CRF	74.70%	Inglês
[9]	2016	Basile et al.	BiLSTM	50.34%	Italiano
[2]	2017	W-NUT-2017	BiLSTM	41.86%	Inglês
[102]	2017	CAp2017	CRF	58.59%	Francês
[79]	2018	Moon et al.	MNER	69.10%	Inglês
[94]	2019	Porcaro et al.	BiLSTM-CRF	69.11%	-
[104]	2020	WNUT-2020	BERTneural	76.60%	Inglês
[84]	2020	W-NUT-2016 W-NUT-2017	BERTweet BERTweet	52.10% 56.50%	Inglês Inglês
[117]	2020	WeiboNER	FGN + LSTM-CRF	71.25%	Chinês
[114]	2021	W-NUT-2016	rede neural + CRF	58.98%	Inglês
[114]	2021	W-NUT-2017	rede neural + CRF	60.45%	Inglês

Tabela 3.1: *Estudos de REN em textos informais [29].*

3.1 Aplicações de REN no domínio legal

O PLN para o domínio jurídico tem seus próprios desafios únicos. Isso se deve à forma como os documentos legais são estruturados, bem como à linguagem específica do domínio que está sendo usada [95]. Embora muitos domínios de texto tenham um uso de linguagem bastante semelhante, a linguagem judicial usa semântica e interpretação de maneira única, o que a torna menos transparente para os leigos compreenderem [13].

A tecnologia que lida com documentos legais tem recebido maior atenção nos últimos anos. Isso pode ser visto pelo número de artigos científicos recentes sendo publicados, a existência dos eventos: *Natural Legal Language Processing* (NLLP), [85], *Conference on Legal Knowledge and Information Systems* (JURIX) [48], e diferentes projetos internacionais que tratam do processamento de linguagem natural para o domínio jurídico [95].

A influência das técnicas computacionais e da ciência da computação no domínio jurídico deu origem à informática jurídica, uma disciplina entre a ciência da computação e o direito, lidando com recuperação de informações jurídicas e design de aplicativos para

especialistas em direito, mas também com experiência em questões relacionadas a TI, como direitos autorais e segurança [13].

No trabalho desenvolvido por Bruckschen [13], por exemplo, é proposto uma abordagem para preencher automaticamente uma ontologia jurídica a partir de textos jurídicos através da tarefa de REN, visando a descoberta de relações semânticas. Com o objetivo final de fornecer um recurso que possa ajudar os gerentes de projeto da indústria de *software* a calcular, entender e reduzir os riscos de privacidade em seus projetos.

Badji [7] focou na identificação de pessoas jurídicas em textos em espanhol e inglês, com foco principal em referências informais a documentos legislativos encontrados em notícias, *Twitter*, contratos ou artigos de periódicos. O trabalho enquadra-se no projeto *H2020 Lynx*, que visa a criação de um *Legal Knowledge Graph* que permita a prestação de serviços relacionados com compliance.

Devemos observar que a pesquisa apresentada nessa dissertação difere dos trabalhos existentes devido a vários aspectos: (i) Estudamos um domínio legislativo em plataforma de mídia social aberta para o cidadão[6] cujo texto é informal e curto, com menos de 500 caracteres. (ii) Até o momento não foi encontrado um *corpus* para reconhecimento de entidades nomeadas em textos gerados pelo usuário (UT - *user-generated text*) para o português. (iii) Além de fornecer um *corpus* de UT para o REN, no domínio legislativo, avaliamos os benefícios de treinar um modelo de uma coleção formal de textos anotados no mesmo domínio, mas de gêneros diferente; e (iv) experimentamos modelos CRF, BiLSTM-CRF, BERT e RoBERTa, que são estado da arte para muitas tarefas de PLN [105].

[6]https ://www.camara.leg.br/enquetes/

CAPÍTULO 4

Metodologia

Na Figura 4.1 são apresentadas as etapas para realização desse trabalho.

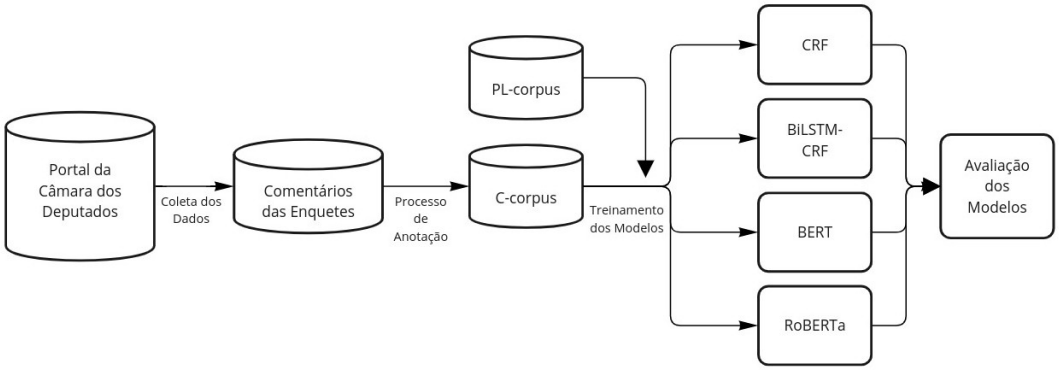


Figura 4.1: *Fluxograma de execução dos procedimentos metodológicos.*

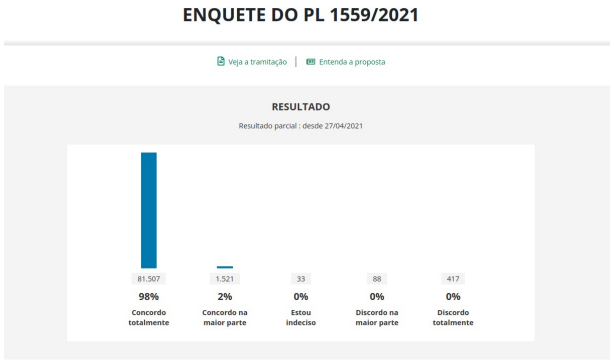
A primeira etapa consistiu na coleta dos comentários relacionados aos projetos de lei no portal da Câmara dos Deputados[1]. Na Figura 4.2 é possível observar uma votação em aberto para os cidadões sobre um determinado projeto de lei no portal da Câmara dos Deputados. A Subseção 4.1 fornece maiores detalhes dessa fase.

Após a obtenção do conjunto de dados, como ilustra a Figura 4.1, foi realizada a anotação dos mesmos, por meio da plataforma *open-source* INCEpTION [57][2], sendo consideradas neste estudo as categorias: *Pessoa, Organização, Local, Evento, Data, Fundamento e Produto de Lei*. Elaborou-se um manual de anotação para auxiliar os anotadores, também foi utilizado um documento compartilhado no *Google Drive* no qual eram colocadas dúvidas relativas à anotação. Todo o processo de anotação e a descrição das categorias das entidades estão detalhados na Subseção 4.1.

Após o processo de anotação, os dados foram exportados da plataforma INCEpTION no formato CoNLL 2002 [57]. O formato de marcação dos *tokens* usado, foi o IOB2

[1] https://www.camara.leg.br/enquetes/
[2] https://inception-project.github.io/

(a) *Votação de uma PL no Portal da Câmera dos Deputados.*



(b) *Resultado da votação da PL.*

(c) *Comentários dos cidadões sobre a PL proposta.*

Figura 4.2: *Portal da Câmera dos Deputados.*

[110], ou seja, as entidades podem receber: a *tag* **B**, no qual indica que o *token* é o início de uma entidade; a *tag* **I**, que indica que o *token* faz parte de uma entidade já iniciada; e a *tag* **O**, indicando que o *token* não pertence a nenhuma categoria das entidades que estão sendo anotadas.

De acordo com a Figura 4.1, na parte de treinamento dos modelos, utilizou-se o PL-corpus, sendo este um dos *corpus* que constitui o UlyssesNER, junto ao C-corpus. Por meio da análise dos trabalhos relacionados, foram selecionados os modelos que obtiveram melhores resultados para a tarefa de REN para obtenção dos resultados experimentais, sendo eles: CRF, BiLSTM-CRF, BERT e RoBERTa. Os detalhes dos experimentos serão apresentados na Subseção 4.2.

Após os experimentos, os modelos foram avaliados utilizando as medidas: precisão, *recall* e F1-score, para comparação da performance dos modelos.

4.1 C-corpus

O processo de anotação do C-corpus[3] foi o mesmo seguido por Albuquerque et al. [4] na criação do UlyssesNER-Br. O *corpus* UlyssesNER-Br contém sete classes ou categorias semânticas. Cinco delas, usadas no *corpus* desta dissertação, foram baseadas no HAREM [100]: *pessoa*, *local*, *organização*, *evento* e *data*. Além disso, foram definidas duas outras classes para o domínio legislativo: *fundamento*, que faz referência a entidades relacionadas a leis, resoluções, decretos, bem como a entidades de domínio específico como projetos de lei e solicitações de trabalho; e *produtos de lei*, que se refere a sistemas, programas e outros produtos criados a partir da legislação. Algumas das categorias também são divididas em tipos. A Tabela 4.1 as resumem.

Tabela 4.1: *UlyssesNER-Br: categorias e tipos [4].*

Categoria	**Tipo**	**Descrição**	**Exemplo**
DATA	—	Data	01 de janeiro de 2020
EVENTO	—	Evento	Eleições de 2018
FUNDAMENTO	FUNDlei	Norma legal	Lei no 8.666, de 21 de junho de 1993
	FUNDapelido	Apelido da norma legal	Estatuto da Pessoa com Deficiência
	FUNDprojetodelei	Projeto de lei	PEC 187/2016
	FUNDsolicitacaotrabalho	Consulta legislativa	Solicitação de Trabalho n^o 3543/2019
LOCAL	LOCALconcreto	Local concreto	Niterói-RJ
	LOCALvirtual	Local virtual	Jornal de Notícias
ORGANIZAÇÃO	ORGpartido	Partido político	PSB
	ORGgovernamental	Organização governamental	Câmara dos Deputados
	ORGnãogovernamental	Organização não governamental	Conselho Reg. de Medicina (CRM)
PESSOA	PESSOAindividual	Individual	Jorge Sampaio
	PESSOAgrupoind	Grupo de indivíduos	Família Setúbal
	PESSOAcargo	Cargo	Deputado
	PESSOAgrupocargo	Grupo cargo	Parlamentares
PRODUTO DE LEI	PRODUTOsistema	Sistema	Sistema Único de Saúde (SUS)
	PRODUTOprograma	Programa	Programa Minha Casa, Minha Vida
	PRODUTOoutros	Outros produtos	Fundo partidário

O *corpus* UlyssesNER-Br foi construído a partir de dois tipos de documentos legislativos: *projetos de lei* - PL-corpus, documentos públicos disponíveis no portal da Câmara na Web [4], e *solicitações de trabalho* - ST-corpus, documentos internos fornecidos pela equipe da Câmara dos Deputados.

Na execução dos resultados experimentais desse trabalho, as solicitações de trabalho não foram utilizadas por conterem informações classificadas como confidenciais.

[3] https://github.com/rosi-pc/UlyssesNER-Br_2.0

[4] https://www.camara.leg.br/buscaProposicoesWeb/

Para fins de legibilidade, nomearemos o *corpus* constituído pelos projeto de lei como PL-corpus.

Para construir este novo *corpus*, denominado C-corpus, foi utilizada a abordagem de [4] no processo de anotação, abordado na Subseção 4.1.2. As entidades anotadas são apresentadas na Subseção 4.1.3.

4.1.1 Composição do *Corpus*

Para a construção do C-corpus, coletou-se 285.920 comentários relacionados a 6.560 projetos de lei diponibilizados no portal da Câmara dos Deputados, do dia 01 de janeiro de 2019 até o dia 31 de dezembro de 2020. Com o intuito de se obter comentários mais longos, restringimos a população de todos os comentários disponibizados no portal da Câmara dos Deputados nesse período, para todos os comentários com 500 caracteres, sendo essa a quantidade máxima de caracteres aceita pelo site. A Figura 4.3 apresenta a quantidade de comentários anotados por projeto de lei.

