UNIFOR

FUNDAÇÃO EDSON QUEIROZ

UNIVERSIDADE DE FORTALEZA - UNIFOR

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA APLICADA

MESTRADO ACADÊMICO EM INFORMÁTICA APLICADA

ANTÔNIO MAURÍCIO BRITO JUNIOR

CDJUR-BR: UMA COLEÇÃO DOURADA DO JUDICIÁRIO BRASILEIRO COM ENTIDADES NOMEADAS REFINADAS

FORTALEZA – CE

Novembro, 2023

ANTÔNIO MAURÍCIO BRITO JUNIOR

CDJUR-BR: UMA COLEÇÃO DOURADA DO JUDICIÁRIO BRASILEIRO COM ENTIDADES NOMEADAS REFINADAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Informática Aplicada do Programa de Pós-Graduação em Informática Aplicada do Universidade de Fortaleza - UNIFOR da FUNDAÇÃO EDSON QUEIROZ, como requisito parcial à obtenção do título de mestre em Informática Aplicada. Área de Concentração: Sistema de Informação

Orientadora: Prof. Vládia Célia Monteiro Pinheiro, D.Sc.

FORTALEZA – CE

Novembro, 2023

Dados Internacionais de Catalogação na Publicação (CIP)

B862c Brito Junior, Antônio Maurício.
CDJUR-BR: uma coleção dourada do judiciário Brasileiro com entidades nomeadas refinadas/ Antônio Maurício Brito Junior. - 2023.
71 f.

Dissertação (Mestrado Acadêmico) – Universidade de Fortaleza. Programa de Mestrado Acadêmico em Informática Aplicada, Fortaleza, 2023.
Orientação: Prof. Dra. Vládia Célia Monteiro Pinheiro.

1. Reconhecimento de entidades nomeadas. 2. Anotação de corpus. 3. Coleção Dourada. 4. Processamento de linguagem natural. I. Pinheiro, Prof. Dra. Vládia Célia Monteiro. II. Título.

CDU 681.3:004.53

Elaborado por Biblioteca Central da Universidade de Fortaleza (UNIFOR)

ANTÔNIO MAURÍCIO BRITO JUNIOR

CDJUR-BR: UMA COLEÇÃO DOURADA DO JUDICIÁRIO BRASILEIRO COM ENTIDADES NOMEADAS REFINADAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Informática Aplicada do Programa de Pós-Graduação em Informática Aplicada do Universidade de Fortaleza - UNIFOR da FUNDAÇÃO EDSON QUEIROZ, como requisito parcial à obtenção do título de mestre em Informática Aplicada. Área de Concentração: Sistema de Informação

Aprovada em: 08 de novembro de 2023

BANCA EXAMINADORA

Prof. Vládia Célia Monteiro Pinheiro, D.Sc. (Orientadora)
Universidade de Fortaleza - UNIFOR

Prof. Evandro Eduardo Seron Ruiz, Ph.D.
Universidade de São Paulo - USP

Prof. João José Vasco Peixoto Furtado, Ph.D.
Universidade de Fortaleza - UNIFOR

Este trabalho é dedicado à minha esposa Vera, meu filho Leonardo (*in memoriam*) e minha filha Melissa que compreenderam a minha ausência durante as horas de dedicação e sempre me incentivaram.

AGRADECIMENTOS

A Deus por me dar o dom da vida e permitir chegar até este momento com Fé e esperança.

Aos meus pais, pelo amor, incentivo e apoio incondicional.

A minha esposa, Vera, por seu apoio, incentivo, compreensão e amor.

A professor Vládia Pinheiro, pela orientação, apoio e confiança.

A esta universidade, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, eivado pela acendrada confiança no mérito e ética aqui presentes.

Aos professores que me acompanharam ao longo do curso e que, com empenho, se dedicam à arte de ensinar.

“Ter fé é assinar uma folha em branco e deixar que Deus nela escreva o que quiser.”

(Santo Agostinho)

RESUMO

Esta dissertação apresenta o desenvolvimento da Coleção Dourada do Judiciário Brasileiro (CDJUR-BR), um corpus formado por 21 entidades refinadas anotadas de forma manual por especialistas em documentos jurídicos. Nosso objetivo principal é preencher uma lacuna existente de processos e recursos linguísticos, disponibilizando uma metodologia própria de criação de um *corpus* de entidades nomeadas (EN) e uma coleção dourada abrangente e robusta, em língua portuguesa, composta por 44.526 anotações, que possa servir ao processo de treinamento e validação de modelos de Inteligência Artificial Legal (*Legal AI*) na esfera do judiciário brasileiro. Neste trabalho, relatamos os critérios de seleção do *corpus* utilizado e ferramentas de anotações, as definições de entidades nomeadas e diretrizes de anotação, os treinamentos de anotadores especialistas, o processo de anotação e as métricas de aferição da concordância entre os anotadores e os resultados dos modelos de reconhecimento de entidades nomeadas (REN) utilizados.

Palavras-chave: Reconhecimento de Entidades Nomeadas · Anotação de Corpus · Coleção Dourada · Processamento de Linguagem Natural · Documentos Legais · Inteligência Artificial Legal · Aprendizado de Máquina · Português.

ABSTRACT

This thesis presents the development of the Golden Collection of the Brazilian Judiciary (CDJUR-BR), a corpus formed by 21 fine-grained entities annotated manually by experts in legal documents. Our main objective is to fill an existing gap in linguistic processes and resources, providing our own methodology for creating a corpus of named entities (NE) and a comprehensive and robust golden collection, in Portuguese, composed of 44,526 annotations, which can serve the training and validation process of Legal Artificial Intelligence (Legal AI) models in the sphere of the Brazilian judiciary. In this work, we report the selection criteria for the used *corpus* and annotation tools, the definitions of named entities and annotation guidelines, the training of expert annotators, the annotation process and the metrics for measuring agreement inter annotators and the results of the named entity recognition models (NER) used.

Keywords: Named-Entity Recognition · Corpus Annotation · Gold Standard Corpora · Natural Language Processing · Legal Documents · Legal Artificial Intelligence · Machine Learning · Portuguese.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

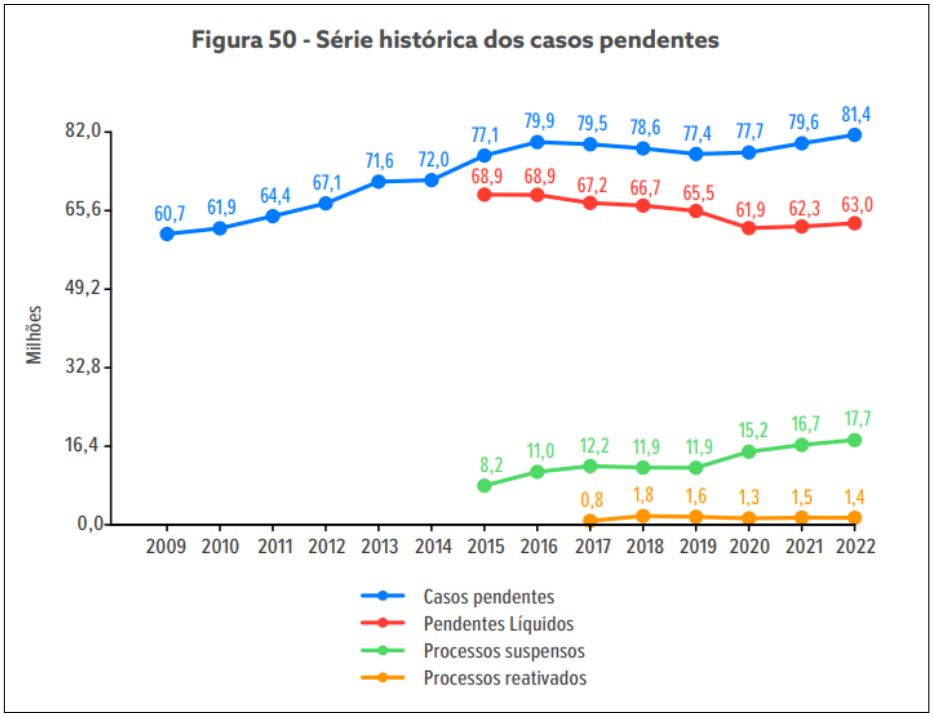
1 INTRODUÇÃO

Este capítulo situa o trabalho no contexto do Sistema Judiciário Brasileiro, abordando os desafios, a motivação para o desenvolvimento deste estudo, a problemática e questões de pesquisa, os objetivos, a metodologia e a estrutura desta dissertação.

1.1 CONTEXTUALIZAÇÃO E MOTIVAÇÃO

A morosidade da justiça vem sendo um dos principais problemas enfrentados pela sociedade brasileira quando precisa recorrer ao Poder Judiciário para a solução de seus litígios. O congestionamento dos tribunais devido ao grande volume de trabalho e ao aumento do número de casos contribuem para essa situação.

Figura 1 – Série histórica dos casos pendentes.



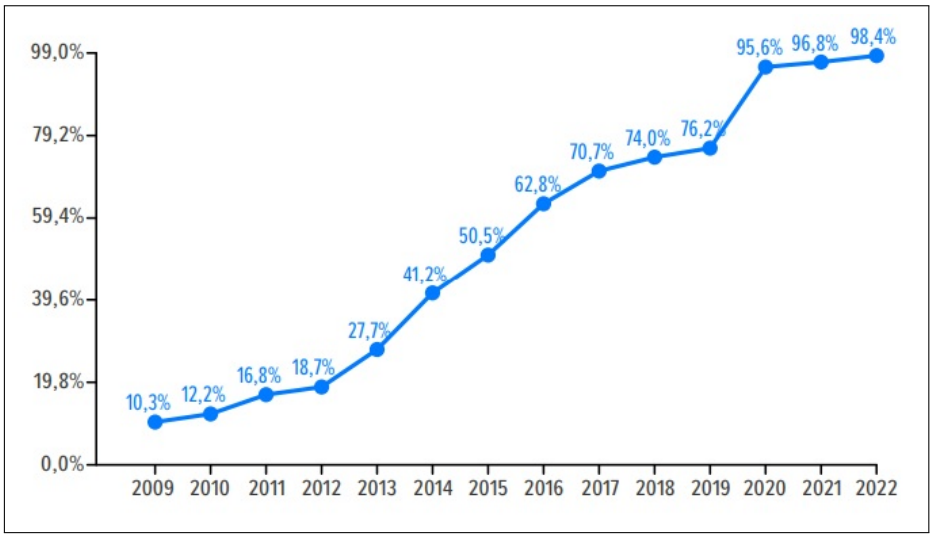
Fonte**:** Conselho Nacional de Justiça, 2023.

Ainda que no período de 2015 a 2020 tenha havido reduções sucessivas nos casos pendentes líquidos, como se pode observar na Figura 1 que apresenta a série históricas dos casos pendentes do Relatório Justiça em números/2023, do CNJ (CNJ, 2023), nos últimos anos houve um leve aumento deste indicador.

É importante enfrentar esses problemas complexos por meio de planejamento e desenvolvimento de novas competências gerenciais e de inovação no âmbito do Judiciário. Nesse

contexto, a Emenda Constitucional n° 45/2004 criou as condições para a criação do Conselho Nacional de Justiça - CNJ. Uma das iniciativas do CNJ foi a implementação do Processo Judicial Eletrônico (PJE), que trouxe benefícios significativos em relação ao processo tradicional em papel. A Figura 2 apresenta a evolução da adoção do processo eletrônico pelo judiciário brasiliero.

Figura 2 – Série histórica do percentual de processos eletrônicos.



Fonte**:** Conselho Nacional de Justiça, 2023.

Apesar das melhorias com o uso do Processo Judicial Eletrônico (PJE), ainda há gargalos de eficiência, principalmente devido a atos processuais que demandam tarefas intelectuais, como a elaboração de petições, recursos e, principalmente, as sentenças. Nesse contexto, a Inteligência Artificial Legal (*Legal AI*) desempenha um papel significativo, sendo capaz de reduzir o trabalho pesado e redundante para profissionais do direito. A maioria das atividades nessa área é apresentada em forma de texto, o que torna o Processamento de Linguagem Natural (PLN) (ZHONG *et al.*, 2020; CHOWDHURY, 2003) uma ferramenta fundamental para automatizar tarefas como classificação de processos (PEIXOTO, 2020), sumarização de documentos (KANAPALA; PAL; PAMULA, 2019; YAMADA; TEUFEL; TOKUNAGA, 2019), geração automática de sentenças e pareceres (ZHONG *et al.*, 2020), e busca de jurisprudências e normas jurídicas (ANGELIDIS; CHALKIDIS; KOUBARAKIS, 2018; BOELLA; CARO; LEONE, 2019).

Uma tarefa fundamental em aplicações de Inteligência Artificial no contexto jurídico é o Reconhecimento de Entidades Nomeadas (REN). Além de realizar uma simples classificação gramatical, o REN tem como objetivo identificar e qualificar se determinado trecho de texto

refere-se a entidades como pessoas, locais, organizações, datas, entre outros, acrescentando um viés semântico ao texto (YADAV; BETHARD, 2019). A literatura científica e os repositórios de códigos de IA são vastos em reconhecedores de entidades nessas categorias, os quais são amplamente empregados e treinados com base em coleções de textos marcados por especialistas humanos (SCHMITT *et al.*, 2019; LI *et al.*, 2020). Essas coleções são compiladas a partir de diversas fontes, incluindo enciclopédias, jornais, e obras literárias, tanto ficcionais quanto não-ficcionais. Esse conjunto de textos anotados, enriquecido com informações valiosas, é frequentemente conhecido como Coleção Dourada (*Golden Collection*) (SCHMITT *et al.*, 2019; JIANG; BANCHS; LI, 2016; ATDAĞ; LABATUT, 2013).

No entanto, textos de domínios específicos, como os produzidos no contexto da prática jurídica, fazem referências a outras entidades que não são trivialmente reconhecidas pelos RENs disponíveis atualmente. Isso se dá pelo fato de que textos jurídicos possuem discurso com termos técnicos, racionalmente ordenado e objetivando uma comunicação clara, precisa e concisa (COAN, 2003). É através do texto que o autor explicita a sua pretensão jurídica quando elabora uma demanda judicial, daí ser comum referências à entidades como leis e doutrinas, réus, vítimas, testemunhas, penalidades, etc, que embasam e buscam dar clareza às peças processuais. Da mesma forma, as respostas às petições legais, sentenças e decisões são produzidas seguindo similar vocabulário, estrutura e referências a estas entidades.

Como em toda área de conhecimento, certas entidades requerem conhecimento técnico do anotador para serem devidamente rotuladas em uma coleção dourada. Na área jurídica não é diferente. A habilidade de rotular que uma determinada citação (e.g. *artigo 5. da CF*) é uma norma legal, muito embora não seja de natureza complexa, necessita de conhecimento e experiência na produção dos documentos. Mais desafiador ainda se torna classificar a relevância da mesma para caracterizar do que se trata o texto jurídico. Trata-se de uma *normal principal* e que define o assunto da petição? Ou se trata de uma *norma acessória* servindo somente para apoiar os argumentos do peticionante? Outro exemplo seria o de classificar o papel semântico das pessoas mencionadas na peça processual: é uma vítima? o réu? o juiz? As respostas a essas perguntas são de natureza interpretativa e requer do anotador conhecimento técnico para fornecê-las.

O contexto supramencionado contribui para que sejam raros os exemplos de extensas coleções douradas para o domínio jurídico (ARAUJO *et al.*, 2018; LEITNER; REHM; MORENO-SCHNEIDER, 2020), o que é um claro obstáculo para o desenvolvimento de aplicações em *Legal AI*. Para a língua portuguesa, há carência de categorizações de legislação, jurisprudência,

provas, penalidades, dos papéis das pessoas em um processo jurídico (juiz, advogado, vítima, réu, testemunha), dos tipos de locais (local do crime, endereço do réu), etc. Neste sentido, ainda persiste a necessidade de uma coleção dourada robusta, anotada com entidades categorizadas com mais detalhes do domínio jurídico, e que abranja diversos documentos de um processo legal, como petições, inquéritos, denúncias, decisões e sentenças. Neste trabalho, essas entidades específicas são denominadas de "entidades refinadas".