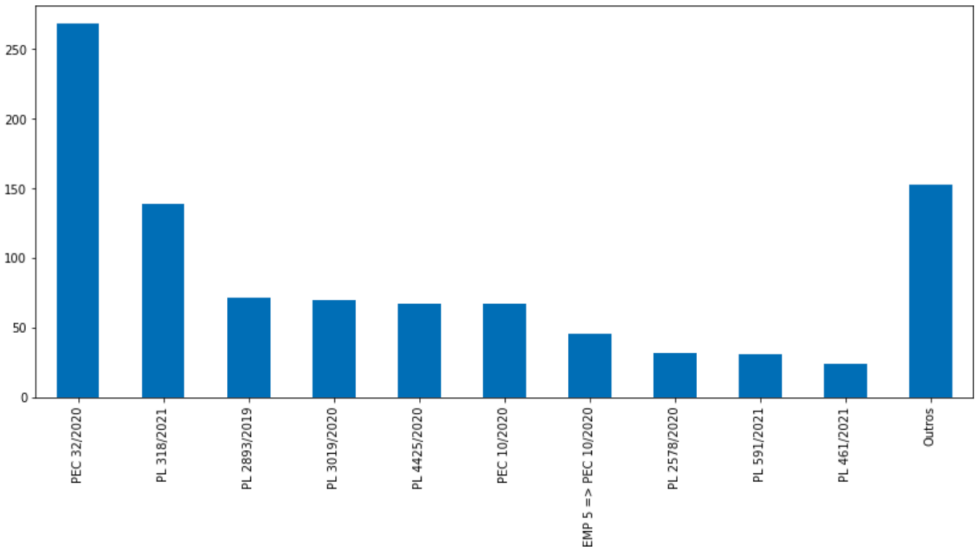


Figura 4.3: *Distribuição de frequência absoluta da quantidade de comentários anotados por projeto de lei.*

O *corpus* UlyssesNER-Br é constituído por dois tipos de documentos formais: *projetos de lei* e *solicitações de trabalhos*. Para realização desse trabalho, foram anotados textos informais, ou seja, textos gerados pelos usuários: comentários dos cidadãos relacionados aos projetos de lei PEC32/2020, PL318/2021, PL2893/2019, PL3019/2020, PL4425/2020, PEC10/ 2020, EMP5=>PEC10/2020, PL2578/2020, PL591/2021, PL461/2021 e outros[5], coletado das enquetes no portal da Câmara dos

[5] A lista dos projetos de lei relacionados aos comentários que constituem o C-corpus está no apêndice A.

Deputados, a fim de analisar se a combinação de textos formais e informais de um mesmo domínio poderia melhorar o desempenho de modelos de REN. Essas enquetes são um instrumento de participação cidadã no processo legislativo, em que o usuário pode votar se é a favor ou contra a proposição, bem como formular um comentário.

4.1.2 Processo de Anotação

Existem várias abordagens para o processo de anotação. Em geral, os anotadores humanos examinam os textos e marcam as partes relevantes que correspondem às entidades nomeadas. Eles seguem diretrizes ou esquemas pré-definidos para garantir consistência na anotação. Em alguns casos, são utilizadas ferramentas de anotação assistida por computador para agilizar o processo e melhorar a eficiência [18] .

No presente trabalho, a anotação foi realizada por três grupos de anotadores, em duas fases que foram realizadas em momentos distintos, em que cada grupo era composto por dois alunos de graduação que eram responsáveis pela anotação, enquanto um terceiro aluno de pós-graduação realizava a curadoria, sendo as divergências discutidas periodicamente.

Na primeira fase, este processo ocorreu em três etapas. Na primeira, foi realizada uma etapa de treinamento, na qual foram anotados 20 comentários por todas as equipes. Na segunda etapa, os documentos eram enviados diariamente às equipes e ao final de cada dia, era realizado o acompanhamento da medida de concordância kappa de Cohen [25], além de frequentes reuniões entre curadores e anotadores. Ao final, 300 comentários foram anotados. Por fim, na última etapa, em que foram anotados 669 comentários, as reuniões foram realizadas apenas quando solicitadas pelos anotadores ou curadores, e o kappa de Cohen foi computado apenas ao final do processo.

Ao final da primeira fase foram anotados 969 comentários. A Tabela 4.2 mostra a quantidade de *tokens* e *tokens* únicos das entidades por categoria e tipo, no C-corpus, na fase 1.

Ao final do processo de anotação, foram realizados experimentos com o conjunto de dados, no qual foi possível notar um baixo desempenho dos modelos para categorias mais raras, portanto, conduziu-se a segunda fase do processo de anotação, similar ao primeiro, ao final da segunda fase foram anotados 300 comentátios, totalizando assim, 1.269 comentários anotados. Na tabela 4.3 é possível analisar a quantidade de *tokens* para cada categoria e tipo no *corpus* final.

A anotação foi realizada utilizando a ferramenta INCEpTION [57][6], que fornece um ambiente para muitas tarefas de anotação em texto escrito. A Figura 4.4 mostra

[6]https://inception-project.github.io/

Tabela 4.2: *C-corpus: Quantidade de tokens e tokens únicos por categorias e tipos [4].*

Categoria	**Tipo**	**Quantidade de *tokens***	**Quantidade de *tokens* únicos**
DATA	—	45	34
EVENTO	—	154	72
FUNDAMENTO	FUNDlei	209	86
	FUNDapelido	180	72
	FUNDprojetodelei	21	12
LOCAL	LOCALconcreto	161	67
	LOCALvirtual	16	15
ORGANIZAÇÃO	ORGpartido	12	9
	ORGgovernamental	376	132
	ORGnãogovernamental	49	38
PESSOA	PESSOAindividual	134	101
	PESSOAgrupoind	7	7
	PESSOAcargo	272	137
	PESSOAgrupocargo	649	200
PRODUTO DE LEI	PRODUTOsistema	10	6
	PRODUTOprograma	5	5
	PRODUTOoutros	609	167

exemplos das entidades anotadas nos comentários. A primeira, segunda e terceira equipe alcançaram o seguinte coeficiente kappa de Cohen: 65%, 87% e 85%, respectivamente, na segunda etapa da fase 1. Na terceira etapa da fase 1 alcançaram 79%, 86% e 83%, em ordem. Já na segunda fase, as equipes alcançaram 72%, 73% e 80%, de maneira respectiva.

Tabela 4.3: *C-corpus: Quantidade de tokens e tokens únicos por categorias e tipos no corpus final.*

Categoria	Tipo	Quantidade de *tokens*	Quantidade de *tokens* únicos
DATA	—	140	62
EVENTO	—	244	105
FUNDAMENTO	FUNDlei	611	194
	FUNDapelido	328	111
	FUNDprojetodelei	306	87
LOCAL	LOCALconcreto	325	147
	LOCALvirtual	84	59
ORGANIZAÇÃO	ORGpartido	51	25
	ORGgovernamental	615	198
	ORGnãogovernamental	95	74
PESSOA	PESSOAindividual	249	177
	PESSOAgrupoind	12	11
	PESSOAcargo	370	163
	PESSOAgrupocargo	813	237
PRODUTO DE LEI	PRODUTOsistema	52	16
	PRODUTOprograma	94	56
	PRODUTOoutros	709	207

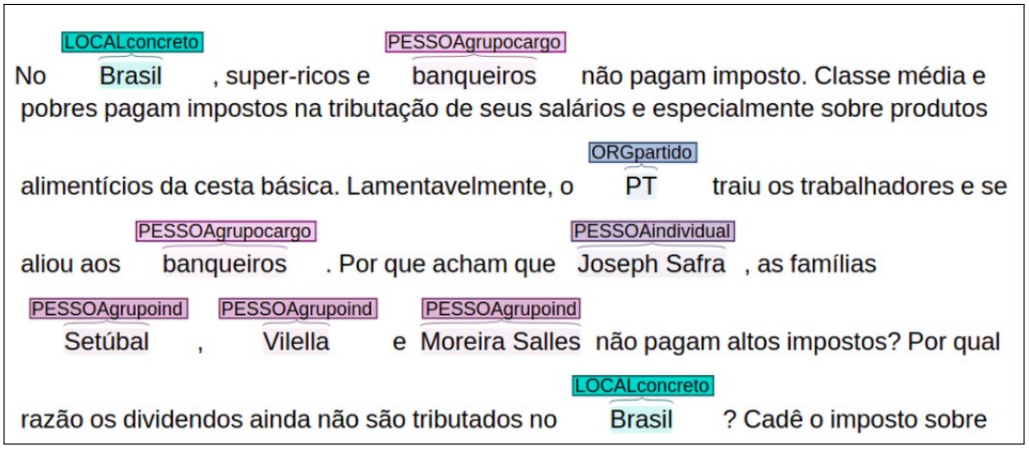


Figura 4.4: *Processo de anotação de entidades usando INCEpTION.*

4.1.3 Categorias das Entidades Anotadas

As entidades anotadas na construção do C-corpus, foram as mesmas que Albuquerque et al. [4]: *pessoa*, *localização*, *organização*, *evento*, *data*, *Fundamento* e *Produto de Lei*. O tipo *FUNDsolicitacaotrabalho* que está enquadrada na categoria *Fundamento*, como é possível observar na Tabela 4.1, não foi anotada no C-corpus, dado que o *corpus* não possui entidades como essa.

Pessoa

Essa categoria enquadra os tipos: *Individual*, *Cargo*, *Grupo Cargo* e *Grupo Indivíduos*.

- ***Individual***: São incluídos nessa categoria nome de pessoas, apelidos, nomes diminutivos, alcunhas ou iniciais.
  Ex.: "...Quer melhorar o pais, apoia o projeto do <*PESSOA TIPO=INDIVIDUAL*> **Sergio Moro** <*PESSOA*> que e acabar com a corrupção,...".
- ***Cargo***: Um posto que é ocupado por uma pessoa, mas que poderá no futuro ser ocupado por outro indivíduo.
  Ex.: "...retirada do <*PESSOA TIPO = CARGO*> **presidente da República** <*PESSOA*> e das categorias jurídicas...".
- ***Grupo cargo***: Abrange todos os termos que implicitamente se referem a um conjunto de pessoas, referidas pelo cargo.
  Ex.: "...reduzir em metade o número de <*PESSOA TIPO=GRUPOCARGO*> **Deputados** <*PESSOA*> e <*PESSOA TIPO=GRUPOCARGO*> **Senadores** <PESSOA>, com a consequente redução...".
- ***Grupo Indivíduos***: Abrange todos os termos que implicitamente se referem a um conjunto de pessoas, referidas pelo nome.
  Ex.: "...as famílias <*PESSOA TIPO=GRUPOIND*>**Setúbal**<*PESSOA*>, <*PESSOA TIPO=GRUPOIND*>**Vilella**<*PESSOA*> e <*PESSOA TIPO=GRUPOIND*>**Moreira Salles**<*PESSOA*> não pagam altos impostos?...".

Local

Essa categoria identifica localizações que foram criadas pelo Homem. Inclui países, bairros, regiões geopolíticas (Rio de Janeiro, Alentejo, bairro dos Anjos, Ásia Menor, Jardim das Amoreiras, Oriente Médio, América Latina, África, Países de Leste). Com exceção do termo "Distrito Federal" quando este refere-se a uma organização político-administrativa, e não a um local concreto.

Ex.: "...aqui na minha cidade de <*LOCAL TIPO=CONCRETO*> **Niterói-RJ** <*LOCAL*>, por determinação de decreto municipal, este fechamento...".

Organização

Essa categoria enquadra os tipos: *Governamental*, *Privado* e *Partido Político*.

- ***Governamental***: Nessa categoria foram classificados todos os órgãos possíveis e existentes no ambiente do governo em geral. Elas podem ser autarquias, ministérios, tribunais, universidades e institutos federais, secretarias de estado, empresas públicas (Ex.: Secretaria de Estado da Cultura, Brasil, Prefeitura de São Paulo, Câmara Municipal de Leiria) entre outras repartições que possam vir ser classificadas como Organização. Termos como: "Federação", "União", "Estados", "Municípios",

"Distrito Federal", "País" quando apareceram sozinhos ou sem um contexto específico, não foram anotados.
Ex.: "...programa de excelência com nota 5 na <*ORGANIZACAO TIPO=GOVERNAMENTAL*> **Capes** <*ORGANIZACAO*> e venho por meio desta pedir...".

- ***Não Governamental***: Para organizações com ou sem fins lucrativos, do setor privado, como empresas, sociedades, clubes (Boavista FC, Círculo de Leitores, Livraria Barata, (discoteca) Kapital) .
  Ex.: "...a <*ORGANIZACAO TIPO=PRIVADO*> **FEDEX** <*ORGANIZACAO*>, <*ORGANIZACAO TIPO=PRIVADO*> **UPS** <*ORGANIZACAO*>, <*ORGANIZACAO TIPO=PRIVADO*> **DHL** <*ORGANIZACAO*> entre outras empresa de encomendas...".
- ***Partido Político***: Grupos organizados, legalmente formados, com base em formas voluntárias de participação numa associação orientada para ocupar o poder político, como PSDB, PT, entre outros.
  Ex.: "...deve ser porque foi o <*ORGANIZACAO TIPO=PARTIDO POLÍTICO*>**PSOL**<*ORGANIZACAO*> que fez, sou de direita...".

Data

Todas as referências a dias, mês e ano. Referências a mês e ano, ou só a ano também devem ser etiquetadas.

Ex.: "...Enquanto isso, os bancos que tiveram lucro na ordem de 60 bilhões em <*DATA*>**2019**<*DATA*>, não colaboram com nenhum centavo...".

Evento

Essa categoria refere-se aos acontecimentos pontuais, organizados ou não (Pandemia do Covid-19, Greve dos Caminhoneiros).

Ex.: "...Deve-se buscar outras maneiras de solucionar a <*EVENTO*>**crise gerada pela pandemia do coronavírus**<*EVENTO*>...".

Fundamento

Nessa categoria se enquadram os tipos: *Lei*, *Projeto de Lei* e *Apelido de Lei*.

- ***Lei***: Nessa classe, se enquadram as leis, resoluções, portarias, decretos, medidas provisórias, nos quais foram referidas pelo o número, sendo utilizadas pelos usuários para fundamentar seus argumentos.

Ex.: "...As propostas visam à redução salarial dos trabalhadores, o que fere o princípio constitucional da irredutibilidade salarial previsto no <*FUNDAMENTO TIPO = LEI*>**art. 37, XV da CF/88**<*FUNDAMENTO*>...".

- ***Projeto de Lei***: Foi etiquetado menções a projetos de lei referidos pelo número, que foram citados pelos usuários.
  Ex.: "...votem a favor deste <*FUNDAMENTO TIPO = PROJETODELEI*> **pl 1263/2020** <*FUNDAMENTO*>...".
- ***Apelido de Lei***: Se enquadra nessa categoria apelido de normas (leis, projetos de leis, PECs) que foram citados por usuários.
  Ex.: "...Fomos prejudicados com a <*FUNDAMENTO TIPO=FUNDapelidodelei*> **reforma da previdência** <*FUNDAMENTO*>, inexistindo praticamente regras de transição...".

Produto de Lei

Nessa categoria se enquadram os tipos: *Sistema*, *Programa* e *Outros*.

- ***Sistema***: Esta classe incorpora os Sistemas criados pelo governo, sociais ou não, tais como SUS (Sistema Único de Saúde), Sistema Único de Assistência Social (SUAS), entre outros.
  Ex.: "...mas é extremamente importante para que o cidadão reconheça o <*PRODUTODELEI TIPO= SISTEMA*> **SUS** <*PRODUTODELEI*>...".
- ***Programa***: Esta classe incorpora os Programas criados pelo governo, sociais ou não, tais como programa Minha Casa Minha Vida (Casa Verde e Amarela), Bolsa Família, entre outros.
  Ex.: "...Deus envia anjos pra nós da algo mais tem vezes que não temos NAO TENHO cargo <*PRODUTODELEI TIPO=PROGRAMA*> **BOLSA FAMÍLIA** <*PRODUTODELEI*>...".
- ***Outros***: Benefícios criados pelo governo para atender tanto a população (seguro desemprego, FGTS, aposentadoria...), como também organizações específicas (os partidos políticos possuem o Fundo partidário, Fundo eleitoral, etc), ou até mesmo indivíduos pertencentes a uma organização específica (auxílio moradia, auxílio alimentação, auxílio paletó para deputados e senadores, entre outros).
  Ex.: "...O mínimo é que os recursos do <*PRODUTODELEI TIPO: OUTROS*> **Fundo Partidário** <*PRODUTODELEI*> e do <*PRODUTODELEI TIPO OUTROS*> **Fundo Especial de Financiamento de Campanhas** <*PRODUTODELEI*> sejam utilizadas no enfrentamento de emergências de saúde pública, de calamidade pública ou de desastres naturais...".

4.2 Configurações Experimentais

Os experimentos foram divididos em quatro partes: Na primeira, foi utilizado o modelo CRF. Na segunda parte, usou-se o modelo BiLSTM-CRF e, em sequência, o *fine-tuning* do BERT e RoBERTa. Posteriormente, avaliou-se as hipóteses: 1^a: treinar modelos de REN usando textos formais e informais aumenta a qualidade das previsões para o conjunto de textos informais no domínio legislativo, e 2^a: Modelos baseados em *transformers* possuem melhor desempenho para sistemas REN no domínio legislativo.

As seguintes *features* foram utilizadas para os experimentos com o modelo CRF, cuja inspiração foi o trabalho de Amaral e Vieira [5]. *1) Palavras de contexto:* Refere-se às palavras que cercam a palavra atual. Neste trabalho, foi considerada a palavra anterior e a seguinte. *2) Palavra minúscula*: a palavra atual em minúscula. *3) Part of speech (POS) tag*: Retorna a classe gramatical da palavra atual. *4) Primeiro caractere maiúsculo*: Indica se o primeiro caractere da palavra atual está em maiúsculo. *5) Dígito*: Indica se todos os caracteres são dígitos. *6) Prefixo e sufixo*: Foram considerados os sufixos e prefixos da palavra atual, até o caractere de terceira posição. *7) Primeira palavra*: Indica se a palavra atual é a primeira palavra da frase. *8) Última palavra*: Indica se a palavra atual é a última palavra da frase. Foram também adicionadas as *features* 2 a 6 da palavra anterior e a palavra posterior da atual.

Nos experimentos com BiLSTM-CRF, foi utilizada uma implementação *open-source* [42], sem alterar os hiperparâmetros propostos pelo autor. Usou-se os *embeddings* GloVe com 300 dimensões fornecidos por [30, 50].

O modelo BERTimbau Base [105] foi utilizado para realizar um pré-treinamento adicional com as sentenças não rotuladas do conjunto de comentários sobre os projetos de lei. A partir desse modelo especializado, realizamos o *fine-tuning* para a tarefa de REN. Os hiperparâmetros para ajustar o modelo BERT para REN foram os mesmos usados por [46], sendo o *batch size* igual a 4, a taxa de aprendizado igual a 2e-5, e 3 épocas, dado que ocorre *overfitting* para números mais altos. Usamos o PyTorch [88] como uma estrutura de aprendizado profundo, o *tokenizer* do modelo pré-treinado, e *optimizer* do *HuggingFace* [116].

O modelo BERTimbau Base foi pré-treinado com o BrWaC [112], o maior *corpus* aberto em português até o momento. Além do seu tamanho, o BrWaC é composto por documentos completos e sua metodologia garante alta diversidade de domínio e qualidade de conteúdo, que são características desejáveis para o pré-treinamento do BERT [105]. O *corpus* final processado possui 17.5 GB de texto bruto.

O modelo RobertaTwitterBR [83] também foi utilizado para realizar um pré-treinamento adicional com as sentenças não rotuladas do conjunto de comentários sobre os projetos de lei, e posteriormente foi realizado o *fine-tuning* para a tarefa de REN. Os

hiperparâmetros para ajustar o modelo RoBERTa para REN foram os mesmos usados no treinamento do BERT. Esse modelo foi pré-treinado com aproximadamente 7 milhões de *tweets*, usando os *corpora Twitter* NPS[7], TweetSentBR [14], PELESent [55], Portuguese *Tweets* for Sentiment Analysis (Kaggle)[8] e *Twitter* e Notícias sobre COVID-19 em português [32]. O *corpora* contêm 126.285.329 *tokens*, em 10 GB de texto [82].

Nos experimentos com BiLSTM-CRF, para o *fine-tuning* do BERT e do RoBERTa, consideramos uma divisão de treinamento-teste-validação de 70%-15%-15%, as sentenças foram selecionadas aleatoriamente para cada tipo com conjunto não sobreposto, e para CRF, consideramos conjuntos de treinamento+validação em fase de treinamento e teste no mesmo cenário. Usamos a implementação CRF do Scikit-learn [89].

Os experimentos realizados com a BiLSTM-CRF foram repetidos 5 vezes, o que significa que os dados de precisão, *recall* e F1-score foram coletados em 5 execuções separadas do experimento e, em seguida, a média e o desvio padrão foram calculados a partir desses resultados.

Enquanto o *fine-tuning* do BERT e do RoBERTa foi repetido 15 vezes para testar se a diferença no desempenho e o tempo de treinamento entre os modelos é estatisticamente significativa. Os resultados de precisão, *recall* e F1-score foram coletados em cada uma dessas 15 execuções do experimento e, em seguida, a média e o desvio padrão foram calculados a partir desses resultados.

[7] https://github.com/verissimomanoel/twitter_nps

[8] https://www.kaggle.com/datasets/augustop/portuguese-tweets-for-sentiment-analysis

CAPÍTULO 5

Resultados Experimentais

5.1 Resultados

Para testar a hipótese de que textos formais auxiliam em melhores predições para textos informais, um modelo foi treinado para cada forma possível de junção dos conjuntos de treinamento e seu desempenho foi avaliado separadamente no conjunto de teste. Os resultados de cada variação dos modelos CRF, BiLSTM-CRF e o *fine-tuning* do BERT e RoBERTa em cada conjunto de teste são apresentados na Tabela 5.1.

O modelo que obteve o melhor desempenho entre os quatro modelos de REN testados foi o *fine-tuning* do BERT, quando utilizado a junção do PL-corpus ao C-corpus como conjunto de treinamento, como pode ser visto na Tabela 5.1. É importante ressaltar que apenas o BERT foi capaz de capturar as relações semânticas entre os textos formais e informais, de forma a obter ganhos no desempenho do modelo quando a junção dos textos foi utilizada no treinamento para os diferentes conjuntos de teste.

A Tabela 5.2 apresenta o *valor de p* do teste de Wilcoxon [115] para verificar a diferença entre o desempenho dos modelos BERT e RoBERTa nas diferentes configurações do conjunto de treinamento e teste, considerando o F1-score. Observa-se que existe uma diferença estatisticamente significativa entre o desempenho dos modelos, exceto quando o C-corpus e o PL-corpus, a nível categoria, pertenciam ao conjunto de treinamento e teste, respectivamente, considerando o nível de significância de 5%.

Ainda na Tabela 5.1, é interessante notar que o conjunto de teste PL-corpus, obteve um desempenho superior, se comparado ao C-corpus. Isso se deve principalmente ao fato de que o PL-corpus possui uma linguagem formal e o número de sentenças disponíveis para treinamento do modelo é muito maior. A presença de uma linguagem menos formal no C-corpus, prejudicou significativamente o aprendizado das relações semânticas entre entidades pelos algoritmos de REN.