Esta dissertação descreve o desenvolvimento da Coleção Dourada do Judiciário brasileiro (CDJUR-BR) contemplando um conjunto de entidades nomeadas refinadas, as quais foram anotadas de forma manual por especialistas em documentos jurídicos. A criação da CDJUR-BR seguiu uma metodologia própria que visou atribuir o caráter de abrangência e robustez à coleção contendo 21 entidades representativas (sob a perspectiva dos especialistas envolvidos) e que possa servir ao processo de treinamento e validação de modelos de *Legal AI* para língua portuguesa. Especialmente, para a rotulação das normas legais e seus artigos foram aplicadas etapas adicionais de refinamento e validação, pois essas entidades formalizam o raciocínio jurídico e são consideradas essenciais a uma série de aplicações de PLN no domínio jurídico.

A motivação deste trabalho é apresentar uma metodologia própria desenvolvida para guiar as atividades de anotação manual de Entidades Nomeadas (EN) do domínio jurídico, em língua portuguesa. Além disso, será disponibilizada uma coleção dourada, produzida para treinamento e validação de algoritmos de aprendizado de máquina utilizados em soluções de *Legal AI*, para o judiciário brasileiro.

1.2 PROBLEMÁTICA E QUESTÕES DE PESQUISA

Durante a realização da revisão bibliográfica, foi constatado que existem poucas coleções douradas disponíveis em língua portuguesa para dar suporte às aplicações de *Legal AI*. As coleções existentes são limitadas e, em grande parte, feitas a partir de *corpus* de notícias ou retirados de *sites* de conhecimentos gerais como a wikipédia. São raros os conjuntos de dados voltados para o domínio jurídico. As coleções existentes são de tamanho limitado e há poucas entidades refinadas (ARAUJO *et al.*, 2018; LEITNER; REHM; MORENO-SCHNEIDER, 2020). Além disso, foi observada a ausência de informações sobre as metodologias de anotação utilizadas. Diante dessa realidade, esta dissertação propõe a elaboração de uma metodologia própria para as anotações manuais de documentos que compõem as peças de um processo

jurídico, e faz uso prático da mesma para criar um coleção dourada com 21 entidades nomeadas refinadas para o judiciário.

Diante desta problemática, as seguintes questões de pesquisa são levantadas para direcionar o desenvolvimento deste trabalho:

- *QP1* - Como elaborar uma metodologia de anotações manuais de entidades nomeadas que contemple as especificidades e complexidades do domínio jurídico?
- *QP2* - A coleção dourada gerada é adequada para o treinamento e validação de modelos de *Legal AI*?

1.3 OBJETIVOS

Os objetivos deste trabalho são a criação de uma coleção dourada de Entidades Nomeadas (EN), em língua portuguesa, no domínio do judiciário brasileiro e a apresentação de uma metodologia própria de anotação para pesquisadores interessados no tema. Acredita-se que a proposta dessa metodologia contribuirá para a validação e reprodução de pesquisas nessa área, além de auxiliar na padronização de anotações e classificação de entidades nomeadas. Adicionalmente, será apresentada a avaliação da coleção dourada na tarefa de reconhecimento de entidades nomeadas, com o objetivo de comprovar a viabilidade prática da coleção criada. Nas seções a seguir, serão definidos os objetivos geral e específicos do trabalho, no escopo que foi definido.

1.3.1 Objetivo Geral

O objetivo deste trabalho é a criação de uma Coleção Dourada com Entidades Nomeadas Refinadas para o domínio do judiciário brasileiro, para uso em modelos de aprendizado de máquina que possibilitem o desenvolvimento de aplicações de IA jurídica em português.

1.3.2 Objetivos Específicos

Dentre os objetivos específicos, que irão suportar o objetivo geral, pode-se elencar:

a) Elaborar metodologia de anotação de entidades nomeadas no domínio jurídico;
b) Realizar a anotação manual de documentos legais por especialista para criar uma coleção dourada de EN;
c) Avaliar o trabalho de anotação através de medidas de concordância entre anotadores;

d) Avaliar a coleção dourada construída na tarefa de REN e comparar com trabalhos básicos de referência no domínio jurídico, em português;
e) Analisar os resultados obtidos com o modelo REN proposto.

1.3.3 Metodologia

Como meio de alcançar os objetivos definidos na seção anterior, vislumbra-se realizar as etapas descritas abaixo:

a) Realizar pesquisa bibliográfica para a fundamentação teórica de anotação de coleção dourada e reconhecimento de entidades nomeadas;
b) Formalizar a metodologia de anotação da coleção dourada;
c) Realizar a anotação baseada na metodologia definida;
d) Avaliar a coleção dourada criada na tarefa de REN;
e) Analisar os resultados obtidos e possíveis melhorias futuras identificadas.

1.3.4 Estrutura do Trabalho

Além deste capítulo introdutório, esta dissertação é composta por quatro outros, assim discriminados:

- No capítulo 2 será abordada a fundamentação teórica que embasa este trabalho e se apresenta conceitos impostantes sobre a anotação de coleção dourada e reconhecimento de entidades nomeadas. Também, serão mostrados os trabalhos relacionados identificados na revisão bibliográfica e o estado da arte sobre reconhecimento de entidades nomeadas no domínio do judiciário brasileiro, com seus recursos linguísticos computacionais disponíveis, tecnologias e ferramentas utilizadas;
- No capítulo 3 será apresentada a metodologia de anotação para a construção da coleção dourada de EN para o domínio do judiciário brasileiro (CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas);
- O capítulo 4 apresenta os cenários de experimentos, os modelos de reconhecimento de entidades nomeadas implementados e analisamos os resultados obtidos na avaliação da coleção dourada na tarefa de REN.
- E, por fim, no capítulo 5 apresenta as conclusões finais, as contribuições realizadas, as limitações do estudo, os possíveis trabalhos futuros e publicação decorrente desta Pesquisa.

Os documentos apresentados como exemplos neste estudo não estão sujeitos a segredo de justiça, sendo utilizados puramente para ilustrar os conceitos discutidos. Além disso, é importante ressaltar que os exemplo de anotações fazem uso de nomes de pessoas, organizações ou endereços fictícios.

2 FUNDAMENTAÇÃO TEÓRICA

Nos dias atuais há uma grande demanda da sociedade por soluções tecnológicas que possam realizar tarefas que antes só eram possíveis de serem realizadas por pessoas. Por isso a IA se tornou largamente utilizada nos mais diversos ramos de atividades. Porém, para que os algoritmos de IA funcionem com eficiência e confiabilidade, eles precisam ser "ensinados" com uma base de conhecimento fidedigno, em uma linguagem computacional compreensível, para que possam "aprender" o trabalho que lhes delegamos. Do ponto de vista da PLN, estas bases de conhecimento são conhecidas como coleções douradas e se referem a um conjunto de dados que foram anotados manualmente e representam "a verdade objetiva" e, também, são usados para avaliar os algoritmos de aprendizado de máquina. Ao longo deste capítulo, discorreremos sobre este e outros conceitos teóricos fundamentais que nos nortearam na realização deste estudo.

2.1 ANOTAÇÃO DE *CORPUS*

Segundo (HOVY; LAVID, 2010) a anotação de *corpus* é o processo de enriquecer um *corpus* adicionando informações linguísticas e outras informações, tais como, identificar e associar partes do texto a categorias pré-definidas ou estruturas gramaticais. As informações que serão anotadas dependerão dos objetivos que se pretende alcançar com o *corpus*. A execução das anotações, em si, pode ser realizada por humanos ou máquinas (ou uma combinação de ambos). Os autores ressaltam que a anotação manual não requer grande esforço de preparação dos documentos e, em pequenos *corpus*, poderá ser realizada com um custo relativo baixo para fenômenos linguísticos complexos. Todavia, para que os algoritmos de PLN baseados em aprendizado de máquina sejam treinados e avaliados de forma confiável, demandam *corpus* grande e de alta qualidade, tornando a atividade de anotação trabalhosa e cara. Por outro lado, anotações automáticas exigirão grande investimento na preparação do *corpus* e na programação do sistema de anotação. Seus resultados em marcações sintáticas de baixo nível são altas e confiáveis, porém em tarefas de marcação semântica de alto nível, como de entidades nomeadas, os resultados ainda não são precisos o bastante para uso prático em várias áreas (TOMANEK; WERMTER; HAHN, 2007; WISSLER *et al.*, 2014).

2.1.1 Coleção Dourada

Uma coleção dourada (ou *corpus* padrão-ouro) é aquela anotada manualmente por, ao menos, dois anotadores, que realizam o trabalho de forma independente e a concordância entre os anotadores é calculada para assegurar alta qualidade (WISSLER *et al.*, 2014). O termo "padrão-ouro" foi, originalmente, utilizado no domínio econômico, usado para caracterizar o sistema monetário em que os países atrelaram o valor de suas moedas à quantidade física de ouro como compromisso de credibilidade (BORDO; ROCKOFF, 1996). Este sistema não é mais usado, no entanto ainda é considerado muito estável. Como as tarefas de PNL dependem em grande parte do domínio e dos resultados pretendidos, são necessárias coleções douradas adequados para cada domínio e tarefa (WISSLER *et al.*, 2014).

2.1.2 Coeficiente Kappa de Cohen

Uma das principais preocupações no processo de criação de uma CD é assegurar a alta confiabilidade das anotações realizadas. Essa questão surge, principalmente, devido à variabilidade das anotações entre os especialistas que realizam o trabalho. Especialmente, nos grandes *corpus*, devido ao fato de que vários anotadores podem interpretar os fenômenos linguísticos de interesse de maneira diferente. Por isso, são previstos treinamentos dos anotadores e medições contínuas dos documentos anotados para que se alcance maior concordância entre os anotadores para os mesmos documentos trabalhados. Para isso, a medida Kappa de Cohen é uma das métricas mais utilizadas. O coeficiente de Kappa (k) é um método estatístico para avaliar o nível de concordância entre dois conjuntos de dados. Pode ser calculado pela equação:

$$k = \frac{C_o - C_e}{1 - C_e} \tag{2.1}$$

Onde C_o é a concordância observada e C_e a concordância esperada ao acaso. Se os anotadores estiverem em pleno acordo, então k = 1. Se não houver acordo, além do que seria esperado ao acaso, k = 0. O resultado de k também poderá ser negativo, apesar de que, na prática, isso é incomum. Porém, quando esse resultado ocorre é um indicador de um problema sério e podem ser interpretados como um completo desacordo entre os anotadores e indicam a necessidade de mais treinamento ou revisão de procedimentos de anotação (MCHUGH, 2012). Em termos objetivos, para avaliar a concordância entre os pares de anotadores em texto, são consideradas os seguintes pontos (WYNER; PETERS; KATZ, 2013):

- Duas anotações são concordantes (ou coincidentes) se cobrirem a mesma extensão de texto em um documento, ou seja, seus pontos inicial e final são idênticos;
- Duas anotações estão sobrepostas se houver interseção de alguma extensão de texto em comum;
- Uma das duas anotações está faltando.

Nos trabalhos da CDJUR-BR, a similaridade de cosseno é utilizada como métrica para comparar a similaridade entre as anotações. Assim, as anotações do mesmo tipo, depois de convertidas em vetores de *tokens* e terem as *stopwords* removidas, são comparadas usando a medida de similaridade de cosseno. Considerou-se que duas anotações são correspondentes quando a similaridade obtida é igual ou superior a 90% (HUANG *et al.*, 2008).

A tabela 0 apresenta uma interpretação sugerida por Cohen para os valores do coeficiente Kappa.

Tabela 0 – Interpretação do coeficiente Kappa de Cohen

Valor de Kappa	Nível de Concordância	% Dados Confiáveis
0 - 0,20	Nenhum	0 - 4%
0,21 - 0,39	Mínima	4 - 15%
0,40 - 0,59	Fraca	15 - 35%
0,60 - 0,79	Moderada	35 - 63%
0,80 - 0,90	Forte	64 - 81%
Acima 0,90	Quase Perfeita	82 - 100%

Fonte**:** MCHUGH, Mary L. Interrater reliability: the kappa statistic. Biochemia medica, v. 22, n. 3, p. 276-282, 2012.

2.2 INTELIGÊNCIA ARTIFICIAL LEGAL - *LEGAL AI*

Legal Inteligência Artificial, ou *Legal AI* na sigla em inglês, é um ramo de estudo que se concentra na aplicação das tecnologias de inteligência artificial para auxiliar advogados e os operadores do judiciário em suas atividades. Dadas as características da atividade jurídica estar intrísicamente relacionada com grandes volume de documentos escritos, como petições, pareceres jurídicos e sentenças, a *Legal AI* faz uso internsivo das técnicas do Processamento de Linguagem Natural e do aprendizado de máquina (ZHONG *et al.*, 2020). Para Surden (SURDEN, 2019), "IA e lei" envolve a aplicação de técnicas matemáticas e de computador para tornar a lei mais compreensível, gerenciável, útil, acessível ou previsível. Algumas soluções baseadas em NLP para uma variedade de tarefas de *Legal AI* já foram desenvolvidas, como previsão de julgamento legal, reconhecimento e classificação de entidades jurídicas, resposta a

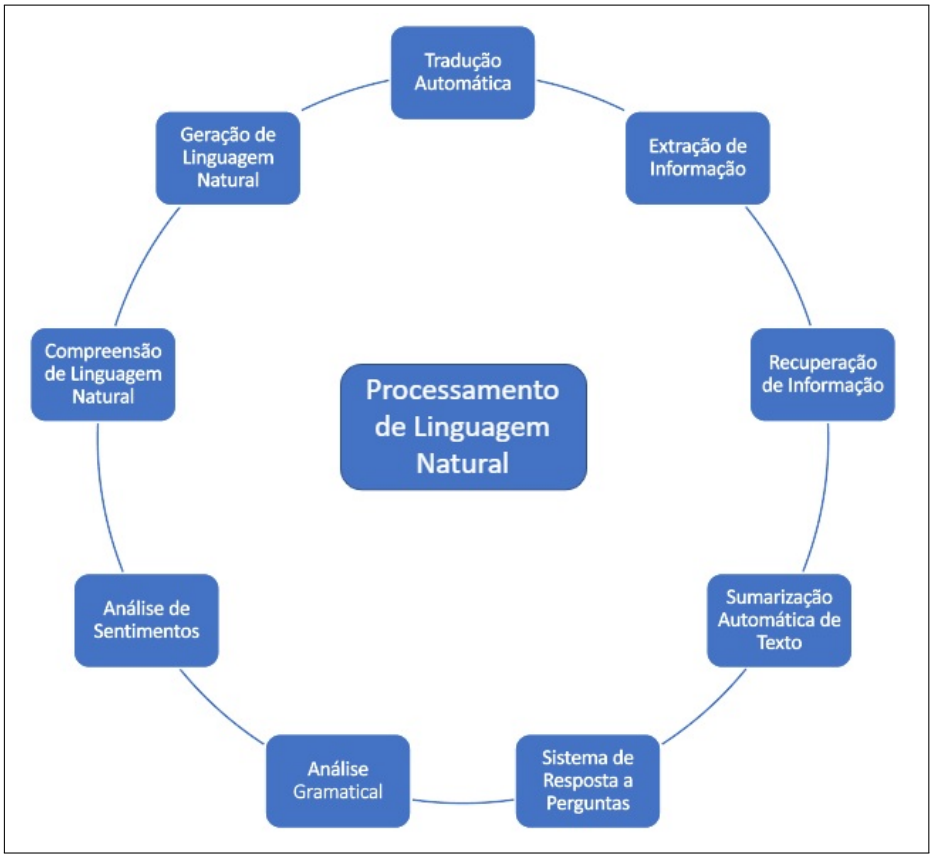
questões jurídicas, resumo legal (ZHONG *et al.*, 2020) e agrupamentos de documentos jurídicos (MARTINS, 2018; FIGUEIREDO, 2022). A despeito das relevantes contribuições que *Legal AI* proporciona, alguns questões atuais relativas a ética permanecem em aberto e gerando muito debate, como (1) o potencial de viés na tomada de decisão algorítmica, (2) a transparência sobre como os sistemas de IA estão tomando suas decisões e (3) problemas potenciais com deferência à tomada de decisão automatizada e computadorizada (SURDEN, 2019).