Tabela 5.1: *Resultados com agregação de conjuntos de dados para treinar os modelos CRF, BiLSTM-CRF+GloVe, o fine-tuning do BERT e do RoBERTa. P foi usado para Precisão, R para Recall e F1 para F1-score. O PL-corpus foi utilizado para projetos de lei e o C-corpus para nosso corpus. O termo "All"significa a junção de PL-corpus e C-corpus no conjunto de treinamento. Os melhores resultados estão em negrito.*

Resultados com a agregação do conjunto de dados para treinamento do modelo CRF

Nível	Conj. Treinamento	Conj. Teste	P	R	F1	Nível	Conj. Treinamento	Conj. Teste	P	R	F1
Categoria	PL-corpus	PL-corpus	84.06	80.97	82.48	Tipo	PL-corpus	PL-corpus	82.78	79.47	81.09
	C-corpus		41.09	37.73	39.34		C-corpus		38.71	34.06	36.23
	All[1]		82.41	79.80	81.09		All		80.52	77.30	78.88
	PL-corpus	C-corpus	60.47	32.58	42.35		PL-corpus	C-corpus	59.24	27.32	37.39
	C-corpus		74.23	60.65	66.76		C-corpus		71.69	59.65	65.12
	All		71.93	61.65	66.40		All		73.35	61.40	66.85

Resultados com a agregação do conjunto de dados para treinamento do BiLSTM-CRF+GloVe

Nível	Conj. Treinamento	Conj. Teste	P	R	F1	Nível	Conj. Treinamento	Conj. Teste	P	R	F1
Categoria	PL-corpus	PL-corpus	85.1±0.12	78.16±0.12	81.48±0.28	Tipo	PL-corpus	PL-corpus	**85.24±1.17**	74.65±4.83	79.57±3.26
	C-corpus		20.63±1.63	19.20±0.81	19.51±0.70		C-corpus		19.32±1.58	18.60±1.39	19.01±0.33
	All		80.12±1.35	77.37±1.95	79.03±1.53		All		79.36±0.79	75.2±0.34	77.97±0.64
	PL-corpus	C-corpus	36.49±2.63	14.34±0.64	19.54±1.54		PL-corpus	C-corpus	35.99±0.89	13.66±1.64	18.82±1.14
	C-corpus		66.12±3.21	25.82±1.87	35.7±1.20		C-corpus		71.5±1.59	23.47±1.20	35.67±2.43
	All		71.63±2.25	39.10±1.47	50.33±1.78		All		**75.74±1.87**	41.39±2.55	53.70±1.25

Resultados com a agregação do conjunto de dados para realizar o *fine-tuning* do BERT

Nível	Conj. Treinamento	Conj. Teste	P	R	F1	Nível	Conj. Treinamento	Conj. Teste	P	R	F1
Categoria	PL-corpus	PL-corpus	84.63±1.16	89.63±0.51	86.94±0.61	Tipo	PL-corpus	PL-corpus	83.04±0.47	86.89±0.27	84.74±0.42
	C-corpus		46.28±1.11	59.13±0.28	51.52±0.94		C-corpus		42.75±1.37	55.80±0.64	48.22 ±0.92
	All		**85.75±0.46**	**89.50±0.15**	**87.44±0.35**		All		82.11±1.46	**87.56±1.47**	**85.95±1.65**
	PL-corpus	C-corpus	68.55±2.41	67.89±0.75	66.42±1.05		PL-corpus	C-corpus	62.43±2.59	58.08±0.32	59.34±1.24
	C-corpus		76.84±0.42	80.06 ±0.68	77.61±0.28		C-corpus		71.68±1.59	76.88±0.43	73.53±0.79
	All		**77.69±0.30**	**81.64±0.60**	**78.65±0.39**		All		71.90±0.35	**76.69±0.58**	**73.60±0.39**

Resultados com a agregação do conjunto de dados para realizar o *fine-tuning* do RoBERTa

Nível	Conj. Treinamento	Conj. Teste	P	R	F1	Nível	Conj. Treinamento	Conj. Teste	P	R	F1
Categoria	PL-corpus	PL-corpus	72.93 ±1.41	82.07 ±0.79	77.08 ±0.87	Tipo	PL-corpus	PL-corpus	69.51 ±0.49	80.33 ±0.30	74.19 ±0.29
	C-corpus		47.55 ±1.99	60.28 ±0.83	52.98 ±1.36		C-corpus		18.55 ±0.57	32.05 ±0.96	23.26 ±0.70
	All		70.05 ±0.66	81.33 ±0.41	75.01 ±0.52		All		65.12 ±0.73	77.96 ±0.54	70.86±0.67
	PL-corpus	C-corpus	46.27±1.51	44.98±1.96	43.74±0.89		PL-corpus	C-corpus	38.48±0.64	36.32±0.76	36.76±0.84
	C-corpus		62.28±0.70	66.55±1.48	63.74±0.90		C-corpus		56.62±0.96	61.88±1.12	58.8±1.02
	All		63.51±1.5	71.93±0.80	67.06±1.23		All		60.47±1.95	66.37±0.48	62.95±0.45

Tabela 5.2: *Teste de Wilcoxon (valor de p) do F1-score entre o modelo BERT e RoBERTa, para 15 execuções. O termo "All"significa a junção de PL-corpus e C-corpus no conjunto de treinamento.*

Nível	Conjunto de Treinamento	Conjunto de Teste	Teste de Wilcoxon (*valor de p*)
Categoria	PL-corpus	PL-corpus	6.1e-05*[1]
	C-corpus		0.2077
	All[1]		6.1e-05*
	PL-corpus	C-corpus	6.1e-05*
	C-corpus		6.1e-05*
	All		6.1e-05*
Tipo	PL-corpus	PL-corpus	6.1e-05*
	C-corpus		6.1e-05*
	All		6.1e-05*
	PL-corpus	C-corpus	6.1e-05*
	C-corpus		6.1e-05*
	All		6.1e-05*

[1] O * foi adicionado quando rejeita-se a hipótese nula de igualdade das distribuições, ao nível de significância de 5%.

Observou-se que os modelos, quando treinados apenas com o C-corpus, obtiveram predições ruins para o PL-corpus. O baixo desempenho foi supostamente devido ao vocabulário informal. É importante ressaltar que o *fine-tuning* do BERT treinado com a junção do C-corpus com o PL-corpus foi o único modelo que relatou ganhos nas predições das *tags* do PL-corpus.

Apesar do desempenho dos modelos CRF e BiLSTM-CRF ter caído ao serem treinados na junção dos *corpora* C-corpus e PL-corpus, para prever as *tags* do PL-corpus, o F1-score não apresentou diferenças significativas. A diferença média máxima foi observada no modelo BiLSTM-CRF, em termos de categoria. Quando treinado apenas com exemplos do PL-corpus, a média do F1-score foi de 81.48%. Ao adicionar exemplos do C-corpus no treinamento, o F1-score médio caiu para 79.03%.

Os modelos treinados apenas com o PL-corpus obtiveram previsões ruins para o C-corpus, o que pode ser explicado pelos diferentes exemplos de entidades encontradas no PL-corpus. No entanto, o *fine-tuning* do BERT treinado com a junção do C-corpus e do PL-corpus obteve uma melhora considerável nas predições do C-corpus, obtendo um F1-score de 78.65% para o nível categoria.

Ao analisar a Tabela 5.1 e 5.2, percebe-se que o *fine-tuning* do RoBERTa apresentou resultados inferiores aos do BERT. Essa diferença pode ser explicada pelo fato de que o modelo RoBERTa utilizado no presente trabalho foi pré-treinado em um *corpus* com tamanho de 10 GB, enquanto o BERT foi pré-treinado com um *corpus* maior, de 17.5 GB. Além disso, a diversidade desses *corpora* também pode ter influenciado, visto que o RoBERTa foi treinado apenas com *tweets*, enquanto o BERT foi pré-treinado com um *corpus* mais diverso em termos de domínio e qualidade de conteúdo.

Na Figura 5.1, é possível analisar a nuvem de palavras para o C-corpus. É

possível observar termos como *lei*, *estabilidade* e *direito* como palavras pertencentes ao domínio legislativo, sendo mais frequentes, e em menor frequência, aparecem palavras como *estado*, *governo*, *servidor*, *político*, *PL*, *PEC*, entre outras. Isso contribui com a hipótese de que o domínio do *C-corpus* pode ter afetado o desempenho do modelo RoBERTa.

Figura 5.1: *Nuvem de palavras do C-corpus.*

É importante destacar que, conforme apresentado na Tabela 5.1, a junção do PL-corpus ao C-corpus para as predições das *tags* do C-corpus contribuíram para melhores predições em todos os modelos utilizados nesse trabalho. Exceto, para o modelo CRF, quando modelado o *corpus* a nível categoria, no qual observa-se uma queda de 0.36% no F1-score quando o PL-corpus é adicionado no conjunto de treinamento.

5.1.1 Tempo para o treinamento dos modelos BERT e RoBERTa

Com o objetivo de comparar o tempo médio necessário para realizar o treinamento (*fine-tuning*) nos modelos BERT e RoBERTa, foram executados 15 execuções para cada possível combinação de conjuntos de treinamento e teste. Uma vez que foram observados indícios que indicam a não normalidade dos dados (Figura 5.2, Figura 5.3 e Figura 5.4), utilizou-se o Teste de Wilcoxon para avaliar se existe diferença significativa entre os tempos de *fine-tuning* dos dois modelos. Os resultados obtidos, incluindo a média ($\bar{x}$) e o desvio padrão amostral (s) dos tempos de treinamento para ambos os modelos, bem como o *valor de p* do teste de Wilcoxon, são apresentados na Tabela 5.3.

Com base nos dados apresentados na Tabela 5.3, é possível perceber que o tempo de *fine-tuning* dos modelos BERT e RoBERTa é maior quando aplicados ao PL-corpus em comparação ao C-corpus, em razão do tamanho do *corpora*.

Também observou-se que o tempo para o treinamento do BERT é menor em todas as combinações de treino e teste, quando é utilizado o C-corpus no conjunto de treinamento. Quando analisado o tempo para o treinamento dos modelos com o PL-corpus

Tabela 5.3: *Tempo de fine-tuning entre o modelo BERT e RoBERTa (em segundos), em que $\bar{\boldsymbol{x}}$ e $\boldsymbol{s}$ representam a média amostral, e o desvio padrão amostral, respectivamente, para 15 execuções. O termo "All"significa a junção de PL-corpus e C-corpus no conjunto de treinamento.*

Nível	Conjunto de Treinamento	Conjunto de Teste	Tempo de treinamento BERT ($\bar{x} \pm s$)	Tempo de treinamento RoBERTa ($\bar{x} \pm s$)	Teste de Wilcoxon (*valor de p*)
Categoria	PL-corpus	PL-corpus	611.20 ±22.01	591.73 ±16.71	0.0215*[1]
	C-corpus		102.20 ±10.48	111.73 ±11.17	0.0215*
	All[1]		648.53 ±34.77	577.4 ±30.84	0.0001*
	PL-corpus	C-corpus	604.27 ±54.04	597.13 ±31.58	0.2523
	C-corpus		111.53 ±9.20	118.73 ±8.21	0.0102*
	All		594.73 ±7.64	543.13 ±26.63	6.1e-05*
Tipo	PL-corpus	PL-corpus	600.4 ±9.825	583.2 ±29.46	0.0051*
	C-corpus		111.0 ±11.56	118.01 ±8.67	0.0255*
	All		620.4 ±51.62	570.66 ±42.49	0.0023*
	PL-corpus	C-corpus	645.33 ±52.06	558.73 ±54.66	0.0006*
	C-corpus		107.93 ±10.33	118.47±3.83	0.0012*
	All		600.26 ±5.28	599.93 ±10.12	0.5994

[1] O * foi adicionado quando rejeita-se a hipótese nula de igualdade das distribuições, ao nível de significância de 5%.