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Com a informatização e a Internet, uma infinidade de informações estão disponíveis e de acesso relativamente barato. Com isso, tratar estes dados e transformá-los em conhecimento para criar novos produtos e serviços tem sido uma estratégia comum utilizada por pessoas, empresas e governos. Grande parte dessas informações são textos livres, feitos para serem compreendidos por pessoas. São dados desestruturados e, portanto, de difícil interpretação computacional. O Processamento de Linguagem Natural é definido por Chowdhury (CHOWDHURY, 2003) como uma área de estudo que busca reunir conhecimento para compreender e fazer uso da linguagem natural humana, seja ela falada ou escrita, objetivando usar os sistemas computacionais para desenvolver aplicações úteis. Gonzalez (GONZALEZ; LIMA, 2003) ressalta que PLN busca tratar os diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, com o intuito de fazer com que o computador possa se comunicar em linguagem natural. Em qualquer tarefa de PLN é importante a compreensão da linguagem. Para Chowdhury (CHOWDHURY, 2003), compreender a linguagem depende equacionar três grandes problemas: primeiramente, compreender o pensamento ou ideia, o segundo entender e decifrar a representação e o significado das entradas linguísticas e o terceiro se refere ao conhecimento do mundo. Dito de outra forma, o fluxo de PLN pode começar a análise de um texto no nível da palavra ou tokens para determinar a sua forma ou estrutura morfológica. É nessa fase que a cada palavra é atribuída a sua classe gramatical (substantivo, verbo, adjetivo, advérbio e assim por diante). O passo seguinte é no nível da sentença para determinar a ordem das palavras e o significado da frase inteira. Ou seja, a análise semântica. E, por fim, a análise segue para o nível do contexto em que o texto está inserido. Para Singh (SINGH, 2018) as principais tarefas de PLN são: Tradução Automática (TA), Extração de Informação (EI), Recuperação de Informação (RI), Sumarização Automática de Texto (SAT), Sistema de Resposta a Perguntas, Analise Gramatical, Análise de Sentimentos, Compreensão de Linguagem Natural (CLN) e Geração de Linguagem Natural

(GLN). Ver ilustração na figura 3.

Figura 3 – Principais tarefas de PLN.



Fonte**:** Elaborado pelo autor

2.3.1 Extração de Informação

A EI é uma das primeiras e fundamentais atividades em tarefas de alto nível em PLN, como Tradução Automática, problemas de pesquisa mais avançados, como Pesquisa de Entidade, Pesquisa Estruturada e Sistemas de Resposta a Perguntas (SINGH, 2018; JIANG, 2012). A EI pode ser definida como um conjunto de técnicas e ferramentas computacionais para encontrar informações escritas em texto livre, ou seja, dados não estruturados, processá-los para convertê-los em um formato que possa ser inteligível e armazenado pelos computadores (SINGH, 2018). Os primeiros sistemas de EI foram baseados em regras linguísticas e padrões predefinidos, manualmente, que "casavam" com os textos e as informações pesquisadas. Esta tecnologia ainda é utilizada e alcançam bom desempenho em determinados domínios de conhecimento. Porém, é difícil desenvolver as regras e estas são muito dependentes do domínio para qual foram criadas. Com o tempo, novos sistemas baseados em estatística e aprendizado de máquina foram

superando as limitações da abordagem por regras. Além disso, a EI foi subdividida em muitos componentes que poderiam ser tratados como um problema de classificação fazendo uso de algoritmos supervisionados de aprendizado de máquina (JIANG, 2012). Nesta dissertação, será dedicada atenção a um dos componentes mais fundamentais na extração de informações, a saber, o reconhecimento de entidades nomeadas.

2.3.1.1 Reconhecimento de Entidades Nomeadas

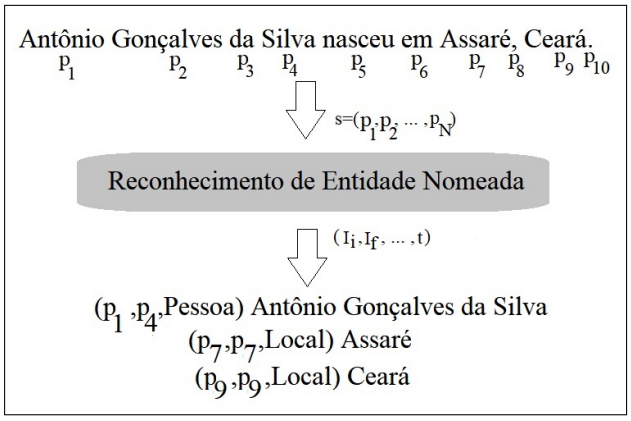
O termo Entidade Nomeadas (EN) foi apresentado durante a realização da Sixth Message Understanding Conference (MUC-6) (GRISHMAN; SUNDHEIM, 1996; NADEAU; SEKINE, 2007) e foi definido pela primeira vez como o reconhecimento de nomes próprios (COATES-STEPHENS, 1992). Para Jiang, entidade nomeada é uma palavra ou sequência de palavras que identifica uma entidade do mundo real em uma categoria estabelecida para um determinado domínio (JIANG, 2012). As categorias de entidades mais comuns registradas são nomes de pessoas, organizações e locais. Também, é comum o registro de entidades geopolíticas, relacionadas a tempo e numéricas (MIKHEEV; MOENS; GROVER, 1999; SINGH, 2018). Novas propostas foram feitas para subdividir essas entidades em categorias mais refinadas. A subcategoria POLÍTICOS foi proposta para a categoria PESSOA por ((FLEISCHMAN; HOVY, 2002)) e CIDADES foi adicionada a categoria LOCAIS (FLEISCHMAN, 2001; LEE; LEE, 2005). Entidades específicas e extensões as categorias existentes foram propostas e tratadas em algumas tarefas demandadas por sistemas de resposta a perguntas em domínios específicos (ROSSET *et al.*, 2008; SEKINE; NOBATA, 2004).

Nesta dissertação, entidades nomeadas refinadas são definidas como referências específicas a entidades do mundo real, minuciosamente identificadas dentro de subcategorias de categorias comuns. Essas referências podem consistir em palavras individuais ou expressões mais longas, abrangendo entidades mais complexas dentro de um domínio específico.

Os sistemas de REN realizam a identificação de entidades nomeadas em texto livres e os classificam em categorias predeterminadas como pessoa, organização e localização (JIANG, 2012). Sistemas REN funcionam como importante etapa de pré-processamento em várias tarefas de PLN como EI, RI e outras (MIKHEEV; MOENS; GROVER, 1999). Jiang (JIANG, 2012) ressalta que a extração de estrutura de informações mais complexas, como relações entre entidades e eventos, são muito dependentes da exatidão de sistemas REN em etapa prévia de processamento, sendo REN considerada uma das tarefas mais fundamentais em EI. Formalmente,

dada uma sequência de palavras $s = (p_1, p_2, ..., p_N)$, o REN deve produzir uma lista de tuplas (I_i, I_f, t_i), cada uma das quais é uma entidade nomeada mencionada em s. Sendo, $I_i \in [1, N]$ e $I_f \in [1, N]$ os índices inicial e final de uma menção de entidade nomeada e t é o tipo de entidade de um conjunto de categorias predefinido. A Figura 4 mostra um exemplo em que um sistema REN reconhece três entidades nomeadas da sentença dada.

Figura 4 – Exemplo de reconhecimento de entidade nomeada.



Fonte**:** Elaborado pelo autor

Os primeiros sistemas REN dependiam de padrões definidos manualmente, baseados em regras ou algoritmos de rotulagem de sequência. Algumas outras soluções REN foram construídas baseadas em listas predefinidas dos valores de entidades nomeadas. Estas soluções são rápidas e precisas, porém tem a limitação do tamanho da lista criada, pois para algumas entidades é impossível definir todos os valores possíveis. Posteriormente, passaram a usar dados rotulados para serem usados no treinamento de algoritmos de aprendizado de máquina supervisionado (SINGH, 2018; JIANG, 2012; LI *et al.*, 2020). Estes sistemas são treinados com grandes coleções de documentos anotados, apresentando exemplos positivos e negativos para que o algoritmo possa aprender as características das EN. O problema desta abordagem é a necessidade de grandes *corpus* anotados por especialistas, tornando-os recursos caros. Nos últimos poucos anos, as soluções de REN usando técnicas de Aprendizado Profundo (AP) passaram a ser dominantes, alcançando desempenho em "estado da arte", também, são atrativas por possibilitarem maior independência de domínio de conhecimento (YADAV; BETHARD, 2019).

Por esse motivo, as pesquisas avançam na alternativa de sistemas REN com aprendizado semi-supervisionado e aprendizado não-supervisionado (NADEAU; SEKINE, 2007). Neste trabalho, foi utilizada a abordagem de aprendizado supervisionado nos sistemas REN

implementados.

2.3.1.2 Métricas de Avaliação para REN

Os resultados de REN são essenciais para diversas aplicações de PLN e para o seu aperfeiçoamento contínuo, por isso é muito importante poder contar com sistemas REN precisos e confiáveis. Normalmente, a avaliação dos sistemas REN é feita comparando os seus resultados com as anotações feitas por especialistas humanos (coleção dourada). Estas avaliações podem ser realizadas verificando se as correspondências são exatas ou com maior flexibilidade. A correspondência exata só considera o reconhecimento como correto se o tipo e os limites textuais de uma entidade, forem exatamente iguais aos da coleção dourada. Por outro lado, as avaliações mais flexíveis consideram um acerto de tipo se houver reconhecimento do tipo correto, independente de haver correspondência exata com os limites do texto, desde que haja a correspondência com os limites definidos na coleção dourada. Da mesma forma, um limite é considerado correto independentemente do tipo que tenha sido associado a uma entidade (LI *et al.*, 2020). Esta forma de avaliação tem a vantagem de considerar todas as possibilidades de acertos e de erros, dando crédito parcial (1 ponto) quando o acerto é somente em uma das categorias (tipo ou texto), considera um crédito dobrado (2 pontos) para os acertos completos e, de forma contrária, penaliza os erros completos em dobro (2 pontos) e os erros simples são penalizados com um ponto. Apesar de mais complexa, esta forma de avaliação pode ser útil em aplicações que não é fundamental a exata identificação de uma entidade (NADEAU; SEKINE, 2007).

No contexto da CDJUR-BR, optou-se por uma avaliação mais flexível em relação aos limites das marcações. São considerados limites corretos quando há interseção entre os limites da predição e os limites definidos na coleção. Contudo, para que o reconhecimento seja validado como preciso em sua totalidade, o sistema de Reconhecimento de Entidades Nomeadas (REN) deve identificar com exatidão o tipo específico da Entidade Nomeada (EN).

Após definir o critério de correspondência que será adotado, a métrica final mais utilizada é a Medida-F (*F-Score*), que é calculada a partir da Precisão (P) e Revocação (R) ou Sensibilidade. Por sua vez, P e R são obtidas a partir do número de Verdadeiros Positivos (VP), Falsos Positivos (FP) e Falsos Negativos (FN):

- Verdadeiro Positivo (VP): Quando o resultado do reconhecimento de uma entidade é igual ao da CD;

- Falso Positivo (FP): Quando o REN identifica uma entidade, porém esta não aparece na CD;
- Falso Negativo (FN): Ocorre quando uma entidade existe na CD, mas não é reconhecida por um sistema REN.

Precisão (P) e Revocação (R) são definidas pelas fórmulas:

$$P = \frac{VP}{(VP + FP)} \tag{2.2}$$

$$R = \frac{VP}{(VP + FN)} \tag{2.3}$$

A Precisão denota a porcentagem de entidades reconhecidas positivas que são positivas reais, ou seja, coincidente com a CD. A Revocação é a porcentagem do total de entidades reconhecidas corretamente pelo sistema (POWERS, 2020; LI *et al.*, 2020).

A Medida-F combina Precisão e Revocação para trazer um número exclusivo que indica a qualidade geral do modelo. Ela é a média harmônica de Precisão e Revocação. A equação que a define, é a seguinte:

$$Medida - F = 2\frac{P * R}{P + R} \tag{2.4}$$

A Medida-F tende para o menor número, minimizando o impacto de grandes valores discrepantes e maximizando o impacto de pequenos valores. Sendo uma medida que tende a privilegiar os sistemas equilibrados (NADEAU; SEKINE, 2007).

Os sistemas REN, em geral, trabalham com várias categorias de entidades, por isso, para considerar o desempenho de todas, é comum realizar o cálculo da média Macro e Micro da Medida-F. Para calcular a Média Macro da Medida-F (MMAMF), primeiro é feito o cálculo da Medida-F de cada entidade e depois se faz a média dos valores obtidos (ver equações 2.5 a 2.7). Para o cálculo da Média Micro (MMIMF), soma-se as quantidades de VP, FP e FN de todas as categorias de entidades e depois calcula a média (ver equações 2.8 a 2.10). A Média Micro da Medida-F é mais sensível ao resultado do reconhecimento de entidades em *corpus* com categorias desbalanceadas (LI *et al.*, 2020).

$$Macro - P = \frac{\sum_{i=1}^{n} P_i}{n} \tag{2.5}$$

$$Macro - R = \frac{\sum_{i=1}^{n} R_i}{n} \tag{2.6}$$

$$MMAMF = 2\frac{Macro - P * Macro - R}{Macro - P + Macro - R} \tag{2.7}$$

$$Micro - P = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} (VP_i + FP_i)} \tag{2.8}$$

$$Micro - R = \frac{\sum_{i=1}^{n} VP_i}{\sum_{i=1}^{n} (VP_i + FN_i)} \tag{2.9}$$

$$MMIMF = 2\frac{Micro - P * Micro - R}{Micro - P + Micro - R}, \tag{2.10}$$

onde n é o número de entidades nomeadas.

A Média Ponderada da Medida-F é uma métrica de desempenho comumente usada para avaliar o desempenho geral de um modelo de classificação. É calculada computando a média ponderada da Medida-F para cada classe, onde os pesos são proporcionais ao número de amostras em cada classe. Para calcular a média ponderada da Medida-F (MPMF), primeiro calculamos a Medida-F para cada classe usando a fórmula citada em 2.4. Em seguida, calcula-se a Média Ponderada da Medida-F usando a seguinte fórmula:

$$MPMF = \frac{(p_1 * Medida - F_1 + p_2 * Medida - F_2 + ... + p_n * Medida - F_n)}{(p_1 + p_2 + ... + p_n)}, \tag{2.11}$$

onde a $Medida - F_1$, $Medida - F_2$, ..., $Medida - F_n$ são as Medida-F para cada classe, e p_1, p_2, ..., p_n são os pesos, calculados como o número de amostras em cada classe dividido pelo número total de amostras.

2.4 TRABALHOS RELACIONADOS

A prática de criar coleções douradas para o contexto jurídico tem na Europa seus maiores exemplos.