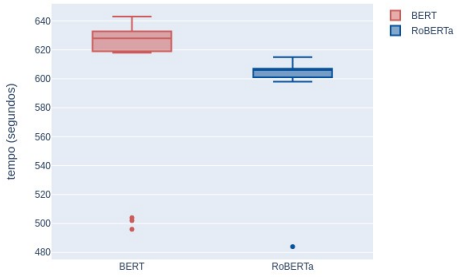
no conjunto de treinamento, observa-se que o RoBERTa possui o menor tempo para o treinamento.

A Tabela 5.3 apresenta o *valor de p* do teste de Wilcoxon, para testar a diferença de tempo para o treinamento dos dois modelos, no qual observa-se uma diferença estatisticamente significativa em quase todos os casos, exceto quando o PL-corpus pertencia ao conjunto de treinamento e o C-corpus ao conjunto de teste, e quando a junção dos *corpus* foi utilizada no treinamento e o C-corpus pertencia ao conjunto de teste.

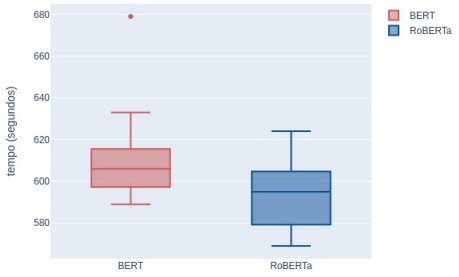
Na Figura 5.2, observa-se a distribuição do tempo de treinamento do BERT e RoBERTa com o PL-corpus no conjunto de treinamento. Em todos as figuras, é possível observar que o BERT demanda um maior tempo para realizar o *fine-tuning* dos modelos. Embora o teste de Wilcoxon tenha rejeitado a hipótese nula de que as médias são iguais quando o C-corpus a nível categoria estava no conjunto de teste (Figura 5.2(a)), é importante observar que essa conclusão pode ter sido influenciada pela presença de *outliers*. De fato, ao visualizar graficamente os dados, é possível notar que há valores extremos que podem estar distorcendo a análise.

A Figura 5.3 apresenta a distribuição do tempo para realizar o *fine-tuning* do BERT e do RoBERTa com o C-corpus no conjunto de treinamento. É possível observar que o RoBERTa requer um tempo significativamente maior para realizar o processo de *fine-tuning* dos modelos, em comparação com o BERT. Além disso, é importante destacar que o tempo de treinamento do BERT apresentou uma maior variabilidade em relação ao RoBERTa, com exceção do caso em que o C-corpus a nível categoria estava no conjunto de teste.

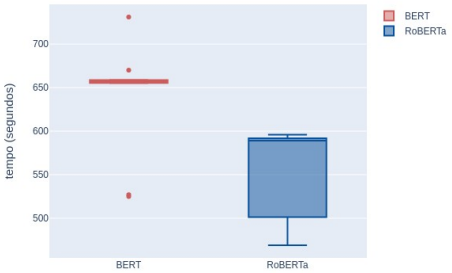
A Figura 5.4 apresenta a distribuição do tempo de treinamento do BERT e do RoBERTa com a adição do PL-corpus ao C-corpus no conjunto de treinamento. É possível observar que o BERT requer um tempo significativamente maior para realizar o processo



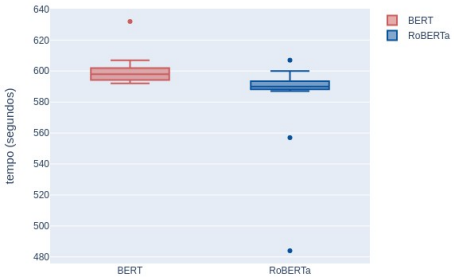
(a) *Conjunto Teste: C-corpus - Categoria*



(b) *Conjunto Teste: PL-corpus - Categoria*

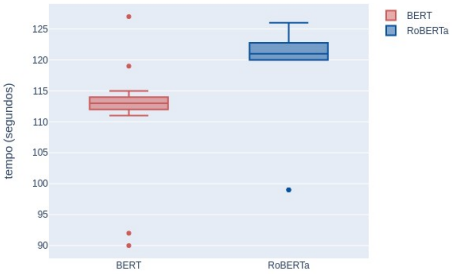


(c) *Conjunto Teste: C-corpus - Tipo*

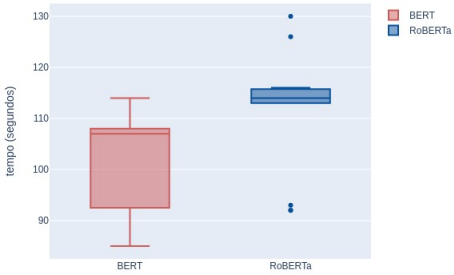


(d) *Conjunto Teste: PL-corpus - Tipo*

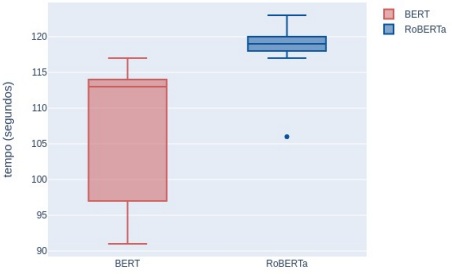
Figura 5.2: *Distribuição do tempo de fine-tuning do BERT e RoBERTa com o PL-corpus no conjunto de treinamento.*



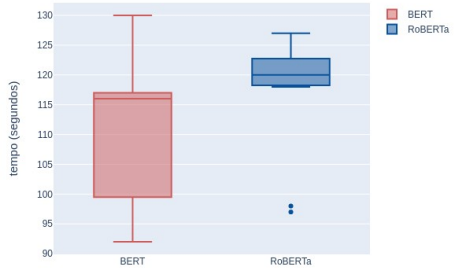
(a) *Conjunto Teste: C-corpus - Categoria*



(b) *Conjunto Teste: PL-corpus - Categoria*



(c) *Conjunto Teste: C-corpus - Tipo*

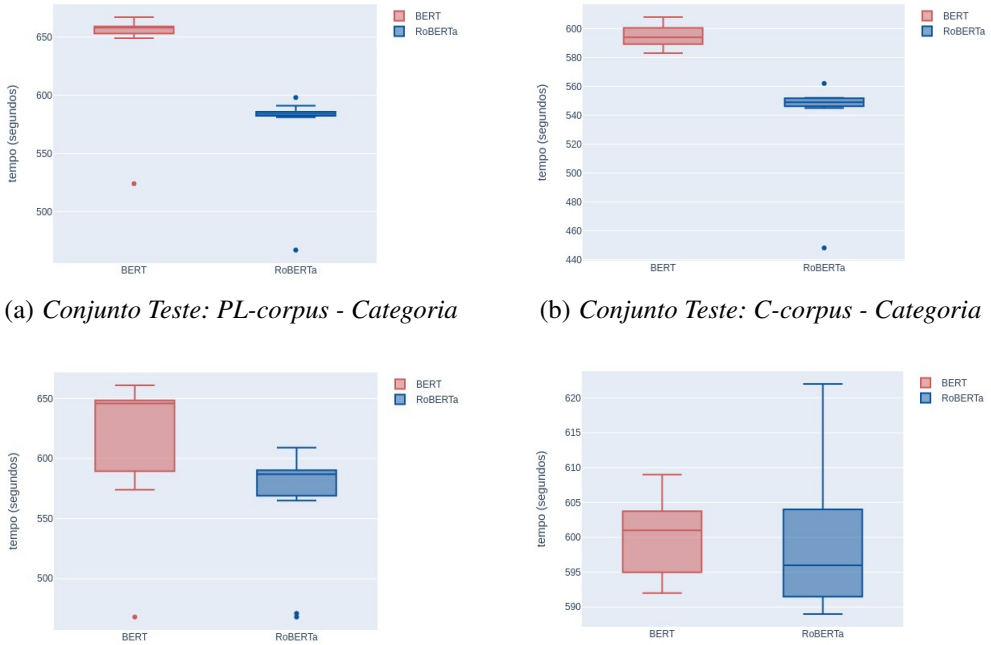


(d) *Conjunto Teste: PL-corpus - Tipo*

Figura 5.3: *Distribuição do tempo de fine-tuning do BERT e RoBERTa com o C-corpus no conjunto de treinamento.*

de *fine-tuning* dos modelos, em comparação com o RoBERTa, com exceção do caso em que o C-corpus a nível tipo estava no conjunto de teste.

É importante ressaltar que o tempo de treinamento dos modelos de processa-



(a) *Conjunto Teste: PL-corpus - Categoria* (b) *Conjunto Teste: C-corpus - Categoria*

(c) *Conjunto Teste: PL-corpus - Tipo* (d) *Conjunto Teste: C-corpus - Tipo*

Figura 5.4: *Distribuição do tempo de fine-tuning do BERT e RoBERTa com o PL-corpus + C-corpus no conjunto de treinamento.*

mento de linguagem natural pode variar significativamente dependendo de diversos fatores, como o tamanho do conjunto de treinamento, a complexidade do modelo e a capacidade de processamento da máquina utilizada. Portanto, antes da escolha de um modelo específico, é necessário analisar cuidadosamente as necessidades do projeto e os recursos disponíveis, a fim de selecionar o modelo que melhor atenda às necessidades específicas da tarefa em questão.

Além disso, é importante destacar que, embora a arquitetura do RoBERTa seja mais complexa do que a do BERT, esse modelo pode apresentar um melhor desempenho em tarefas de processamento de linguagem natural, devido à sua capacidade de lidar com informações mais complexas. No entanto, no presente estudo, a superioridade do RoBERTa em relação ao BERT não foi observada, possivelmente devido à diferença no tamanho e na diversidade dos *corpora* utilizados para pré-treinar os modelos.

5.2 Discussão

Com base nos resultados encontrados, podemos concluir que o treinamento do modelo REN com frases de textos formais e informais foi crucial para garantir previsões precisas em textos informais. Além disso, a junção dos conjuntos de dados PL-corpus e C-corpus para o modelo BERT levou a um desempenho significativamente melhor

no conjunto PL-corpus. Essas descobertas destacam a importância de usar conjuntos de dados diversificados e representativos durante o treinamento dos modelos, para que eles possam generalizar bem para diferentes tipos de dados de entrada.

O modelo CRF, BiLSTM-CRF e o RoBERTa treinados com a combinação do PL-corpus e C-corpus não obtiveram a melhor pontuação absoluta no conjunto de teste do PL-corpus. No entanto, eles se destacaram por sua robustez em todos os casos. Ou seja, mesmo não tendo a melhor pontuação, esses modelos apresentaram um desempenho consistente e estável em todas as métricas avaliadas. Isso sugere que esses modelos também podem ser uma boa escolha para casos em que a estabilidade e a confiabilidade são importantes.

Considerando o melhor modelo em cada caso, ou seja, o *fine-tuning* do BERT treinado pela junção do PL-corpus ao C-corpus, foi analisado o desempenho de cada modelo em cada categoria e tipo de entidades nomeadas.

Na análise por categorias, a Figura 5.5(a) e a Figura 5.5(b) mostram que, para o PL-corpus, as categorias “DATA” e “PESSOA” apresentaram maior F1-score, comparadas com as demais categorias, sendo 98.42% e 96.72%, respectivamente. Assumimos que a pouca quantidade de exemplos “EVENTO” induziu o modelo a ter mais dificuldades em identificar e classificar corretamente as entidades com esta categoria, no qual apresentou um F1-score de 52.65%. Para o C-corpus, foi possível verificar uma menor oscilação nas métricas das categorias, variando em torno de 70.00% em todos os casos.

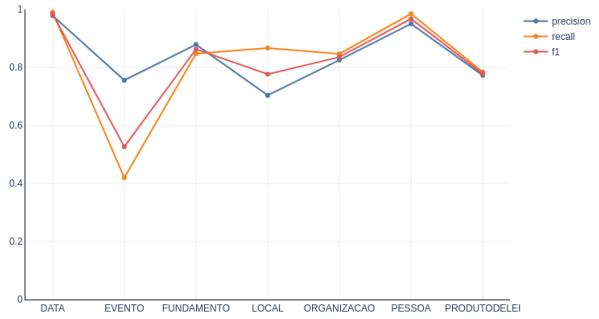
A Figura 5.5(c) mostra a análise no PL-corpus a nível tipo, na qual é possível observar que o modelo BERT obteve os melhores resultados para as entidades que pertencem às classes "DATA", "PESSOAindividual"e "PESSOAcargo", com F1-score de 98.12%, 97.39% e 95.50%, respectivamente. O modelo apresentou um baixo *recall* para a categoria "FUNDprojetodelei", o que pode estar relacionado à presença de poucas entidades pertencentes a essa classe nos documentos do tipo analisado.

Por fim, as entidades “EVENTO”, “FUNDprojetodelei”, “ORGpartido” obtiveram os melhores resultados no C-corpus, com F1-score acima de 80.00% (Figura 5.5(d)). É possível que as entidades da classe “LOCALvirtual” e “ORGnaogovernamental” tenham apresentado um desempenho inferior em relação às outras categorias devido a uma quantidade insuficiente de exemplos anotados.

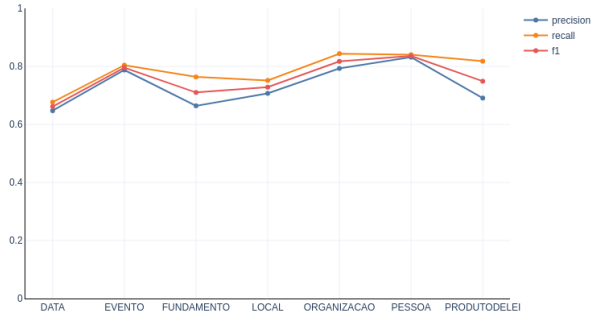
Os resultados obtidos indicam que a **hipótese 1** de que “a junção de textos formais a textos informais ajuda os modelos REN a identificar entidades em textos informais, no domínio legislativo“, foi confirmada. Além disso, foi observado que a junção dos dois conjuntos de dados (PL-corpus e C-corpus) no conjunto de treinamento foi benéfica para o modelo BERT, pois também ajudou a identificar as entidades nos textos formais.

Com base nas informações apresentadas, pode-se afirmar que a **hipótese 2** de que

“os modelos baseados na arquitetura *transformers* apresentam um desempenho superior para sistemas REN“, foi confirmada, uma vez que o modelo BERT apresentou os melhores resultados para todas as configurações diferentes de treinamento e teste. No entanto, o modelo RoBERTa não se destacou em relação ao CRF e ao BiLSTM-CRF, sugerindo que o tamanho e a qualidade do corpora utilizado no pré-treinamento podem ter influenciado no desempenho deste modelo para o REN.

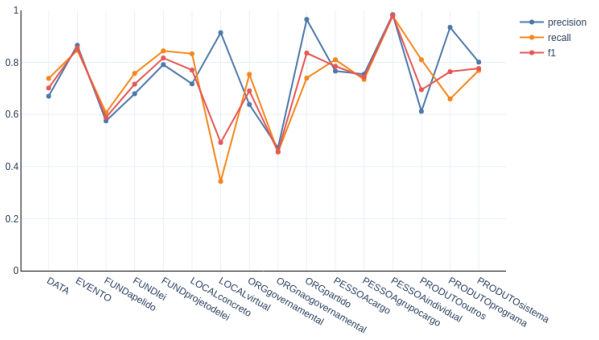


(a) *PL-corpus: categorias*



(b) *C-corpus: categorias*

(c) *PL-corpus: tipos*



(d) *C-corpus: tipos*

Figura 5.5: *Resultados do fine-tuning do BERT: Precisão, F1-score e Recall por categorias e tipos.*

CAPÍTULO 6

Conclusão

Este trabalho apresentou o "C-corpus", um *corpus* de comentários de cidadãos relacionados aos projetos de lei brasileiros para Reconhecimento de Entidades Nomeadas. Os textos foram anotados e curados manualmente por 3 equipes em 3 fases, e contêm entidades nomeadas que representam organização, pessoas, produto de lei, local, fundamento, evento e data. Além disso, enriquecemos o *corpus* UlyssesNER-Br, em que foi possível analisar a combinação de textos formais e informais em sistemas REN.

Por fim, treinamos o CRF, BiLSTM-CRF e o *fine-tuning* do BERT e RoBERTa. Os resultados obtidos indicaram que a **hipótese 1**: "a junção de textos formais a textos informais ajuda os modelos REN a identificar entidades em textos informais, no domínio legislativo", foi confirmada. Além disso, também observou-se que a junção dos dois *corpora* no conjunto de treinamento ajudou a identificar as entidades nos textos formais.

Embora o BERT tenha obtido resultados superiores, é importante ressaltar que o modelo CRF e BiLSTM-CRF também forneceram resultados satisfatórios, o que está de acordo com a literatura sobre a capacidade desses modelos em fornecer bons resultados para a tarefa de REN. Isso destaca a eficácia dos modelos tradicionais, mesmo em comparação com abordagens mais recentes baseadas em modelos de linguagem pré-treinados como o BERT.

A **hipótese 2** de que "os modelos baseados na arquitetura *transformers* apresentam um desempenho superior para sistemas REN", também foi confirmada, uma vez que o modelo BERT obteve os melhores resultados, com um F1-score médio de 78.65% na análise por categorias e 73.60% na análise por tipos, no C-corpus.

Foi observado que o modelo RoBERTa não obteve resultados superiores ao BERT, como era esperado, e sugere-se que isso ocorreu devido ao tipo de dados utilizados no pré-treinamento de cada modelo. Uma vez que o RoBERTa foi treinado em um *corpus* menor composto por *tweets*, pode não ter sido suficiente para capturar e modelar adequadamente as relações semânticas entre as palavras específicas do domínio legislativo.

6.1 Trabalhos Futuros

A ausência de entidades em algumas classes do C-corpus, como "LOCALvirtual"e "ORGnaogovernamental", afetou a performance do modelo BERT. É possível considerar o uso de técnicas de aprendizado de máquina que visem treinar modelos de REN com um número reduzido de exemplos rotulados, como o *Few-Shot NER*, como uma opção mais adequada para esse *corpus*. Com essa abordagem, seria possível obter resultados satisfatórios mesmo com um conjunto de dados menor, o que pode ser útil em situações em que há poucos dados rotulados disponíveis [31].

Também é possível testar o desempenho de arquiteturas variantes do BERT mais recentes, como o DeBERTa [51], porém seria recomendável realizar o pré-treinamento deste modelo em um *corpus* em português, dado que foram realizados testes com a versão multilíngue deste modelo, e os resultados obtidos para o REN usando o PL-corpus e o C-corpus foram insatisfatórios.

6.2 Sumário das Principais Contribuições

As três principais contribuições deste trabalho são:

1. um *corpus* de textos informais para REN, em português, no domínio legislativo, denominado C-corpus, que é uma extensão de UlyssesNER-Br [4]. O C-corpus foi coletado de uma plataforma online que permite a todos os cidadãos brasileiros interagir e expressar suas opiniões sobre projetos de lei em discussão no parlamento;
2. uma investigação da hipótese de que modelos de REN treinados usando documentos de diferentes tipos (textos formais e informais) pode aumentar a qualidade das previsões, em comparação com a modelagem de cada gênero separadamente;
3. o treinamento e avaliação dos modelos: CRF, BiLSTM-CRF, BERT e RoBERTa para os *corpora* em análise.

6.3 Publicações Geradas

Durante o processo de desenvolvimento deste trabalho, foram realizadas as seguintes contribuições:

- ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G. D.; SOUZA, E. P. R.; SILVA, N. F. F. D.; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A.; SIQUEIRA, F. A.; TARREGA, J. P. M.; BEINOTTI, J. V. P.; DIAS, M. D. S.; SILVA, M.; GARDINI, M.; SILVA, V. A. P. D.; CARVALHO, A. C. P. D. L. F. D.; OLIVEIRA, A. L. I. D. **Ulyssesner-br: a corpus of brazilian**

legislative documents for named entity recognition. In: International Conference on Computational Processing of the Portuguese Language - PROPOR. Springer, 2022.

- COSTA, R.; ALBUQUERQUE, H. O.; SILVESTRE, G.; SILVA, N. F. F.; SOUZA, E.; VITÓRIO, D.; NUNES, A.; SIQUEIRA, F.; TARREGA, J. P.; BEINOTTI, J. V.; DIAS, M. S.; PEREIRA, F. S. F.; SILVA, M.; GARDINI, M.; SILVA, V.; CARVALHO, A. C. P. L. F.; OLIVEIRA, A. L. I. **Expanding UlyssesNER-Br Named Entity Recognition Corpus with Informal User-generated Text**. In: Marreiros, G.; Martins, B.; Paiva, A.; Ribeiro, B.; Sardinha, A., editors, Progress in Artificial Intelligence, p. 767–779, Cham, 2022. Springer International Publishing.
- ALBUQUERQUE, H. O.; SOUZA, E.; GOMES, C.; PINTO, M. H. D. C.; FILHO, R. P.; COSTA, R.; LOPES, V. T. D. M.; SILVA, N. F. D.; CARVALHO, A. C. D.; OLIVEIRA, A. L. **Named entity recognition**: a survey for the portuguese language. Sociedad Española para el Procesamiento del Lenguaje Natural, (70):171–185, 2023.

Referências Bibliográficas

[1] AGARWAL, O.; NENKOVA, A. **The utility and interplay of gazetteers and entity segmentation for named entity recognition in English**. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 3990–4002, Online, Aug. 2021. Association for Computational Linguistics. 28, 30

[2] AGUILAR, G.; MAHARJAN, S.; LÓPEZ-MONROY, A. P.; SOLORIO, T. **A multi-task approach for named entity recognition in social media data**. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 148–153, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 28, 30

[3] AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. **Contextual string embeddings for sequence labeling**. In: *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638–1649, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. 18, 19

[4] ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G. D.; SOUZA, E. P. R.; SILVA, N. F. F. D.; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A.; SIQUEIRA, F. A.; TARREGA, J. P. M.; BEINOTTI, J. V. P.; DIAS, M. D. S.; SILVA, M.; GARDINI, M.; SILVA, V. A. P. D.; CARVALHO, A. C. P. D. L. F. D.; OLIVEIRA, A. L. I. D. **Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition**. In: *International Conference on Computational Processing of the Portuguese Language - PROPOR*. Springer, 2022. 12, 16, 34, 35, 37, 38, 55

[5] AMARAL, D. O. F. D.; VIEIRA, R. **Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields**. *Linguamática*, 6(1):41–49, Jul. 2014. 42

[6] ANGELIDIS, I.; CHALKIDIS, I.; KOUBARAKIS, M. **Named entity recognition, linking and generation for greek legislation**. In: *JURIX*, p. 1–10. IOS Press, 2018. 16

[7] BADJI, I. **Legal entity extraction with ner systems**. Master's thesis, Universidad Politécnica de Madrid, 2018. 16, 31

[8] BALDWIN, T.; DE MARNEFFE, M. C.; HAN, B.; KIM, Y.-B.; RITTER, A.; XU, W. **Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition**. In: *Proceedings of the Workshop on Noisy User-generated Text*, p. 126–135, Beijing, China, July 2015. Association for Computational Linguistics. 13, 27

[9] BASILE, P.; CAPUTO, A.; GENTILE, A.; RIZZO, G. **Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task**, p. 40–47. 01 2016. 27, 30

[10] BENAJIBA, Y.; ROSSO, P. **Arabic named entity recognition using conditional random fields**. 2008. 18

[11] BENDER, O.; OCH, F. J.; NEY, H. **Maximum entropy models for named entity recognition**. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 148–151, 2003. 13

[12] BIKEL, D. M.; SCHWARTZ, R. M.; WEISCHEDEL, R. M. **An algorithm that learns what's in a name**. *Machine Learning*, 34:211–231, 2004. 18