TRABALHO 01. O estudo realizado por (LEITNER; REHM; MORENO-SCHNEIDER, 2020), desenvolveu um conjunto de dados, em alemão, de entidades nomeadas e expressões temporais, a partir de 750 documentos de decisões judiciais publicadas online pelo Ministério Federal da Justiça e Defesa do Consumidor da Alemanha. Este conjunto de dados é parte de um esforço da União Europeia (UE) para apoiar, especialmente, PMEs que desejam atuar em outros mercados da UE, oferecendo serviços relacionados à *compliance*. No processo de criação do conjunto de dados, 54.000 entidades foram anotadas manualmente, mapeadas para 19 classes semânticas (pessoa, juiz, advogado, país, cidade, rua, paisagem, organização, empresa, instituição, tribunal, marca, lei, decreto, norma jurídica europeia, regulamento, contrato, decisão judicial e literatura jurídica). O artigo não apresenta metodologia para a etapa de anotação, mas descreve que foram desenvolvidas instruções específicas para as anotações. Estas instruções foram utilizadas para que um segundo anotador realizasse marcações em uma parte, não especificada, dos documentos. Para os documentos que tiveram duas anotações foi alcançada a concordância entre anotadores de 0,89 no coeficente Kappa. Eles também relatam que alcançaram a melhor Medida-F de 95,46 com um modelo de rede neural BI-LSTM e relatam que experimentos realizados com o BERT não apresentaram melhorias nos resultados.

TRABALHO 02. No trabalho desenvolvido por (ANGELIDIS; CHALKIDIS; KOUBARAKIS, 2018), foram anotadas 254 partes do Diário do Governo Grego, relativos a leis, decretos presidenciais, decisões ministeriais, regulamentos, como também os assuntos referentes a decisões relacionadas ao planejamento urbano, rural e ambiental entre os anos de 2.000 e 2.017. As anotações envolveram 6 tipos de entidades: PESSOA, para qualquer nome de pessoa citada nos documentos; ORGANIZAÇÃO, para qualquer referência a organização pública ou privada; ENTIDADE GEOPOLÍTICA, para qualquer referência a uma entidade geopolítica (por exemplo, país, cidade, unidade administrativa grega, etc.); MARCO GEOGRÁFICO, para especificar referências a entidades geográficas como bairros, estradas, praias, que constam principalmente de regulamentações relativas a planejamentos urbanísticos e topográficos; REFERÊNCIA À LEGISLAÇÃO, qualquer referência à decretos presidenciais, leis, decisões, regulamentos e diretivas da União Europeia ou Grega; REFERÊNCIA A DOCUMENTOS PÚBLICOS, qualquer referência a documentos ou decisões que tenham sido publicadas por uma instituição pública que não sejam consideradas uma fonte primária de legislação. O objetivo deste trabalho foi o reconhecimento de entidades nomeadas (REN) para enriquecer um grafo de conhecimento da legislação grega com informações mais detalhadas sobre as EN descritas. O artigo não menciona as atividades realizadas no processo de anotação manual das EN. Nos experimentos realizados,

eles relatam que alcançaram a Média Macro para a Medida-F de 0,88 na avaliação do Modelo REN desenvolvido.

TRABALHO 03. Huang et al. (HUANG *et al.*, 2020) realizaram o reconhecimento de entidades nomeadas para documentos de julgamento chineses com base nos modelos BI-LSTM e CRF, obtendo, no geral, 75,35 de Medida-F. Para tanto, precisaram trabalhar com as particularidades do idioma chinês que não há limites óbvios entre as palavras (como nas línguas ocidentais). Para resolver esse problema, eles propuseram uma abordagem nova, construindo vetores de caracteres e vetores de frases e os fundiram antes de enviá-los ao modelo BI-LSTM para treinamento. Para realizar os experimentos, foi construído um conjunto de dados anotado manualmente a partir de vários documentos judiciais, como processos criminais, civis e administrativos, obtidos da Rede de Documentos Judiciais Chineses. Os tipos de entidade anotados incluem nomes de pessoas, organizações, crimes, leis e regulamentos e penalidades. No total, foram feitas 40.737 anotações entre as diversas EN. No artigo, não há menção à metodologia adotada para a anotação do *corpus* gerado.

TRABALHO 04. No trabalho pioneiro realizado por (ARAUJO *et al.*, 2018), os autores disponibilizaram um conjunto de dados de entidades nomeadas, chamado de LENER-BR, construído a partir de anotações manuais de 66 documentos jurídicos de diversos tribunais brasileiros, entre eles o Supremo Tribunal Federal, Superior Tribunal de Justiça, Tribunal de Justiça de Minas Gerais e Tribunal de Contas da União. Adicionalmente, foram incluídos quatro documentos legislativos, como a Lei Maria da Penha, totalizando 70 documentos anotados. As entidades categorizadas foram "ORGANIZACAO"para organizações, “PESSOA” para pessoas físicas, “TEMPO” para entidades temporárias, “LOCAL” para localizações, "LEGISLACAO"para leis e "JURISPRUDENCIA"para decisões sobre processos judiciais. Ao todo, foram feitas 12.248 anotações de EN. O trabalho não cita as atividades realizadas durante o processo de anotação e nem se foram realizadas avaliações de concordância entre anotadores, porém relata que obteve Medida-F geral de 92,53%. Para as entidades específicas do domínio jurídico, obteve Medida-F de 97,00% e 88,82% para entidades de Legislação e Jurisprudência, respectivamente.

TRABALHO 05. O estudo realizado por (ALBUQUERQUE *et al.*, 2022), desenvolveu um conjunto de dados de entidades nomeadas, chamado de UlyssesNER-BR, a partir de 154 projetos de lei e 800 consultas legislativas da Câmara dos Deputados do Brasil, contendo dezoito tipos de entidades estruturadas em sete classes ou categorias semânticas. Baseadas no HAREM (SANTOS; CARDOSO, 2006) foram definidas 5 classes típicas: PESSOA, LOCALIZAÇÃO, ORGANIZAÇÃO, EVENTO e DATA. Além dessas, foram definidas duas classes semânticas

específicas para o domínio legislativo: FUNDAMENTOS DO DIREITO e PRODUTO DA LEI. A categoria de FUNDAMENTOS DO DIREITO faz referência a entidades relacionadas a leis, resoluções, decretos, bem como a entidades de domínio específico, como projetos de lei, que são propostas de lei em discussão no parlamento, e consultas legislativas, também conhecidas como solicitações de trabalho feitas pelo parlamentares. A entidade PRODUTO DA LEI refere-se a sistemas, programas e outros produtos criados a partir da legislação.

Os autores relatam que o processo de anotação ocorreu em três etapas. A primeira etapa foi usada como treinamento prático dos anotadores. Nas duas demais etapas, as anotações foram avaliados quanto a concordância entre anotadores usando a medida Kappa de Cohen. Ao final do processo de anotação, as equipes alcançaram a média geral no kappa de Cohen de 90%. Para as anotações foi usada a ferramenta Inception (KLIE *et al.*, 2018). Não há detalhes da quantidade de anotações anotadadas. Os modelos de aprendizado de máquina Hidden Markov Model (HMM) e Conditional Random Fields (CRF) foram usados para avaliar o *corpus*. Os resultados mostraram que o modelo CRF teve melhor desempenho na tarefas de REN, com pontuação média de Medida-F de 80,8% na análise por categorias e 81,04% na análise por tipos.

TRABALHO 06. Em (CASTRO *et al.*, 2019) o autor criou um *corpus* de EN no domínio jurídico, voltado para o Tribunal do Trabalho do Brasil, objetivando avaliar arquiteturas baseadas em Redes Neurais Profundas para desenvolver um modelo REN. Para a construção do *corpus*, foram selecionados 1.305 documentos dos tipos atas de audiências, sentenças e acórdãos. Estes documentos foram selecionados a partir de processos distribuídos em todas as 24 regiões da justiça do trabalho brasileira, entre os anos de 2008 e 2018. O processo de anotação foi realizado de maneira semi-automática: Primeiramente, foram anotados 76 documentos por um anotador e revisado pelo autor do trabalho para checar erros e garantir a aderência dos critérios e padrões de anotação. Em seguida, estes documentos foram usados para treinar a primeira versão do modelo automático de extração de entidades jurídicas. Para apoiar as anotações, foi utilizado o *software* Webanno. Posteriormente, o trabalho do anotador passou a ser mais uma revisão das anotações sugeridas pelo *software*. Foram anotadas as seguintes EN: FUNÇÃO, EN para indicar à função ou papel das pessoas mencionadas nos documentos. FUNDAMENTO, é a EN que especifica todo e qualquer dispositivo jurídico que possa ser referenciado nos documentos. LOCAL, EN para definir um endereço parcial ou completo. ORGANIZAÇÂO, EN para especificar organizações jurídicas formais ou não. PESSOA, EN para definir qualquer nome de pessoa física. TRIBUNAL, EN para identificar o Tribunal citado no documento. As entidades VALOR_ACORDO, VALOR_CAUSA, VALOR_CONDENACAO e VALOR_CUSTAS são

usadas para especificar os diversos valores presentes nos processos trabalhistas e, VARA, é EN para identificar a Vara citada no documento jurídico. No total, foram anotadas 4.578 EN. O *corpus* criado foi usado para treinar um modelo de REN para o domínio da Justiça do Trabalho do Brasil com o desempenho geral de 93.81% na Medida-F, utilizando um modelo baseado em Redes Neurais Profundas e o vetor de palavras Glove. A tabela 1 apresenta um resumo comparativo dos trabalhos analisados.

Tabela 1 – Resumo de trabalhos relacionados.

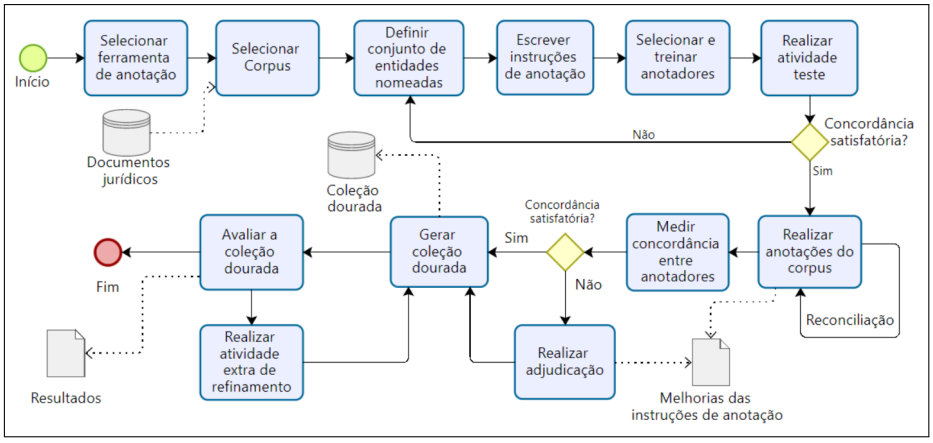
Trabalho	Nome da Coleção	Idioma	Domínio	Quantidade Anotações	Entidades Nomeadas	Apresenta Metodologia?	Kappa	F1 (REN)
01	-	Alemão	Jurídico	54.000	PESSOA, JUIZ, ADVOGADO, PAÍS, CIDADE, RUA, PAISAGEM, ORGANIZAÇÃO, EMPRESA, INSTITUIÇÃO, MARCA, DECRETO, LEI, NORMA JURÍDICA EUROPEIA, REGULAMENTO, CONTRATO, NORMA, DECISÃO JUDICIAL, TRIBUNAL, LITERATURA JURÍDICA	Não	0,89	95,46
02	-	Grego	Administrativo e jurídico	-	PESSOA, ORGANIZAÇÃO, ENTIDADE GEOPOLÍTICA, MARCO GEOGRÁFICO, REFERÊNCIA À LEGISLAÇÃO, REFERÊNCIA DOCUMENTOS PÚBLICOS	Não	-	88,00
03	-	Chinês	Jurídico	40.737	PESSOAS, ORGANIZAÇÕES, CRIMES, LEIS, REGULAMENTOS, PENALIDADES	Não	-	73,35
04	LENER-BR	Português	Jurídico	12.248	PESSOA, ORGANIZAÇÃO, TEMPO, LOCAL, LEGISLAÇÃO, JURISPRUDÊNCIA	Não	-	92,53
05	UlyssesNER-BR	Português	Jurídico	-	PESSOA, LOCALIZAÇÃO, ORGANIZAÇÃO, EVENTO, DATA, FUNDAMENTOS DO DIREITO, PRODUTO DA LEI	Não	0,90	80,80
06	-	Português	Jurídico	4.578	FUNÇÃO, FUNDAMENTO, LOCAL, ORGANIZAÇÃO, PESSOA, TRIBUNAL, VALOR_ACORDO, VALOR_CAUSA, VALOR_CONDENACAO, VALOR_CUSTAS, VARA	Não	-	93,81

Fonte**:** Produzido pelo autor.

3 CDJUR-BR: UMA COLEÇÃO DOURADA DO JUDICIÁRIO BRASILEIRO COM ENTIDADES NOMEADAS REFINADAS

Este capítulo descreve as atividades da metodologia proposta e executada para geração da CDJUR-BR, consistindo de seleção da ferramenta de anotação, seleção dos documentos que comporão o *corpus*, definição das entidades a serem anotadas, criação das instruções de anotação, seleção e treinamento dos anotadores, definição dos critérios de concordância entre anotadores, atividade teste, anotação e, por fim, avaliação e refinamentos das anotações. Um comitê com três professores da área do direito e dois da computação foi formado com o objetivo de definir, juntamente com os especialistas no domínio, os principais parâmetros da CDJUR-BR bem como zelar pela adequada aplicação da metodologia. A figura 5 ilustra as etapas da metodologia.

Figura 5 – Etapas da metodologia.



Fonte**:** Elaborado pelo autor.

3.1 SELECIONAR FERRAMENTA DE ANOTAÇÃO

Realizou-se uma pesquisa de mercado para identificar e avaliar ferramentas de anotação existentes, considerando critérios que fossem relevantes para o cenário de anotação deste projeto. O conjunto de ferramentas de anotação inicial é composto pelas ferramentas Inception/Webanno, Annotation Lab, Sinapses (CNJ), Doccano, Brat, LightTag, Label Studio, Labelbox, Tagtog, Superannotate, Telus International/Playment, CVAT, Sloth e Dataturk. Em seguida, foi especificado um conjunto de critérios para avaliar quais dessas ferramentas seriam mais adequadas. A lista de critérios que foram analisados são:

- Tipo de Dado: quais tipos de dados a ferramenta possibilita anotar (texto, imagem, vídeo,

áudio...).

- Perfis de Usuário: quais perfis de usuário é possível criar nesta ferramenta (anotador, curador, administrador...).
- Fases do Processo de Anotação que Controla: diferentes fases que a ferramenta controla (distribuição dos documentos, curadoria, avaliação de concordância entre anotadores, importação e exportação dos documentos, etc.).
- Anotação Automática: se é possível realizar anotação automática de documentos.
- Formato do arquivo de saída: quais os formatos de arquivo que é possível exportar as anotações e documentos.
- Permite colaboração: se é possível utilizar a ferramenta de uma forma colaborativa.
- Custo e suporte técnico: Avaliar o custo de licença para uso da ferramenta e/ou de disponibilidade de suporte técnico.

Ao final da avaliação, foram selecionadas e consideradas adequadas às necessidades do projeto as ferramentas Annotation Lab[1], Inception (KLIE *et al.*, 2018) e Tagtog (CEJUELA *et al.*, 2014). A ferramenta Tagtog foi adotada para este projeto por sua facilidade de iniciação das atividades e disponibilidade de suporte técnico.

3.2 SELECIONAR DOCUMENTOS DO *CORPUS*

Manning e Schutze preceituam que uma amostra é representativa se o que for encontrado para a amostra também for válido para a população em geral. Eles também defendem que nesta etapa dos trabalhos o foco deve ser a seleção dos documentos representativos do domínio em questão e baseado em critérios os mais objetivos possíveis (MANNING; SCHUTZE, 1999). Para os documentos da CDJUR-BR, os critérios definidos foram representatividade e qualidade. Quanto a representatividade, foram selecionados documentos das classes-CNJ (SILVA; HOCH; RIGHI, 2013) mais representativas dos processos de primeiro grau. Conjuntamente, essas classes representaram mais de 85% dos encerrados em 2019 no TJCE. São elas: Procedimento Comum Cível, Procedimento do Juizado Especial Cível, Execução Fiscal, Execução de Título Extrajudicial, Inquérito Policial, Ação Penal - Procedimento Ordinário e outras[2]. A quantidade de documento por classe foi definida de forma proporcional a frequencia real de processos em cada classe, com exceção para a classe Ação Penal que, devido a relevância que a mesma possui,

[1] https://www.johnsnowlabs.com/annotation-lab/

[2] Outras Classes compreendem: Relaxamento de Prisão, Execução da Pena, Alimentos, Medidas Protetivas de Urgência Criminal, Busca e Apreensão em Alienação Fiduciária.

teve um peso maior em relação às outras classes. Para o critério qualidade, foram definidos os documentos com 80% ou mais de palavras válidas da língua portuguesa e com mais de 50 *tokens*.