[13] BRUCKSCHEN, M.; NORTHFLEET, C.; SILVA, D.; BRIDI, P.; GRANADA, R.; VIEIRA, R.; RAO, P.; SANDER, T. **Named entity recognition in the legal domain for ontology population**. 05 2010. 30, 31

[14] BRUM, H. B.; NUNES, M. D. G. V. **Building a sentiment corpus of tweets in brazilian portuguese**. In: *International Conference on Language Resources and Evaluation - LREC*. European Language Resources Association, 2018. 43

[15] CAMPOS, D.; MATOS, S.; OLIVEIRA, J. L. **Biomedical named entity recognition: A survey of machine-learning tools**. In: Sakurai, S., editor, *Theory and Applications for Advanced Text Mining*, chapter 8. IntechOpen, Rijeka, 2012. 13

[16] CANO BASAVE, A.; PREOTIUC-PIETRO, D.; RADOVANOVIC, D.; WELLER, K.; DADZIE, A.-S. **microposts2016: 6th workshop on making sense of microposts: Big things come in small packages**. p. 1041–1042, 04 2016. 27, 30

[17] CANO BASAVE, A.; RIZZO, G.; VARGA, A.; ROWE, M.; STANKOVIC, M.; DADZIE, A.-S. **Making sense of microposts (microposts2014) named entity extraction linking challenge**. volume 1141, 04 2014. 26, 30

[18] CANO BASAVE, A.; VARGA, A.; ROWE, M.; STANKOVIC, M.; DADZIE, A.-S. **Making sense of microposts (msm2013) concept extraction challenge**. In: Cano, A.; Rowe, M.; Stankovic, M.; Dadzie, A.-S., editors, *MSM2013 : concept extraction*

challenge at Making Sense of Microposts 2013, CEUR workshop proceedings, p. 1–15. CEUR-WS.org, 2013. Cano Basave, AE, Varga, A, Rowe, M, Stankovic, M Dadzie, A-S: Making sense of microposts (MSM2013) concept extraction challenge. Proc. of the workshop on 'Making Sense of Microposts' co-located with the 22nd international World Wide Web conference (WWW'13), Rio de Janeiro, Brazil, 13 May, ceur-ws.org/Vol-1019/msm2013-challenge-report.pdf; Making sense of microposts ; Conference date: 13-05-2013. 26, 30, 36

[19] CAPELLARO, L.; CASELI, H. **Análise de polaridade e de tópicos em tweets no domínio da política no brasil**. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 47–55, Porto Alegre, RS, Brasil, 2021. SBC. 16

[20] CARDELLINO, C.; TERUEL, M.; ALEMANY, L. A.; VILLATA, S. **A low-cost, high-coverage legal named entity recognizer, classifier and linker**. In: *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, p. 9–18, New York, NY, USA, 2017. Association for Computing Machinery. 16

[21] CASTRO, P. V. Q. D. **Deep learning for named entity recognition in legal domain**. Master's thesis, Universidade Federal de Goiás, 2018. 16, 24

[22] CHIU, J. P.; NICHOLS, E. **Named Entity Recognition with Bidirectional LSTM-CNNs**. *Transactions of the Association for Computational Linguistics*, 4:357–370, 07 2016. 19

[23] CHRISTHIE, W.; REIS, J. C. S.; MORO, F. B. M. M.; ALMEIDA, V. **Detecção de posicionamento em tweets sobre política no contexto brasileiro**. In: *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil, 2018. SBC. 16

[24] CIURLINO, V. H. **Bertbr: A pretrained language model for law texts**, 2021. 19

[25] COHEN, J. **A coefficient of agreement for nominal scales**. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 36

[26] COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. P. **Natural language processing (almost) from scratch**. *CoRR*, abs/1103.0398, 2011. 18, 19

[27] COLLOVINI, S.; BONAMIGO, T. L.; VIEIRA, R. **A review on relation extraction with an eye on portuguese**. *Journal of the Brazilian Computer Society*, 19:553–571, July 2013. 16

[28] CORTIZ, D.; SILVA, J.; CALEGARI, N.; FREITAS, A.; SOARES, A.; BOTELHO, C.; RÊGO, G.; SAMPAIO, W.; BOGGIO, P. **A weakly supervised dataset of fine-grained emotions in portuguese**. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 73–81, Porto Alegre, RS, Brasil, 2021. SBC. 16

[29] COSTA, R.; ALBUQUERQUE, H. O.; SILVESTRE, G.; SILVA, N. F. F.; SOUZA, E.; VITÓRIO, D.; NUNES, A.; SIQUEIRA, F.; PEDRO TARREGA, J.; VITOR BEINOTTI, J.; DE SOUZA DIAS, M.; PEREIRA, F. S. F.; SILVA, M.; GARDINI, M.; SILVA, V.; DE CARVALHO, A. C. P. L. F.; OLIVEIRA, A. L. I. **Expanding ulyssesner-br named entity recognition corpus with informal user-generated text**. In: Marreiros, G.; Martins, B.; Paiva, A.; Ribeiro, B.; Sardinha, A., editors, *Progress in Artificial Intelligence*, p. 767–779, Cham, 2022. Springer International Publishing. 12, 29, 30

[30] **Núcleo interinstitucional de linguística computacional: Repositório de word embeddings do nilc**. URL: `http://www.nilc.icmc.usp.br/embeddings`, 2017. 42

[31] DAS, S. S. S.; KATIYAR, A.; PASSONNEAU, R. J.; ZHANG, R. **Container: Few-shot named entity recognition via contrastive learning**, 2022. 55

[32] DE MELO, T.; FIGUEIREDO, C. M. **A first public dataset from brazilian twitter and news on covid-19 in portuguese**. *Data in Brief*, 32:106179, 2020. 43

[33] DEMARTINI, G.; IOFCIU, T.; DE VRIES, A. P. **Overview of the inex 2009 entity ranking track**. In: Geva, S.; Kamps, J.; Trotman, A., editors, *Focused Retrieval and Evaluation*, p. 254–264, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. 13

[34] DERCZYNSKI, L.; BONTCHEVA, K.; ROBERTS, I. **Broad Twitter corpus: A diverse named entity recognition resource**. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, p. 1169–1179, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. 28, 30

[35] DERCZYNSKI, L.; NICHOLS, E.; VAN ERP, M.; LIMSOPATHAM, N. **Results of the WNUT2017 shared task on novel and emerging entity recognition**. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*, p. 140–147, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 14, 28

[36] DERCZYNSKI, L.; NICHOLS, E.; VAN ERP, M.; LIMSOPATHAM, N. **Results of the WNUT2017 shared task on novel and emerging entity recognition**. In:

Proceedings of the 3rd Workshop on Noisy User-generated Text, p. 140–147, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. 14

[37] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 12, 18, 19, 21, 22, 23, 29

[38] DOZIER, C.; KONDADADI, R.; LIGHT, M.; VACHHER, A.; VEERAMACHANENI, S.; WUDALI, R. **Named entity recognition and resolution in legal text**. In: *Semantic Processing of Legal Texts*, p. 27–43. Springer, 2010. 16

[39] FLORIAN, R.; ITTYCHERIAH, A.; JING, H.; ZHANG, T. **Named entity recognition through classifier combination.** In: *In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 168–171, 2003. 18

[40] FREITAG, D.; MCCALLUM, A. **Information extraction with hmms and shrinkage**. 1999. 18

[41] FREITAG, D.; MCCALLUM, A. **Information extraction with hmm structures learned by stochastic optimization**. In: *AAAI/IAAI*, 2000. 18

[42] GENTHIAL, G. **Tensorflow - reconhecimento de entidade nomeada**, 2017. 42

[43] GIACAGLIA, G. **How transformers work: The neural network used by open ai and deepmind**. URL: https://towardsdatascience.com/transformers-141e32e69591, 2019. 21

[44] GRAVES, A. **Sequence transduction with recurrent neural networks**, 2012. 21

[45] GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference- 6: A brief history**. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. 13

[46] GUILLOU, P. **Finetuning of the specialized version of the language model bertimbau on a token classification task (ner) with the dataset lener-br**. URL: https://github.com/piegu/language-models, 2022. 42

[47] GUMIEL, Y.; LEE, I.; SOARES, T.; FERREIRA, T.; PAGANO, A. **Sentiment analysis in portuguese texts from online health community forums: Data, model and evaluation**. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 64–72, Porto Alegre, RS, Brasil, 2021. SBC. 16

[48] HAGE, J. **Introduction. Papers from the Jurix '95 Conference**, volume 05. 1997. 30

[49] HAMMES, L.; FREITAS, L. **Utilizando bertimbau para a classificação de emoções em português**. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 56–63, Porto Alegre, RS, Brasil, 2021. SBC. 16

[50] HARTMANN, N.; FONSECA, E. R.; SHULBY, C.; TREVISO, M. V.; RODRIGUES, J. S.; ALUÍSIO, S. M. **Portuguese word embeddings: Evaluating on word analogies and natural language tasks**. *CoRR*, abs/1708.06025, 2017. 42

[51] HE, P.; LIU, X.; GAO, J.; CHEN, W. **Deberta: Decoding-enhanced bert with disentangled attention**, 2021. 55

[52] HECK, A. R. **Processamento de linguagem natural aplicado a reconhecimento de entidades nomeadas em textos legais em português brasileiro**, 2022. 21

[53] HUANG, Z.; XU, W.; YU, K. **Bidirectional lstm-crf models for sequence tagging**, 2015. 18, 19, 20, 21

[54] JIANG, R.; BANCHS, R. E.; LI, H. **Evaluating and combining name entity recognition systems**. In: *Proceedings of the Sixth Named Entity Workshop*, p. 21–27, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 24

[55] JUNIOR, E. A. C.; MARINHO, V. Q.; DOS SANTOS, L. B.; BERTAGLIA, T. F. C.; TREVISO, M. V.; BRUM, H. B. **Pelesent: Cross-domain polarity classification using distant supervision**, 2017. 43

[56] KAZAMA, J.; TORISAWA, K. **Exploiting wikipedia as external knowledge for named entity recognition**. p. 698–707, 01 2007. 13

[57] KLIE, J.-C.; BUGERT, M.; BOULLOSA, B.; ECKART DE CASTILHO, R.; GUREVYCH, I. **The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation**. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, p. 5–9. Association for Computational Linguistics, 2018. 32, 36

[58] KONKOL, M. **Named entity recognition**. Master's thesis, Faculty of Mechanical Engineering - University of West Bohemia, Pilsen, 2015. 13

[59] KOROTEEV, M. V. **BERT: A review of applications in natural language processing and understanding**. *CoRR*, abs/2103.11943, 2021. 21

[60] LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. **Conditional random fields: Probabilistic models for segmenting and labeling sequence data.** In: *In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, p. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 18, 20

[61] LAMPLE, G.; BALLESTEROS, M.; SUBRAMANIAN, S.; KAWAKAMI, K.; DYER, C. **Neural architectures for named entity recognition**. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 260–270, San Diego, California, June 2016. Association for Computational Linguistics. 19

[62] LE, H. T.; TRAN, L. V. **Automatic feature selection for named entity recognition using genetic algorithm**. *Proceedings of the Fourth Symposium on Information and Communication Technology*, 2013. 20

[63] LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. **A dataset of German legal documents for named entity recognition**. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 4478–4485, Marseille, France, May 2020. European Language Resources Association. 17

[64] LI, J.; SUN, A.; HAN, J.; LI, C. **A survey on deep learning for named entity recognition**, 2018. 14, 18

[65] LI, L.; ZHENG, S.; WANG, Q. **Roberta and stacked bidirectional gru for fine-grained chinese named entity recognition**. In: *2021 6th International Conference on Mathematics and Artificial Intelligence*, ICMAI 2021, p. 95–100, New York, NY, USA, 2021. Association for Computing Machinery. 19

[66] LIMSOPATHAM, N.; COLLIER, N. **Bidirectional LSTM for named entity recognition in Twitter messages**. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 145–152, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. 27, 30