Na construção de um *corpus* de linguagem natural, o ideal seria aderir aos princípios teóricos da amostragem estatística e da inferência. Infelizmente, as abordagens padrões para amostragem estatística dificilmente são aplicáveis à construção de uma coleção dourada. Primeiro, muitas vezes é muito difícil delimitar a população total de forma rigorosa. Os livros didáticos sobre métodos estatísticos quase sempre se concentram em populações claramente definidas. Em segundo lugar, não existe uma unidade óbvia de linguagem que possa ser feita uma amostra e que possa ser usada para definir a população (MCHUGH, 2012).

Na escolha específica dos documentos que compuseram a CDJUR-BR os critérios adotados foram a relevância e representatividade, em termos de conteúdo, dos documentos num processo judicial, determinadas por uma equipe de especialistas do domínio jurídico. Foram selecionados os seguintes documentos: Petição Inicial, Petição, Denúncia, Inquérito, Decisão, Sentença, Despacho e Alegações Finais. A seleção final foi randômica dentre um conjunto de 80 mil documentos dos arquivos do TJCE que atenderam aos critérios acima explanados. A Tabela 2 apresenta aquantidade de documentos que compõem o *corpus* a ser anotado para a CDJUR-BR, por tipo de documento e classe, totalizando 1.216 documentos.

Tabela 2 – Composição do *corpus*.

Tipo de Documento	**Proc. Comum Cível**	**Juizado Especial Cível**	**Execução Fiscal**	**Execução Extra-judicial**	**Inquérito Policial**	**Ação Penal**	**Outras**
Petição Inicial	35	30	35	33	0	16	29
Petição	20	18	20	19	0	33	57
Denúncia	0	0	0	0	12	85	35
Inquérito	0	0	0	0	53	54	31
Decisão	21	21	20	0	22	71	32
Sentença	20	20	18	20	22	30	48
Despacho	11	12	13	30	2	6	22
Alegações Finais	0	0	0	0	2	115	23
Total por Classe	107	101	106	102	113	410	277
Total Geral							1.216

Fonte**:** Produzido pelo autor.

3.2.1 Tratamento do Texto

Dependendo da fonte do *corpus*, pode haver várias formatações e conteúdos com os quais não se pode lidar, e é apenas "lixo" que precisa ser retirado. Isso pode incluir: cabeçalhos, quebras de páginas, códigos tipográficos, tabelas e diagramas, dados ilegíveis no arquivo, etc. Se os dados vieram do OCR (*Optical Character Recognition*), o processamento pode ter introduzido problemas como cabeçalhos, rodapés, tabelas, figuras e notas de rodapé, quebrando os parágrafos do texto. Geralmente, também, haverá erros de OCR onde as palavras foram reconhecidas incorretamente. Se o seu programa destina-se a lidar apenas com texto em português, outros tipos de conteúdo, como tabelas e imagens, precisam ser considerados lixo. Muitas vezes, é necessário um filtro para remover o conteúdo indesejado antes de iniciar qualquer processamento adicional. Se deve ter bastante cuidado com tratamentos no texto original, como transformar todas as letras em maiúsculas ou minúsculas, bem como retirar pontuações. Pois, letras maiúsculas podem facilitar a identificação de nomes próprios, por exemplo. E, a pontuação poderá ajudar na identificação do final de frases, como acontece com os sinais de ponto final (.), interrogação (?) ou exclamação (!).

Para os trabalhos da CDJUR-BR, optou-se por manter o texto o mais fidedigno possível. Para isso, foram retirados dos documentos os caracteres inseridos pelo OCR para representar quebra de página e fim de arquivo.

3.3 DEFINIR CONJUNTO DE ENTIDADES NOMEADAS

As categorias de entidades nomeadas (ou classes semânticas) mais comuns são pessoa, organização e localização (MIKHEEV; MOENS; GROVER, 1999). Os sistemas de REN para estas categorias mais amplas, alcançaram alta acurácia e são muito úteis em diversas aplicações práticas. Porém, neste trabalho as aplicações de respostas a perguntas e recuperação de informações demandaram categorizações de entidades nomeadas mais refinadas para o domínio jurídico. Sendo definidas seis categorias de EN: Pessoa, Prova, Pena, Endereço, Sentença e Norma. Abaixo, descreve-se cada categoria e as EN que as compõem.

Pessoa: Esta categoria abrange as pessoas físicas ou jurídicas que normalmente são mencionadas nos documentos jurídicos. Foram definidas nove EN nesta categoria:

a) PES-AUTOR: EN que identifica o autor pessoa física ou jurídica de um processo.
b) PES-REU: EN que identifica o réu pessoa física ou jurídica de um processo.
c) PES-JUIZ: EN que identifica o juíz responsável pela condução e julgamento do

processo.

d) PES-ADV: En que identifica os advogados envolvidos nos processos.
e) PES-AUTORIDADE-POLICIAL: EN que identifica a autoridade policial anunciada nos autos do processo.
f) PES-OUTROS: EN que identifica qualquer outra pessoa mencionada nos autos.
g) PES-PROMOTOR-MP: EN que identifica o promotor de justiça representante do MP.
h) PES-TESTEMUNHA: EN que identifica as testemunhas citadas nos autos.
i) PES-VITIMA: EN que identifica a vítima em um processo penal.

Prova: Esta categoria compreende todas as modalidades de provas mencionadas nos documentos jurídicos.

Pena: Esta categoria foi criada para agregar as punições previstas na lei para os atos ilícitos praticados.

Endereço: Esta categoria abrange os diversos endereços mencionados nos autos dos processos. Foram definidas seis EN nesta categoria:

a) END-AUTOR: EN que identifica o endereço do autor do processo quando este for mencionado nos autos.
b) END-DELITO: EN que identifica o endereço de uma infração ou delito que seja mencionado nos autos.
c) END-REU: EN que identifica o endereço do réu mencionado nos autos.
d) END-TESTEMUNHA: EN que identifica o endereço de testemunhas que sejam mencionadas nos autos.
e) END-VÍTIMA: EN que identifica o endereço de vítimas mencionadas nos autos.
f) END-OUTROS: EN definida para identificar outros endereços que sejam mencionados nos documentos processuais.

Sentença: De acordo com o CPC (Código de Processo Civil), a sentença é o pronunciamento por meio do qual o juiz "põe fim à fase cognitiva do procedimento comum, bem como extingue a execução". Isso significa que, por meio da sentença, o juiz decide a questão trazida ao seu conhecimento, pondo fim ao processo na primeira instância. Esta categoria foi criada para identificar os atos pelo qual o juíz encerra uma das fases do processo, podendo ou não resolver o litígio.

Norma: O termo norma foi usado para categorizar, em um sentido mais amplo, o ordenamento jurídico (compreendendo leis e regulamentos). Para tanto, foram definidas três EN:

a) NOR-PRINCIPAL: EN que identifica as referência mencionadas de leis diretamente relacionadas ao assunto principal do processo.
b) NOR-ACESSÓRIA: EN que identifica as referência mencionadas de leis que contextualizam o documento jurídico em questão.
c) NOR-JURISPRUDÊNCIA: EN que identifica as referências mencionadas sobre decisões sobre determinada matéria proferidas, notadamente, por tribunais superiores.

3.4 ESCREVER INSTRUÇÕES DE ANOTAÇÃO

Uma vez definidas as entidades nomeadas a serem anotadas, inicia-se a elaboração das instruções de anotação. As instruções são diretrizes que deverão ser seguidas pelos anotadores com o propósito de alcançar maior concordância nas anotações realizadas entre os anotadores. Quanto maior a concordância, pressupõe-se maior a qualidade da coleção dourada que será criada.

No caso da criação da CDJUR-BR, as instruções foram sendo aprimoradas continuamente ao longo do trabalho por meio de reuniões entre as equipes de anotadores e a equipe de gestão dos trabalhos. Como resultado dessas reuniões, as dúvidas dos anotadores eram esclarecidas e exemplos mais detalhados foram compartilhados em novas versões das instruções para evitar mal-entendidos.

Nas duas seções a seguir, são exemplificadas as diretrizes utilizadas para as anotações dos documentos da CDJUR-BR.

3.4.1 Diretrizes de anotação para cada entidade.

Norma (Legislação, jurisprudência, doutrinas)

Ao realizar a anotação, deve-se selecionar a menção completa à norma jurídica (artigos, numeração, siglas, capítulos, etc).

Não é necessário marcar trechos das citações.

Analisar se é uma norma principal ou acessória para o caso.

a) NOR-PRINCIPAL: Normas que regulam o objeto principal do processo.
    - Exemplo de anotação: 8º denunciado (Douglas): (i) do *artigo 288, parágrafo único, do Código Penal*; (ii) do *artigo 157, § 3º, in fine, na forma do artigo 29, ambos do Código Penal, de acordo com o previsto no artigo*

1º, inciso II, da Lei n.º 8.072/90; (iii) do *artigo 157, § 2º, incisos I e II, na forma do artigo 29, ambos do Código Penal*, por TRÊS VEZES.

b) NOR-ACESSORIA: Normas de contextualização do processo.

- Exemplo de anotação: O MINISTÉRIO PÚBLICO, pelo Promotor de Justiça infraassinado, vem, a presença de V.ª Ex.ª, com fulcro no *inciso I, do artigo 129, da Constituição da República*.

c) NOR-JURISPRUDÊNCIA: É o conjunto de decisões que refletem a interpretação majoritária do tribunal.

- Exemplo de anotação: "A imunidade profissional contemplada no art. 133, da Constituição Federal, não é absoluta, sofrendo restrições legais. A lei apenas protege o advogado com relação às ofensas irrogadas no exercício da profissão em razão de discussão da causa, não socorrendo os seus excessos (art. 142, I, do CP e art. 7º, § 2º, da Lei 8.906/94)"*(RHC nº 12.458/SP, Relator o Ministro Jorge Scartezzini, DJU 29/9/2003)* 3. Ordem denegada

Pessoas

Ao realizar a anotação, deve-se selecionar a menção completa à pessoa física ou jurídica.

a) PES-AUTOR: Demandante/Autor da Ação - Pessoa Física ou Jurídica.

- Exemplo de anotação: O autor da Ação, o Sr. *João Araújo Monte Neto*, vem perante V.Sia...
- Exemplo de anotação: *XPTO BRASIL COMERCIAL LTDA*., sociedade empresária limitada inscrita no CNPJ/MF sob o n. 39.346.861/0001-61

b) PES-REU: Demandada Pessoa Física ou Jurídica.

- Exemplo de anotação: *Feliberto Braga*, já devidamente qualificadas nos autos em epígrafe, por intermédio de seu advogado

c) PES-VITIMA.

- Exemplo de anotação: FINALIDADE: APURAR SUPOSTO CRIME PREVISTO NO ARTIGO 149-A, IV DO CPB. INDICIADO(S): Em apuração - VÍTIMA: o *GOVERNO DO ESTADO DO CEARÁ*

d) PES-JUIZ.

- Exemplo de anotação: Voto proferido pelo i. Desembargador *Vladimir Abel da Silva*, membro do E. TJ/MS

e) PES-ADVOG: Advogado das partes de um processo.

- Exemplo de anotação: CONSTRUTORA S.A, pessoa jurídica de direito privado, sociedade empresária limitada, inscrita no CNPJ/MF sob o n.º 07.345.234/0001-45, com sede à Rodovia BR-116, km 27, S/N, 1º andar, bairro Jaboti, CEP 60.000-000, vem por meio de seu advogado Sr. *Pedro Paulo Reis*, OAB Número 766, propor a presente Ação de...

f) PES-TESTEMUNHA: Testemunha.

- Exemplo de anotação: Inclusive, cumpre ressaltar que, no depoimento de *MARIA VILMA*, mães das outras duas envolvidas nos aparentes crimes.

g) PES-AUTORID-POLICIAL: Autoridade policial.

- Exemplo de anotação: 1. Seja anexada a Requisição citada e seus anexos; 2. A seguir, voltem-me os autos conclusos para ulteriores deliberações. CUMPRA-SE. IGUATU, 13 de Abril de 2019 *MARCOS SANDRO MAZARÉ DE LIRA* Delegado

h) PES-PROMOTOR-MP: Representante do MP.

- Exemplo de anotação: Assinado digitalmente *EDNA LOPES* Promotora de Justiça(Respondendo mediante Portaria n.)

i) PES-OUTROS: Outras pessoas citados no processo.

- Exemplo de anotação: *FRANCISCO DE ASSIS*: Que nunca ouviu falar que Wesley Flavio estava envolvido com o trafico de drogas...

Endereço

Selecionar o endereço completo, ou seja, incluir o nome da rua ou avenida, número predial (podendo ser um número único ou um número seguido de uma letra), complemento (como um bloco, apartamento, sala, entre outros), bairro, cidade ou município, estado ou UF e código postal (CEP).

Identificar se o endereço é do reu, do autor, do delito, da testemunha, e outros.

Não precisa marcar endereços de cabeçalhos e rodapés.

a) END-AUTOR.

- Exemplo de anotação: SOFIA LIMA CRUZ, brasileira, menor...residente e domiciliada na *Rua Cel Luiz David de Souza, nº 72, Torre A, ap 1402, CEP 60110-000, Fortaleza/CE*, por intermédio de seu advogado que a presente subscreve...

b) END-REU.

- Exemplo de anotação:

c) END-DELITO.

- Exemplo de anotação: O delito ocorreu na *Rua Djalma Almeida, nº 750, Bairro Salinas*, em frente à praça Luíza Távora...

d) END-TESTEMUNHA.

- Exemplo de anotação: A parte autora arrola as testemunhas José Wellington e Fabíola Costa, domiciliados respectivamente na *Rua Antônio Lima 188* e na *Avenida Washington Soares 2504* Como testemunha que presenciou a prática do delito, o Ministério Público indica Rodrigo da Silva, vendedor ambulante, domiciliado na *Rua Teotonio Fonseca Lobo 235*

e) END-VÍTIMA.

- Exemplo de anotação: A vítima do delito, Sra. Fatima Barbosa, reside nesta cidade, na *Rua Tereza de Castro 786*

f) END-OUTROS.

- Exemplo de anotação: As partes deverão comparecer à audiência de conciliação, a acontecer no dia 12/08/2019, na *R. Des. Floriano Benevides Magalhães, 220*;

Sentença

Realizar esta anotação nos documentos do tipo sentenciais ou finais.

Selecionar o trecho de texto que expressa a definição sobre o mérito da causa – não a justificativa, mas a sentença em si.

- Exemplo de anotação: Isso posto, com fundamento no artigo 373, I, do NCPC, *JULGO IMPROCEDENTES* os pedidos iniciais e, por consequência, resolvo o mérito do processo, com resolução do mérito, o que faço com amparo no artigo 487, I, do NCPC.

Pena

Selecionar o trecho de texto que expressa a pena aplicada, não a justificativa, mas a pena em si.

- Exemplo de anotação: "fixo esta no pagamento de *10 dias-multa*, cada um no equivalente a um trigésimo do salário mínimo vigente, em observância ao disposto no art. 60 do código penal"
- Exemplo de anotação: "substituo a pena privativa de liberdade por duas restritivas de direito, a saber, *prestação pecuniária no importe de dois salários mínimos*, e *limitação de fim de semana* pelo período equivalente à

privação de liberdade".

Prova

Selecionar o trecho do texto com o nome ou menção à(s) prova(s). Ter atenção as diversas possibilidades de provas, tais como: provas testemunhais, provas documentais, provas periciais, provas circunstanciais e provas materiais.

- Exemplo de anotação: "Em razão da contradição nos depoimentos, procedeu-se à *acareação* em Juízo dos réus Erivaldo e Pedro Henrique"
- Exemplo de anotação: Segundo o *laudo médico de autopsia*... / De acordo com a *autópsia* acostada nos autos...

3.4.2 Diretrizes Gerais.

a) Não marcar a pontuação após o termo ou expressão a ser destacado (Ex. ponto final, vírgula, dois pontos, etc.).
b) Quando existirem vários autores ou réus e uma parte do nome deles for substituída pelo termo “outros”, não marcar o termo “outros” (Ex. Fulano de tal e outros).
c) Ao marcar pessoas, não incluir no destaque os respectivos pronomes de tratamento (Ex. Sr, Srª, V. Ex.ª, etc.) ou referência a titulação ou cargo (Ex. Dr., Dra., Des. Desa., etc.).
d) Ao marcar pessoas, não incluir no destaque referência posterior ao seu cargo (Ex. “Juiz de direito da Comarca X”, “Promotor de justiça”, “OAB nº X.XXX”, Procurador do Município, etc.).
e) Ao marcar o endereço, não incluir no destaque expressões que o antecedem (Ex. “residente”, “domiciliado”, “com sede”.
f) Não marcar artigos e preposições que antecedem dispositivos legais ou nome de pessoas.
g) Quando os dados de um julgado vierem entre parênteses, não marcar os parênteses.
h) Quando existirem várias normas citadas juntas, não marcar os conectivos (Ex. “combinado com”, “c/c”, “bem como”, “e”, “e seguintes”, “ss”, etc.).
i) Não marcar enunciados de fóruns como norma jurisprudencial (Ex. FPPC, FONAJE, FONANAMEC, etc.).
j) Ao marcar uma norma jurisprudencial, não incluir referências à destaques (Ex.

"grifou-se", "grifo nosso", "destaques no original", etc.).

k) Marcar todo o parágrafo do dispositivo e não apenas o seu núcleo.

3.5 SELECIONAR E TREINAR ANOTADORES

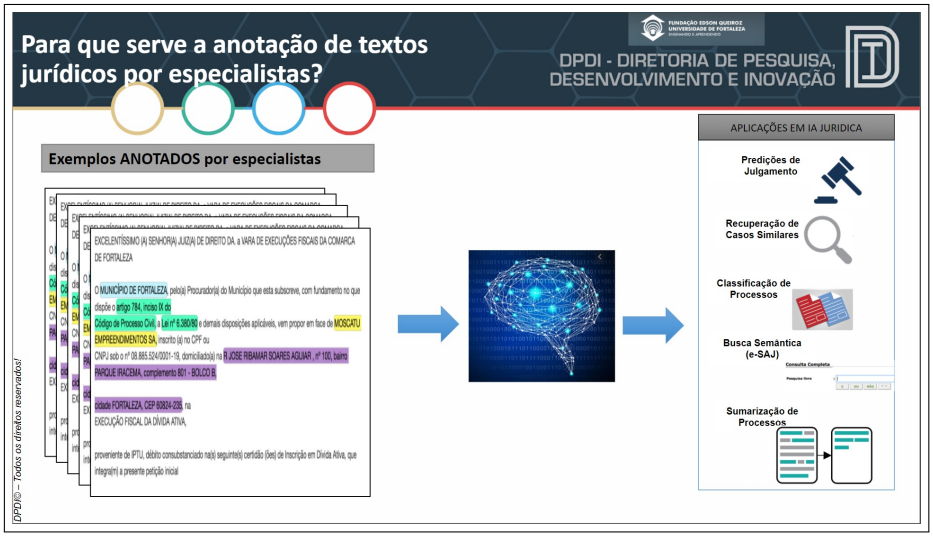
No processo de anotação, segundo (HOVY; LAVID, 2010), a abordagem geral adotada é se usar anotadores que são razoavelmente semelhantes em educação e cultura, realizar treinamento e disponibilizar um manual bastante específico para que se consiga a boa correspondência nas anotações.

Neste trabalho, foram formadas duas equipes de especialistas no domínio jurídico, selecionadas entre os funcionários dos parceiros do projeto para realizar as anotações: Uma composta, inicialmente, por 14 juízes estaduais, de ambos os sexos, com experiência entre 5 e 15 anos. Uma outra equipe composta por 19 promotores de justiça e técnicos judiciários, de ambos os sexos, com experiência entre 5 e 10 anos. Entre o início do treinamento e a fase de realização das anotações, 2 juízes e 1 promotor tiveram que se afastar por licenças trabalhistas e 2 técnicos judiciários se retiraram por mudança de vínculo empregatício.

O treinamento durou 90 minutos e teve como objetivo fornecer habilidades de anotação para os especialistas, que já possuíam amplo conhecimento dos procedimentos e da linguagem jurídica presentes nos documentos. Os seguintes tópicos foram abordados: Contextualização do projeto; Conceitos básicos de como se dar o treinamento de um algoritmo de aprendizado supervisionado; Para que serve a anotação de textos jurídicos por especialistas; Resumo da metodologia de anotação, *Corpus* para anotação; Conjunto de entidades nomeadas; Avaliação da concordância entre anotadores; Metas e prazos para a realização das anotações e treinamento prático do software Tagtog (CEJUELA *et al.*, 2014) que foi utilizado para as anotações.

As equipes dos TJCE e MPCE foram treinadas em momentos distintos e as aulas aconteceram por meio de ensino a distância devido as restrições impostas pela pandemia de Covid-19. Além do treinamento nos tópicos relacionados acima, foi proposta uma atividade prática de anotação como meio de avaliar a assimilação dos conceitos apresentados, avaliar a viabilidade das EN definidas e das instruções propostas, bem como, obter subsídios para implementar melhorias no processo como um todo. Figura 6 mostra um dos slides utilizados no treinamento.

Figura 6 – Exemplo de um slide de treinamento.



Fonte**:** Elaborado pelo autor.

3.6 REALIZAR ATIVIDADE TESTE

Os anotadores anotaram uma amostra de 87 documentos de todos os tipos e de todas as classes de processos, de forma a determinar a viabilidade da metodologia e das instruções de anotação. A concordância entre anotadores, determinada pela medida Kappa, alcançou 0,58. Como resultado desta fase, houve uma revisão das instruções e se chegou à definição final das orientações apresentadas. A atividade de teste foi fundamental para o trabalho de anotação, pois permitiu que os anotadores interagissem mais. Por meio de reuniões e recursos digitais de comunicação, foi possível esclarecer dúvidas sobre interpretação dos textos jurídicos, além de entender o objetivo dessas anotações. Os gestores e especialistas em PLN apoiaram a equipe e isso trouxe mais confiança e precisão ao processo de anotação. Em resumo, a atividade de teste tornou o trabalho dos anotadores mais eficiente e confiável.

3.7 REALIZAR ANOTAÇÕES DO *CORPUS*

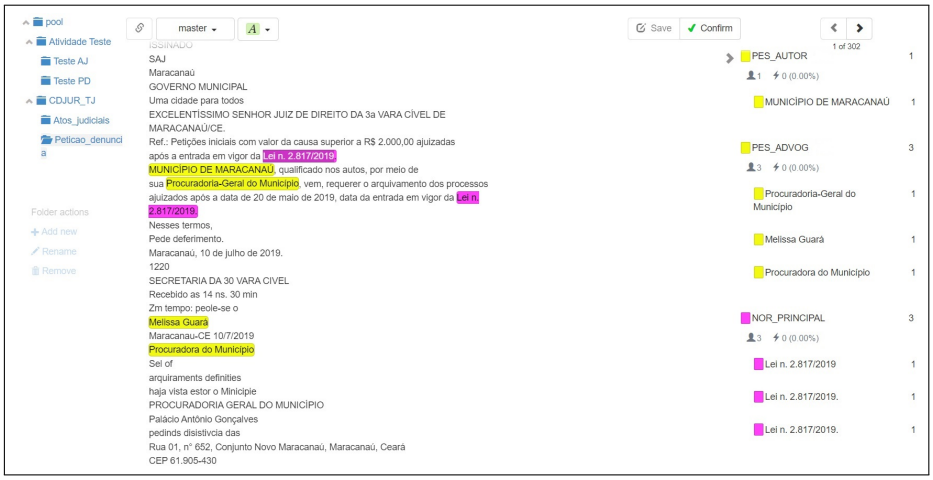
Foram configurados dois projetos no Tagtog para que as equipes do TJCE e MPCE pudessem trabalhar independentes nas anotações dos documentos definidos. O gerenciamento e suporte técnico às equipes de anotadores foi único, possibilitando melhor acompanhamento do ritmo das atividades, bem como a otimização dos recursos humanos envolvidos. Os documentos foram divididos aleatoriamente entre os anotadores de forma equitativa, garantindo que cada documento fosse anotado por um par de anotadores, através de um recurso automático disponível na ferramenta adotada. Em média, cada anotador da equipe do TJCE anotou 102 documentos e

na equipe do MPCE, a média foi de 80 documentos por anotador.

Inicialmente, havia a meta de anotar 1.216 documentos em 2 meses. Entretanto, esse objetivo não foi alcançado e foi necessário mais 15 dias para a conclusão.

Os anotadores tinham liberdade tática para escolher a melhor abordagem ao realizar o seu trabalho. Alguns optaram por anotar seguindo a ordem do texto e identificar as diferentes entidades ao longo do caminho, enquanto outros preferiram procurar por ocorrências de uma entidade em todo o texto antes de retornar ao início e identificar e anotar outra entidade. Esse processo continuava até que todo o documento fosse anotado. Em resumo, havia diferentes maneiras de realizar as anotações e cada pessoa utilizava a tática que considerava mais eficiente para concluir o trabalho. A figura 7 ilustra a anotação de um documento utilizando o Tagtog, destacando algumas entidades anotadas.

Figura 7 – Exemplo de anotação de documento.



Fonte**:** Elaborado pelo autor.

Diariamente, eram divulgadas as quantidades de documentos concluídos por cada anotador e o total alcançado no dia anterior para acompanhar o progresso e incentivar a equipe. No início da semana, verificava-se o coeficiente kappa para avaliar a concordância entre os anotadores. Assim, era possível corrigir ações e manter a qualidade e produtividade necessárias para atingir os objetivos.

A anotação de documentos jurídicos com muitas entidades nomeadas pode ser desafiadora devido à alta concentração necessária para compreender e interpretar o papel de cada entidade no documento. Por exemplo, a categoria Pessoa tem 9 entidades nomeadas e qualquer desatenção durante a marcação no texto pode levar a erros na identificação, como confundir uma testemunha com uma vítima. A identificação de uma entidade nomeada da

categoria Norma pode ser ainda mais complexa. Apesar de existirem apenas três entidades nessa categoria, diferenciá-las entre si exige todo o conhecimento e experiência jurídicas anteriores do anotador. É importante que o anotador tenha habilidade em indicar se uma determinada lei é o ordenamento principal ou se tem o propósito de contextualizar e embasar determinados atos ou ações relacionados ao objeto do processo.

3.8 MEDIR CONCORDÂNCIA ENTRE ANOTADORES

Durante o processo de anotação, foram realizadas medições para avaliar a concordância entre os anotadores, utilizando o coeficiente Kappa. Essa rotina de avaliação foi importante para assegurar mais qualidade a coleção dourada criada e, em caso de baixa concordância, identificar os motivos e tomar as medidas corretivas necessárias. Quando a concordância entre anotadores é alta, o coeficiente de Kappa é maior e indica que as instruções foram bem compreendidas, facilitando a identificação das anotações. Por outro lado, quando se tem baixa concordância, o coeficiente de Kappa será baixo, indicando que a tarefa de anotação é complexa ou pode não ter sido claramente definida. A Tabela 3 mostra os resultados obtidos por categoria de entidade, com 785 documentos que obtiveram coeficiente de Kappa igual ou superior a 0,50. A categoria Pessoa obteve o coeficiente Kappa 0,79, indicando a maior concordância entre os anotadores. A concordância é considerada "moderada"na interpretação sugerida por Cohen (Ver a Tabela 0). Ainda que esta categoria seja composta por 9 EN (Ver Tabela 4) o que poderia ter levado os anotadores a dúvidas entre as diversas partes representadas em um processo jurídico, seu processo de anotação não se mostrou uma tarefa complexa para os especialistas diferenciar as pessoas citadas. Na categoria Prova, percebe-se que os avaliadores tiveram dificuldades em chegar a um consenso nas marcações. Isso pode ter ocorrido em razão da grande variedade de tipos de provas que essa categoria engloba, o que torna complexo o processo de concordância entre os avaliadores. Essas dificuldades acabaram refletindo em um coeficiente de Kappa menor (0.46). Em Pena, o coeficiente de Kappa alcançou 0,64. Esta foi a segunda menor concordância observada. Dois pontos que merecem destaques são a complexidade de se identificar conceitualmente uma pena e, determinar os limites da anotação no texto, pois nos documentos jurídicos, é comum as penas serem descritas em textos mais longos.

A categoria Endereço alcançou Kappa de 0,73. Nos autos processuais, alguns endereços não aparecem com precisão. Por exemplo, local do delito, frequentemente são citados locais genéricos e não um endereço padrão, contendo rua e número da edificação. Apesar de

Tabela 3 – Resultados das anotações por categorias

Categoria	Anotações	Kappa
Pessoa	15.149	0.79
Prova	1.696	0.46
Pena	205	0.64
Endereço	2.041	0.73
Sentença	106	0.75
Norma	6.216	0.76
Total	**25.413**	**0,69**

Fonte**:** Produzido pelo autor.

situações como essa não serem as mais frequentes, elas ocorrem em muitos casos, implicando em uma dificuldade adicional para os anotadores chegarem a consenso em alguns documentos. A categoria Sentença alcançou uma boa concordância entre anotadores, com Kappa de 0,75. As sentenças costumam ser descritas em textos mais longos, no entanto, os anotadores conseguiram muito consenso nas anotações. A categoria Norma, formada por 3 EN, obteve Kappa médio de 0,76. É um resultado muito bom dada a importância dessa categoria, pois as referências às leis e outros atos jurídicos são uma característica intrínseca dos documentos jurídicos.

É importante salientar que os resultados foram obtidos antes da revisão feita por um terceiro avaliador (processo de adjudicação), o que se pressupõe que a CDJUR-BR final tem maior consistência e confiabilidade.

3.9 REALIZAR RECONCILIAÇÃO E ADJUDICAÇÃO

Durante a atividade de teste, os especialistas tiveram a oportunidade de compartilhar suas experiências e dúvidas ao anotar textos jurídicos, buscando ter um consenso em casos problemáticos. Isto foi possível graças às reuniões e ao grupo de comunicação entre os anotadores e a equipe de gestão, permitindo a uniformização de conceitos e interpretações. Como resultado, houve uma maior concordância entre os anotadores em suas marcações. Esta prática é conhecida como estágio de reconciliação (HOVY; LAVID, 2010) e ocorreu durante toda a fase de anotação. Porém, como em qualquer tarefa complexa, o acordo total nunca é possível, mesmo após as reconciliações. Então, quando se deu o fim desta etapa, seguiu-se a estratégia de adjudicação, que consiste em um terceiro especialista revisar os casos de desacordo e decidir (HOVY; LAVID, 2010). Foram selecionados 166 documentos com base no critério em que o Kappa foi inferior a 0,50. As revisões consistiram em um terceiro revisor receber as anotações dos dois anotadores iniciais (uma união das anotações). O revisor pôde decidir entre aceitar uma das anotações já feitas ou criar uma nova, diferente das anteriores. Além disso, no decorrer das atividades, alguns

dos especialistas selecionados deixaram os trabalhos, com isso, 176 documentos remanescentes foram anotados pela equipe de adjudicação.