[67] LITAKE, O.; SABANE, M.; PATIL, P.; RANADE, A.; JOSHI, R. **Mono versus multilingual BERT: A case study in hindi and marathi named entity recognition**. In: *Lecture Notes in Networks and Systems*, p. 607–618. Springer Nature Singapore, 2023. 19

[68] LIU, X.; ZHANG, S.; WEI, F.; ZHOU, M. **Recognizing named entities in tweets**. In: *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, p. 359–367, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 26, 30

[69] LIU, X.; ZHOU, M.; WEI, F.; FU, Z.; ZHOU, X. **Joint inference of named entity recognition and normalization for tweets**. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, p. 526–535, USA, 2012. Association for Computational Linguistics. 26

[70] LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. **Roberta: A robustly optimized bert pretraining approach**, 2019. 19, 23, 29

[71] LOPEZ, C.; PARTALAS, I.; BALIKAS, G.; DERBAS, N.; MARTIN, A.; REUTENAUER, C.; SEGOND, F.; AMINI, M. **Cap 2017 challenge: Twitter named entity recognition**. *CoRR*, abs/1707.07568, 2017. 29

[72] LUZ DE ARAUJO, P. H.; DE CAMPOS, T. E.; DE OLIVEIRA, R. R. R.; STAUFFER, M.; COUTO, S.; BERMEJO, P. **LeNER-Br: a dataset for named entity recognition in Brazilian legal text**. In: *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), p. 313–323, Canela, RS, Brazil, September 24-26 2018. Springer. 16

[73] MA, X.; HOVY, E. **End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF**. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1064–1074, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 19

[74] MCCALLUM, A.; FREITAG, D.; PEREIRA, F. C. N. **Maximum entropy markov models for information extraction and segmentation**. In: *In Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, p. 591–598, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. 18

[75] MCCALLUM, A.; LI, W. **Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.** In: *In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, p. 188–191, troudsburg, PA, USA, 2003. Association for Computational Linguistics. 18

[76] MCCALLUM, A. **Efficiently inducing features of conditional random fields**. *UAI-03*, 10 2012. 20

[77] MINTZ, M.; BILLS, S.; SNOW, R.; JURAFSKY, D. **Distant supervision for relation extraction without labeled data**. In: *Proceedings of the Joint Conference of the*

47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, p. 1003–1011, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics. 26

[78] MONTEMURRO, M. **Beyond the zipf–mandelbrot law in quantitative linguistics**. *Physica A: Statistical Mechanics and its Applications*, 300:567–578, 11 2001. 14

[79] MOON, S.; NEVES, L.; CARVALHO, V. **Multimodal named entity recognition for short social media posts**, 2018. 29, 30

[80] NADEAU, D.; SEKINE, S. **A survey of named entity recognition and classification**. *Lingvisticae Investigationes*, 30:3–26, 2007. 13, 24

[81] NAGARAJAN, B. M. **Understanding user-generated content on social media**. Master's thesis, Wright State University, 2010. 14, 15

[82] NETO, M. V. D. S. **Tweet sentiment analysis via transfer learning in low resource scenarios**. Master's thesis, Universidade Federal de Goiás, 2022. 43

[83] NETO, M. V. D. S. **Hugging face twitter roberta br**. URL: https://huggingface.co/verissimomanoel/RobertaTwitterBR. 42

[84] NGUYEN, D. Q.; VU, T.; TUAN NGUYEN, A. **BERTweet: A pre-trained language model for English tweets**. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 9–14, Online, Oct. 2020. Association for Computational Linguistics. 29, 30

[85] NIKOLAOS ALETRAS, ANDROUTSOPOULOS ION, L. B. A. M.; DANIEL PREOTIUC-PIETRO, E. **Proceedings of the natural legal language processing workshop**. URL: http://ceur-ws.org/Vol-2645/, 2020. 30

[86] OLIVEIRA, B. S. N. **Aprendizado profundo para reconhecimento de entidades nomeadas em narrativas de roubos**, 2020. 13

[87] PARDO, T.; DURAN, M.; LOPES, L.; FELIPPO, A.; ROMAN, N.; NUNES, M. **Porttinari - a large multi-genre treebank for brazilian portuguese**. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, p. 1–10, Porto Alegre, RS, Brasil, 2021. SBC. 16

[88] PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L.; DESMAISON, A.; KOPF, A.; YANG, E.; DEVITO, Z.; RAISON, M.; TEJANI, A.; CHILAMKURTHY, S.; STEINER, B.; FANG, L.; BAI, J.; CHINTALA, S. **Pytorch: An imperative style, high-performance**

deep learning library. In: Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 42

[89] PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn: Machine learning in python**. *J. Mach. Learn. Res.*, 12(null):2825–2830, nov 2011. 43

[90] PENG, N.; DREDZE, M. **Named entity recognition for Chinese social media with jointly trained embeddings**. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 548–554, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. 28, 30

[91] PETASIS, G.; CUCCHIARELLI, A.; VELARDI, P.; PALIOURAS, G.; KARKALETSIS, V.; SPYROPOULOS, C. D. **Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods**. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, p. 128–135, New York, NY, USA, 2000. Association for Computing Machinery. 13

[92] PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. **Deep contextualized word representations**. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, p. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 18, 19

[93] PIROVANI, J. P. C. **CRF+LG : uma abordagem híbrida para o reconhecimento de entidades nomeadas em português**. PhD thesis, Universidade Federal do Espírito Santo, Vitória, 2019. 13

[94] PORCARO, L.; SAGGION, H. **Recognizing musical entities in user-generated content**. *CoRR*, abs/1904.00648, 2019. 29, 30

[95] PUAIS, V.; MITROFAN, M.; GASAN, C. L.; CONESCHI, V.; IANOV, A. **Named entity recognition in the romanian legal domain**. In: *Proceedings of the Natural Legal Language Processing Workshop 2021*, p. 9–18, 2021. 30

[96] QUARESMA, P.; GONÇALVES, T. **Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents**, p. 44–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 16

[97] RITTER, A.; CLARK, S.; MAUSAM.; ETZIONI, O. **Named entity recognition in tweets: An experimental study**. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1524–1534, USA, 2011. Association for Computational Linguistics. 26, 30

[98] RIZZO, G.; CANO BASAVE, A.; PEREIRA, B.; VARGA, A. **Making sense of microposts (microposts2015) named entity recognition linking challenge**. 05 2015. 27, 30

[99] Rowe, M.; Stankovic, M.; Dadzie, A.; Hardey, M., editors. **Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011**, volume 718 de **CEUR Workshop Proceedings**. CEUR-WS.org, 2011. 26

[100] SANTOS, D.; CARDOSO, N. **A golden resource for named entity recognition in portuguese**. In: *Computational Processing of the Portuguese Language*, p. 69–79, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 34

[101] SAVELKA, J.; ASHLEY, K. D. **Detecting agent mentions in us court decisions.** In: *JURIX*, p. 39–48. IOS Press, 2017. 16

[102] SILEO, D.; PRADEL, C.; MULLER, P.; VAN DE CRUYS, T. **Synapse at cap 2017 ner challenge: Fasttext crf**. *arXiv preprint arXiv:1709.04820*, 2017. 29, 30

[103] SOBHANA, N.; PABITRA, M.; GHOSH, S. **Conditional random field based named entity recognition in geological text**. *International Journal of Computer Applications*, 1, 02 2010. 18

[104] SOHRAB, M. G.; DUONG NGUYEN, A.-K.; MIWA, M.; TAKAMURA, H. **mgsohrab at WNUT 2020 shared task-1: Neural exhaustive approach for entity and relation recognition over wet lab protocols**. In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 290–298, Online, Nov. 2020. Association for Computational Linguistics. 30

[105] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **Bertimbau: Pretrained bert models for brazilian portuguese**. In: Cerri, R.; Prati, R. C., editors, *Intelligent Systems*, p. 403–417, Cham, 2020. Springer International Publishing. 31, 42

[106] STRAUSS, B.; TOMA, B.; RITTER, A.; DE MARNEFFE, M.-C.; XU, W. **Results of the WNUT16 named entity recognition shared task**. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 138–144, Osaka, Japan, Dec. 2016. The COLING 2016 Organizing Committee. 27

[107] SUTTON, C.; MCCALLUM, A. **An introduction to conditional random fields**, 2010. 19, 20

[108] TABASSUM, J.; XU, W.; RITTER, A. **WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols**. In: *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 260–267, Online, Nov. 2020. Association for Computational Linguistics. 28

[109] TJONG KIM SANG, E. F.; DE MEULDER, F. **Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition**. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147, 2003. 13

[110] TJONG KIM SANG, E. F.; VEENSTRA, J. **Representing text chunks**. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 173–179, Bergen, Norway, June 1999. Association for Computational Linguistics. 33

[111] VENEROSO, J. M. D. F. **Named entity recognition on the web**, 2019. 18, 20

[112] WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. **The brWaC corpus: A new open resource for Brazilian Portuguese**. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). 42

[113] WAITELONIS, J.; SACK, H. **Named entity linking in tweets with kea**. 04 2016. 27

[114] WANG, X.; JIANG, Y.; BACH, N.; WANG, T.; HUANG, Z.; HUANG, F.; TU, K. **Improving named entity recognition by external context retrieving and cooperative learning**, 2021. 28, 30

[115] WILCOXON, F. **Individual comparisons by ranking methods**. *Biometrics*, 1:196–202, 1945. 25, 44

[116] WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; VON PLATEN, P.; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; LE SCAO, T.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. **Transformers: State-of-the-art natural language processing**. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online, Oct. 2020. Association for Computational Linguistics. 42

[117] XUAN, Z.; BAO, R.; JIANG, S. **Fgn: Fusion glyph network for chinese named entity recognition**, 2020. 28, 30

[118] YAMADA, I.; TAKEDA, H.; TAKEFUJI, Y. **Enhancing named entity recognition in twitter messages using entity linking**. In: *Proceedings of the Workshop on Noisy User-generated Text*, p. 136–140, 2015. 27, 30

[119] YOU, Y.; LI, J.; REDDI, S.; HSEU, J.; KUMAR, S.; BHOJANAPALLI, S.; SONG, X.; DEMMEL, J.; KEUTZER, K.; HSIEH, C.-J. **Large batch optimization for deep learning: Training bert in 76 minutes**, 2020. 12, 23

[120] ZIPF, G. **Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology**. Addison-Wesley Press, 1949. 14

APÊNDICE A

Lista dos projetos de lei relacionados aos comentários anotados

Projetos de Lei: PEC32/2020, PL318/2021, PL2893/2019, PL4425/2020, PL591/2021, PL461/2021, EMP11=>PL6438/2019, EMP4=>PEC10/2020, EMP5=>PEC10/2020, MPV905/2019, MPV927/2020, MPV943/2020, PDL101/2020, PDL156/2020, PDL164/2020, PEC10/2020, PEC101/2003, PEC108/2019, PEC149/2019, PEC206/2019, PL1029/2020, PL1036/2020, PL106/2020, PL1075/2020, PL1089/2020, PL1090/2020, PL1133/2020, PL1142/2020, PL1154/2020, PL1253/2020, PL1263/2020, PL1361/2015, PL1383/2020, PL1405/2020, PL1429/2020, PL1473/2020, PL1486/2020, PL1598/2020, PL1615/2019, PL1957/2020, PL1973/2020, PL2017/2020, PL2125/2020, PL2159/2020, PL2295/2000, PL2464/2020, PL2489/2020, PL2498/2020, PL2513/2020, PL2578/2020, PL2633/2020, PL2717/2019, PL3019/2020, PL4399/2019, PL4405/2019, PL5491/2019, PL597/2020, PL6371/2019, PL6381/2019, PL6460/2019, PL663/2020, PL675/2020, PL7816/2017, PL790/2020, PL80/2007, PL832/2019, PL866/2020, PL918/2020, PL936/2020, PLP101/2020, PLP149/2019, PLP34/2020, PLP39/2020 e REQ580/2020.