3.10 GERAR COLEÇÃO DOURADA

Para disponibilizar o *dataset* final da CDJUR-BR, foi desenvolvido um formato próprio de arquivo que facilitasse a troca de dados entre sistemas e aplicativos. Para isso, foi utilizado o formato JSON (*JavaScript Object Notation*) devido à sua simplicidade, portabilidade, leveza, interoperabilidade e ao fato de ser nativo da linguagem de programação JavaScript, uma das linguagens mais populares e amplamente utilizadas na programação web. Daí, a escolha por esse formato se torna uma excelente escolha para o compartilhamento pretendido.

O formato do arquivo da CDJUR-BR tem uma estrutura semelhante a um dicionário do Python, em que as chaves (ou *keys*) do dicionário são os código dos documentos que compõem a coleção. E, cada documento, possui uma lista de entidades anotadas. Ou seja, cada item da lista é um dicionário com a entidade nomeada, a indicação da posição de início e de fim e o texto da anotação. A Figura 8 mostra um exemplo do formato de arquivo JSON da CDJUR-BR.

Figura 8 – Formato do arquivo da CDJUR-BR.


```
{ "279_DEC_91847485": [{"SENTENCA": {"inicio": 1425,
                                     "fim": 1529,
                                     "texto": "acolho a manifestação ministerial
                                     relativamente a este inquérito policial
                                     e\ndetermino 0 seu arquivamento"}},
                       {"END_DELITO": {"inicio": 571,
                                     "fim": 618,
                                     "texto": "terreno baldio localizado no bairro
                   dom lustosa"}},
                       {"PES_AUTOR": { "inicio": 644,
                                     "fim": 664,
                                     "texto": "0 ministério público"}},
                       {"PROVA": {"inicio": 377,
                                  "fim": 475,
                                  "texto": "inquérito policial registrado sob 0 n°
                                  110-541/2019, oriundo da delegacia do\n10° distrito
                                  policial"}}]}
```


Fonte**:** Elaborado pelo autor.

Para gerar a CDJUR-BR foi desenvolvido um aplicativo que lia os arquivos fornecidos pela ferramenta de anotação, decodificava o formato *anndoc* que é usado pelo Tagtog e faz a correspondência de pares de anotações consensuais feitas pelos especialistas. Para melhor aproveitamento dos trabalhos, quando duas anotações do mesmo tipo eram coincidentes (ou

sobrepostas), o texto final da anotação foi uma união entre o par de anotações originais. A Figura 9 ilustra o processo de seleção do texto final das anotações consensuais, no qual o texto final é determinado pelos limites inicial e final mínimos e máximos das duas anotações realizadas (no exemplo, os limites são 1796 e 1839).

O aplicativo, também, selecionava as anotações revisadas e as feitas individualmente por um dos revisores, além de ser responsável pelos cálculos de concordância entre anotadores para avaliar o coeficiente Kappa das anotações.

Figura 9 – Exemplo de união de anotações consensuais.

Anotador A: {"SENTENÇA": {"início": 1808, "fim": 1819, "texto": "arquivamento"}}

Anotador B: {"SENTENÇA": {"início": 1796, "fim": 1839, "texto": "determino o arquivamento deste procedimento"}}

Anotação final: {"SENTENÇA": {"início": 1796, "fim": 1839, "texto": "determino o arquivamento deste procedimento"}}

Fonte**:** Elaborado pelo autor.

3.11 REALIZAR ATIVIDADES EXTRAS DE REFINAMENTO

A categoria Norma recebeu atenção especial dos revisores devido a sua importância para os objetivos deste trabalho e para o domínio jurídico. Após os primeiros experimentos realizados com os sistemas de reconhecimento de entidades nomeadas (REN) desenvolvidos especificamente para a CDJUR-BR, verificou-se que havia muitas entidades reconhecidas pelos sistemas que não foram identificadas pelos anotadores e, portanto, não constavam na coleção dourada. Diante dessa constatação, a equipe de revisores recebeu um relatório com as normas identificadas pelo REN e analisaram se eram, de fato, normas e a que tipo específica deveriam ser associadas, ou seja, se seriam uma norma principal, acessória ou uma jurisprudência. Este processo de melhoria e refinamento propiciou uma ampla revisão das diretrizes de caráter jurídico das entidades, como também, quanto aos critérios relacionados a limites das anotações. Ao final dessa etapa, foram adicionadas 4.338 novas entidades de normas jurídicas.

Adicionalmente, foi desenvolvido um aplicativo em Python para realizar a correção automática de documentos em que o erro se limitasse a discordância quanto aos limites inicial e final da anotação. Para isso, a equipe gestora definiu que fosse feita a união das sentenças marcadas. Quando o aplicativo encontrava discordância entre entidades anotadas, gerava um

relatório apontando os documentos e as entidades que apresentavam as discordâncias. Com isso, os revisores puderam realizar suas atividades corretivas com maior praticidade, além de terem a liberdade de realizar novas anotações se assim julgassem necessárias. A Figura 10 exemplifica o relatório de revisão usado. No exemplo, o documento apresenta dois erros: O primeiro é uma discordância entre as entidades END_DELITO e END_TESTEMUNHA. O limite inicial da marcação é na posição 35.959 e a posição final é a 36.011, portanto a marcação tem 52 caracteres (incluindo espaços). Diante dessa informação o revisor analisaria o documento para decidir qual EN era a correta. O segundo erro é entre as entidades NOR_ACESSORIA e NOR_PRINCIPAL, com início em 49.111 e limite final em 49.151 (comprimento de 40 caracteres). O início e fim da anotação foi incluida para dar uma noção de onde se encontrava a anotação no documento.

A Tabela 4 apresenta estatísticas da coleção dourada.

Figura 10 – Exemplo do Relatório de Revisão.

Documento: 279_INQ_81801320 **Erros Identificados:** 02

#001

Entidades Nomeadas	Limite Inicial	Limite Final
END_DELITO;END_TESTEMUNHA	35959	36011

Anotação: rua do anjo branco, n° 1124 - cambeba - fortaleza/ce

#002

Entidades Nomeadas	Limite Inicial	Limite Final
NOR_ACESSORIA;NOR_PRINCIPAL	49111	49151

Anotação: § 2° do art. 2°, da lei n°. 12.830/2013.

Documento: 279_DEC_94390954 **Erros Identificados:** 05

Fonte**:** Elaborado pelo autor.

Tabela 4 – Estatística da coleção dourada.

Categoria	Anotações	%	Entidade Nomeada	Anotações	%
Pessoa	24.844	55,80	PES-ADVOG	735	1,65
			PES-AUTOR	1.259	2,83
			PES-AUTORID-POLICIAL	2.012	4,52
			PES-JUIZ	576	1,29
			PES-OUTROS	6.003	13,48
			PES-PROMOTOR-MP	363	0,82
			PES-REU	8.773	19,70
			PES-TESTEMUNHA	2.967	6,66
			PES-VITIMA	2.156	4,84
Prova	3.318	7,45	PROVA	3318	7,45
Pena	407	0.91	PENA	407	0,91
Endereço	2.065	4,64	END-AUTOR	132	0,30
			END-DELITO	466	1,05
			END-OUTROS	355	0,80
			END-REU	693	1,56
			END-TESTEMUNHA	295	0,66
			END-VITIMA	124	0,28
Sentença	172	0.39	SENTENÇA	172	0,39
Norma	13.720	30,81	NOR-ACESSORIA	5.767	12,95
			NOR-JURISPRUDÊNCIA	1.823	4,09
			NOR-PRINCIPAL	6.130	13,77
Total	44.526	100	**Total**	44.526	100

Fonte**:** Produzido pelo autor.

4 AVALIAÇÃO EXTRÍNSECA DA CDJUR-BR

Entre os objetivos deste trabalho está a criação da Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas (CDJUR-BR). Este objetivo teve como requisito fundamental a confiabilidade e consistência da coleção criada para assegurar sua viabilidade no treinamento de algoritmos de aprendizado de máquina utilizado nas soluções de *Legal AI*. Dada a importância estratégica destes requisitos, os esforços foram balizados pelas Questões de Pesquisa definidas na subseção 1.2. A *QP1* (Como elaborar uma metodologia de anotações manuais de entidades nomeadas que contemple as especificidades e complexidades do domínio jurídico?) foi respondida nas etapas de desenvolvimento das anotações, descritas na metodologia aqui apresentada (Ver Capítulo 3) e validada pelas avaliações de concordância entre anotadores (alcançando o coeficiente de Kappa geral de 0,69) e por meio das etapas de conciliação, adjudicação e atividades extras de refinamento com uso do REN treinado na CDJUR-BR, numa abordagem *human-in-the-loop* (rever as subseções 3.8, 3.9 e 3.11) que possibilitaram a adição de 19.113 anotações a coleção dourada final.

Para responder a *QP2* (A coleção dourada gerada é adequada para o treinamento e validação de modelos de *Legal AI*?), a CDJUR-BR foi avaliada na tarefa de reconhecimento de entidades nomeadas (REN). Os sistemas REN são cruciais para a maioria das aplicações que utilizam Processamento de Linguagem Natural (PLN). Isso faz com que avaliar a CDJUR-BR na tarefa de REN seja importante, já que a *Legal AI* é baseada em PLN. Para tanto, foram definidos alguns cenários de experimentos com sistemas REN. Os diferentes cenários criados, possibilitaram avaliar a performance da CDJUR-BR com classificadores distintos, bem como compará-la, quando isso foi possível, com a coleção de referência LENER-BR. Os detalhes dos cenários de experimentos serão apresentados a seguir.

4.1 CENÁRIOS DE EXPERIMENTOS

Para avaliar a CDJUR-BR, realizou-se vários experimentos em cenários que diferem quanto aos conjuntos de dados utilizados para treino, validação e teste e quanto as estratégias para representar as entidades. Em todos os cenários, foram utilizados os modelos descritos na seção 4.2 e as métricas Precisão, Revocação e Medida-F nas avaliações. Como há uma grande diferença entre as quantidades de anotações por categoria de entidades nomeadas, desenvolveu-se uma heurística para manter a mesma propoporção de exemplos da coleção completa quando a dividi em conjuntos de treino, validação e teste. Dessa forma, evitou-se que os conjuntos de

validação e teste ficassem com poucos exemplos, especialmente nas categorias Pena e Sentença. Os conjuntos de treino, validação e teste ficaram com 68,07%, 15,21% e 16,72% dos exemplos, respectivamente.

- **Cenário 1 - C1. Reconhecimento das entidades refinadas da CDJUR-BR.** Neste cenário, utiliza-se os dados da CDJUR-BR para treinar os modelos para o reconhecimento das entidades definidas na CDJUR-BR. O propósito com este cenário, é demonstrar a viabilidade da CDJUR-BR para o treinamento de modelos REN no domínio jurídico em língua portuguesa.
- **Cenário 2 - C2. Reconhecimento das categorias da CDJUR-BR.** As entidades refinadas da CDJUR-BR foram agrupadas nas seguintes categorias: Pessoa (categoria formada por todas as entidades refinadas que se referem a Pessoa, ou seja, todos os tokens representando Pessoa foram unicamente etiquetados como Pessoa), Legislação (categoria formada pelas entidades NOR-ACESSÓRIA e NOR-PRINCIPAL), Jurisprudência (categoria formada pela entidade NOR-JURISPRUDÊNCIA) e Local (categoria formada por todas as EN refinadas que se referem a Endereço). Em resumo, nesse cenário treina-se e se avalia o modelo com a CDJUR-BR, porém, o reconhecimento é em nível de categorias. Com isso, esse cenário nos possibilitará fazer comparações com a coleção LENER-BR.
- **Cenário 3 - C3. Reconhecimento das categorias de entidades da LENER-BR a partir de modelo treinado com LENER-BR.** Neste cenário, os modelos REN foram treinados utilizando o conjunto de dados de treino do LENER-BR, que contém 6 diferentes EN: Pessoa, Jurisprudência, Tempo, Local, Legislação e Organização. Esse canário nos apresentará o desempenho que os modelos alcançarão com a LENER-BR para se ter como referência comparativa de desempenho do REN.
- **Cenário 4 - C4. Reconhecimento das entidades do LENER-BR a partir de modelo treinado com CDJUR-BR.** Neste cenário, agrupa-se as entidades refinadas do conjunto de treino da CDJUR-BR nas categorias Pessoa, Legislação, Jurisprudência e Local e se treinam os modelos REN. Na fase de teste, foram avaliados o reconhecimento das entidades Pessoa, Legislação, Jurisprudência e Local com os dados do LENER-BR. Esse cenário possibilitará avaliar a capacidade de generalização do modelo REN treinado com a CDJUR-BR quando usado com outros documentos (no caso, os documentos que compuseram a LENER-BR).
- **Cenário 5 - C5. Reconhecimento das categorias de entidades da CDJUR-BR a partir de modelo treinado com LENER-BR.** Nesse cenário, os modelos REN foram

treinados utilizando o conjunto de dados de treino do LENER-BR. Porém, na fase de teste, avaliou-se os modelos no reconhecimento das seguintes categorias da CDJUR-BR: Pessoa, Legislação (categoria formada pelas entidades NOR-ACESSÓRIA e NOR-PRINCIPAL), Jurisprudência (categoria formada pela entidade NOR-JURISPRUDÊNCIA) e Local (categoria formada por todas as EN refinadas que se referem a Endereço). Esse cenário buscará demonstrar quão capaz serão os modelos treinados com a LENER-BR em reconhecer entidades de outra coleção (no caso, os documentos que compuseram a CDJUR-BR). Adicionalmente, os resultados permitirão comparar a capacidade de generalização dos modelos REN treinados com a CDJUR-BR e LENER-BR (ao se comparar os resultados obtidos no C4 com os resultados obtidos no C5).

4.2 MODELOS PARA O REN

Para realizar os experimentos, foram desenvolvidos três modelos distintos de sistemas REN. Utilizou-se um modelo com o SPACY para estabelecer uma linha de base para os experimentos. O SPACY é amplamente utilizado, não requer muito conhecimento e tempo para ser construído e consegue resultados razoáveis na tarefa de REN. Os outros dois modelos, BI-LSTM + CRF e BERT conseguem alcançar o desempenho estado-da-arte. A escolha destes modelos visou verificar qual deles tirará melhor proveito com os dados disponíveis. Seguem os detalhes de cada modelo:

- **Bidirectional Long Short-Term Memory (BI-LSTM) acrescido de uma Camada CRF (BI-LSTM + CRF)** (GRAVES; SCHMIDHUBER, 2005; HOCHREITER; SCHMIDHUBER, 1997; LAFFERTY; MCCALLUM; PEREIRA, 2001). A entrada do modelo é uma sequência de representações vetoriais de palavras individuais construídas a partir da concatenação de embeddings de palavras e embeddings de nível de caractere. Para a tabela de pesquisa de palavras, foi usado o Glove (PENNINGTON; SOCHER; MANNING, 2014) (vetor de palavras pré-treinadas em um corpus multi-gênero formado por textos em português do Brasil e da Europa). O modelo foi executado em 10 épocas e lotes com 10 amostras de tamanho; foi usado o otimizador SGD com uma taxa de aprendizado de 0,015.
- **Bidirectional Encoder Representations from Transformers (BERT)** (DEVLIN *et al.*, 2018). Foi utilizada a abordagem baseada em ajuste fino com o modelo pré-treinado BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), 10 épocas e tamanho de lote de 8 amostras. Como otimizador se usou o ADAM com uma taxa de aprendizado de $1x10^{-5}$.

- **SPACY**. Treinou-se o modelo para o componente NER do pipeline do SPACY (HONNIBAL; MONTANI, 2017), iniciando a partir do pacote em Português *pt_core_news_sm*.

4.3 RESULTADOS E DISCUSSÕES

A Tabela 5 apresenta os resultados obtidos no conjunto de teste para o Cenário 1 (C1) usando REN desenvolvidos com BI-LSTM + CRF, SPACY e BERT. O modelo com o BERT, comparativamente, obteve o melhor desempenho na grande maioria das entidades, alcançando a MMAMF de 0,58. Das 21 entidades da CDJUR-BR, 09 (nove) alcançaram Medida-F $\geq$ 0,70. Entretanto, 12 entidades tiveram Medida-F inferior a 0,70. Para alguns casos, como END-AUTOR, END-VÍTIMA e SENTENÇA, a causa foi o pequeno número de exemplos anotados para as entidades e os Falsos Positivos (FP) do tipo “O”, os quais representaram mais de 60% dos erros de predição. Para as entidades PES-ADVOG, PES-AUTOR, PES-OUTROS, PES-TESTEMNHA e PES-VÍTIMA, além dos FP com tokens do tipo “O”, também houve uma quantidade equivalente dos erros entre entidades da mesma categoria. Já para a entidade PROVA, a precisão foi excelente (0,87), com poucos FP. No entanto, a cobertura ficou baixa, em 0,33, com FN nas entidades Normas, Prova, Pessoa e, principalmente, nos tokens tipo "O".

O reconhecimento de entidades nomeadas com os modelos BI-LSTM + CRF e SPACY obtiveram desempenhos inferiores nos experimentos realizados, alcançando MMAMF de 0,55 e 0,42, respectivamente. Apesar do modelo BI-LSTM + CRF ter obtido um desempenho um pouco inferior ao BERT, as entidades NOR-JURISPRUDÊNCIA (0,90), PENA (0,56), PES-AUTOR (0,59), PES-JUIZ (0,79) e PROVA (0,47) obtiveram resultados iguais ou melhores de Medida-F dentre os modelos avaliados. Já o modelo implementado com o SPACY, obteve resultado igual para PES-AUTOR (0,59) e foi melhor para a entidade SENTENÇA (0,29). Ao realizar uma análise sintética dos resultados obtidos pelo melhor modelo (BERT), percebe-se que grande parte dos erros são de predições de *tokens* do tipo "O"(formato IOB, (RAMSHAW; MARCUS, 1999)). Esse tipo de ocorrência correspondeu a quase 30% dos erros verificados. Nas entidades que formam categorias, também, observa-se erros por ambiguidade de entidades da mesma categoria: Na categoria Endereço os erros de ambiguidade chegam a 16%, em Norma 8% e em Pessoa os erros de ambiguidade na mesma categoria alcançam 16%.

Para o aprofundamento das análises realizada, fez-se uso da matriz de confusão, que é uma tabela usada para avaliar o desempenho de um modelo de classificação. Ela resume os resultados mostrando as contagens de verdadeiros positivos (VP), verdadeiros negativos

(VN), falsos positivos (FP) e falsos negativos (FN). Esses valores ajudam a avaliar a Acurácia, Precisão, Revocação e Medida-F de um modelo, que são métricas importantes para compreender o desempenho da classificação das diferentes categorias. a Figura 11 apresenta a matriz de confusão para modelo BERT no cenário C1.

A seguir, será analisado o melhor resultado obtido por cada EN no Cenário 1 e as análises comparativas dos demais cenários (C2, C3, C4 e C5).

Tabela 5 – Resultados de Medida-F para o reconhecimento das entidades refinadas (C1) utilizando os modelos BI-LSTM+CRF, SPACY e BERT

Entidade Nomeada	BI-LSTM+CRF	SPACY	BERT	Suporte
END-AUTOR	**0.56**	0.31	0.33	18
END-DELITO	0.72	0.45	**0.73**	61
END-OUTROS	0.00	0.02	0.16	81
END-REU	0.55	0.59	**0.71**	152
END-TESTEMUNHA	0.27	0.26	**0.67**	68
END-VITIMA	0.06	0.00	**0.22**	27
NOR-ACESSÓRIA	0.79	0.79	**0.82**	990
NOR-JURISPRUDÊNCIA	**0.90**	0.87	0.89	333
NOR-PRINCIPAL	0.67	0.71	**0.77**	791
PENA	**0.56**	0.39	0.50	82
PES-ADVOG	0.54	0.22	**0.63**	122
PES-AUTOR	**0.59**	**0.59**	0.56	169
PES-AUTORID-POLICIAL	0.87	0.66	**0.90**	300
PES-JUIZ	**0.79**	0.50	0.78	83
PES-OUTROS	0.54	0.44	**0.58**	1.210
PES-PROMOTOR-MP	0.81	0.27	**0.88**	57
PES-REU	0.64	0.57	**0.71**	1.503
PES-TESTEMUNHA	0.57	0.45	**0.64**	519
PES-VÍTIMA	0.33	0.23	**0.46**	405
PROVA	**0.47**	0.29	0.34	461
SENTENÇA	0.00	**0.29**	0.00	11
MMIMF	0.64	0.55	**0.67**	7.443
MMAMF	0.53	0.42	**0.58**	7.443
MPMF	0.62	0.54	**0.67**	7.443

4.3.1 Análise dos Resultados para o Cenário 1.

O modelo BERT alcançou o melhor desempenho, ainda assim, na análise detalhada se pode constatar que teve maior dificuldade na desambuiguação de entidade do tipo "O". Estes tipos de *tokens* representaram mais de 60% das predições. Os FN do tipo "O"representaram mais

da metade desse tipo de erro. Abaixo, apresenta-se uma análise dos resultados das entidades para o C1.

END-AUTOR: O modelo BI-LSTM + CRF obteve o melhor desempenho, com Medida-F de 0,56. As predições apresentam uma Precisão moderada devido o alto índice de falsos positivos. As predições são confundidas, principalmente, com com *tokens* não-anotados ("O") ou END-REU. A Revocação é razoável (0,56). A pequena quantidade de exemplos pode contribur para a baixa performance do modelo para essa EN.

END-DELITO: Esta EN foi a de melhor desempenho na categoria, alcançando a Medida-F de 0,73. Esse resultado veio pela excelente Revocação (0,93), porém, a Precisão foi moderada (0,59). As predições são confundidas, principalmente, com *tokens* não-anotados ("O") ou END-REU.

END-OUTROS: Esta EN foi a de pior resultado no REN para os endereços. A Precisão e a Revocação foram muito baixas (0,13 e 0,20, respectivamente), tendo uma altíssima quantidade de falsos positivos (87%) do tipo "O".

END-REU: Esta entidade obteve muita harmonia entre a Precisão (0,70) e a Revocação (0,72). A maior quantidade de erros acontece entre as entidades da mesma categoria.

END-TESTEMUNHA: Alcança boa Revocação (0,71), porém a Precisão é moderada (0,63). As predições erradas são, em sua maioria, entre as entidades de mesma categoria, mas também *tokens* "O".

END-VITIMA: Essa EN teve o desempenho muito baixo (Medida-F de 0,22). Precisão (0,44) e Revocação (0,15) foram baixas, com muitos FP e FN da mesma categoria.

Resumo da Análise para Endereço: Apesar de algumas EN refinadas apresentarem baixa Medida-F, o REN obteve bom desempenho nessa categoria, alcançando Medida-F de 0,72. Este resultado se explica por vários erros de predição ocorrerem entre entidades da mesma categoria e, pelo fato das EN de pior desempenho contarem com poucos exemplos (suporte).

NOR-ACESSÓRIA: O REN obteve excelente desempenho em reconhecer essa entidade (Medida-F de 0,82). O seu desempenho foi melhor em função da Revocação elevada (0,86), tendo a maior parte dos FN apontados na Norma Principal e nos *tokens* tipo "O". A Precisão foi mais baixa (0,79), principalmente, pelos FP do tipo "O"e classificados como Norma Principal.

NOR-JURISPRUDÊNCIA: Esta norma é a que o NER, com o modelo BI-LSTM + CRF, obtenhe melhor desempenho (Medida-F de 0,90). A Precisão alcançada de 0,86, mostra erros de FP em *tokens* sem marcação ("O") e, curiosamente, em PES-OUTROS. Uma explicação para esse erro pode estar relacionada à possibilidade de que o nome do relator da norma seja

frequentemente encontrado nas anotações de jurisprudência. A Revocação alcançou 0,95, sendo os poucos FN de *tokens* tipo "O".

NOR-PRINCIPAL: Apresenta um bom desempenho (Medida-F 0,77). A Precisão de 0,72 puxou a Medida-F para baixo. As predições se confundem muito com as outras entidades da mesma categoria mas, principalmente, com os *tokens* sem anotação ("O"). Quanto a Revocação (0,84), a maioria dos erros estão nas entidades da mesma categoria.

Figura 11 – Matriz de Confusão do modelo BERT no Cenário C1

Real \ Predicto	END_AUTOR	END_DELITO	END_OUTROS	END_REU	END_TESTEMUNHA	END_VITIMA	NOR_ACESSÓRIA	NOR_JURISPRUDÊNCIA	NOR_PRINCIPAL	PENA	PES_ADVOG	PES_AUTOR	PES_AUTORID_POLICIAL	PES_JUIZ	PES_OUTROS	PES_PROMOTOR_MP	PES_REU	PES_TESTEMUNHA	PES_VITIMA	PROVA	SENTENÇA	O
END_AUTOR	**12**	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
END_DELITO	0	**57**	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
END_OUTROS	3	1	**16**	8	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48
END_REU	8	9	3	**110**	8	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
END_TESTEMUNHA	0	2	2	11	**48**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
END_VITIMA	0	1	1	8	7	**4**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
NOR_ACESSÓRIA	0	0	0	0	0	0	**856**	3	81	0	0	0	0	0	0	0	0	0	0	0	0	50
NOR_JURISPRUDÊNCIA	0	0	0	0	0	0	2	**329**	1	0	0	0	0	0	0	0	0	0	0	0	0	1
NOR_PRINCIPAL	0	0	0	0	0	0	98	2	**666**	0	0	0	0	0	0	0	0	0	0	0	0	25
PENA	0	0	0	0	0	0	0	0	0	**35**	0	0	0	0	0	0	0	0	0	0	0	47
PES_ADVOG	0	0	0	0	0	0	0	0	0	0	**84**	1	0	1	0	0	0	0	0	0	0	36
PES_AUTOR	0	0	0	0	0	0	0	0	0	0	0	**112**	0	0	1	0	4	0	1	0	0	51
PES_AUTORID_POLICIA	0	0	0	0	0	0	0	0	0	0	0	0	**280**	0	6	0	2	5	0	0	0	7
PES_JUIZ	0	0	0	0	0	0	0	0	0	0	0	0	0	**75**	1	0	0	0	0	0	0	7
PES_OUTROS	1	0	0	0	0	0	0	11	0	0	13	21	2	1	**673**	0	120	37	53	2	0	276
PES_PROMOTOR_MP	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	**50**	0	0	0	0	0	5
PES_REU	0	1	0	0	0	0	0	0	0	0	0	17	0	0	33	0	**1078**	43	105	18	0	208
PES_TESTEMUNHA	0	0	0	0	1	0	0	0	0	0	0	0	2	0	64	0	30	**320**	22	0	0	80
PES_VITIMA	0	0	0	0	0	0	0	0	0	0	0	0	1	0	10	0	97	16	**190**	11	0	80
PROVA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	7	0	**162**	0	290
SENTENÇA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	11
O	31	25	99	17	8	0	133	62	182	23	47	76	38	33	324	6	212	59	47	307	0	**466261**

Fonte**:** Elaborado pelo autor.

Resumo da Análise para Norma: Esta categoria foi a que obteve a mais alta Medida-F (0,90). A Precisão é de 86%, sendo os erros mais comuns as predições de *tokens* tipo "O". A Revocação alcançou 0,96, com a maioria dos FN de entidades da mesma categoria.

Pena: A Precisão do modelo está moderada (0,60), sendo a maior quantidade de erros (FP) de predições de tipo "O". A Revocação é ruim, alcançando 0,43, com os *tokens* "O"indicados como FN. A medida-F de 0,56 foi obtida com o modelo BI-LSTM + CRF. Os resultados, possivelmente,

foram prejudicados pela pequena quantidade de exemplos disponíveis.

PES-ADVOG: A Precisão foi de 58%, com o modelo confundindo *tokens* não-anotados ("O") e com PES-OUTROS. A Revocação alcançou 0,69, sendo os *tokens* "O"a maioria dos FN.

PES-AUTOR: As predições desta entidade tiveram muitas ambiguidades com entidades da mesma categoria e FP do tipo "O", levando a Precisão a 49%. Já a Revocação de 0,66 se caracterizou por muitos FN do tipo "O". A Medida-F de 0,59 foi obtida com os modelo BI-LSTM + CRF e SPACY.

PES-AUTORID-POLICIAL: Esta EN é a que o REN obteve melhor desempenho, com Medida-F de 0,90. A Precisão alcançada de 0,87, mostram erros de FP em *tokens* sem marcação ("O") e alguns poucos da mesma categoria. A Revocação foi excelente (0,93), sendo a maior parte dos FN da própria categoria.

PES-JUIZ: Apresenta um bom desempenho (Medida-F 0,79) com o modelo BI-LSTM + CRF. A Precisão de 0,83 se caracterizou por erros nos *tokens* "O". A Revocação (0,79) reduziu a Medida-F e, os erros estão nas entidades de mesma categoria e , também, nos *tokens* sem anotação ("O").

PES-OUTROS: Obteve resultados bastante intrigantes, pois a Precisão e a Revocação foram, apenas, moderadas, apesar dessa EN ter muitos exemplos anotados. A Precisão de 0,60 se caracteriza por muitos FP da mesma categoria, mas principalmente, de *tokens* tipo "O". A Revocação, também, foi mediana, alcançando 0,56, com muitos FN da mesma categoria, do tipo "O"e até em NOR-JURISPRUDÊNCIA.

PES-PROMOTOR-MP: Entidade obteve excelente resultado no REN, com Medida-F de 0,88. A Precisão alcançada de 0,89, mostra poucos erros de FP em *tokens* sem marcação ("O"). A Revocação, também, foi excelente (0,88), com alguns poucos FN da própria categoria e de *tokens* sem anotação ("O").

PES-REU: Esta EN obteve excelente desempenho (Medida-F 0,71) e harmonia entre Precisão (0,70) e Revocação (0,72). A maior quantidade de FP é semelhante entre entidades de mesma categoria e tipo "O". Da mesma forma, os FN também são equilibrados entre as entidades de mesma categoria e os *tokens* tipo "O".

PES-TESTEMUNHA: Com Medida-F de 0,64, essa entidade teve muitos erros de predições da mesma categoria (PES-REU, PES-OUTROS, PES-VITIMA). Mas, também, confundio com os *tokens* tipo "O". Com isso a Precisão ficou moderada (0,66). A maior quantidade de FN foi semelhante ao ocorrido com a Precisão, porém a maior quantidade de erros ocorreu com os

tokens sem anotação, levando a Revocação a 0,62.

PES-VITIMA: Esta entidade foi a de pior desempenho na categoria (Medida-F 0,46), apesar da boa quantidade de exemplos disponíveis desta entidade (405 anotações de suporte). A maior quantidade de erros de Predição (FP) estão entre as entidades da mesma categoria, porém, há muitos erros com *tokens* não anotados. Com isso, a Precisão foi, apenas, 0,45. A Revocação foi de 0,47, com muitos FN da mesma categoria, de *tokens* tipo "O"e alguns da entidade Prova.

Resumo da Análise para Pessoa: Esta categoria alcançou Medida-F de 0,81, que é um resultado muito bom. A maior quantidade de FP e FN estão nos *tokens* tipo "O", porém, há uma quantidade quase equivalentes dos mesmos erros entre entidades da mesma categoria.

Vale destacar os bons resultados de classificação alcançados pelas entidades PES-AUTORID-POLICIAL, PES-PROMOTOR-MP, PES-JUIZ e PES-ADVOG. Estas entidades não foram as que dispuseram de mais anotações para treinar o REN. No entando, foram as que alcançaram os mais altos índices da Medida-F. Em contraste com PES-VÍTIMA, PES-TESTEMUNHA e PES-OUTROS que dispunham de muitas anotações e, no entando tiveram o pior desempenho no REN. Uma hipótese que explica esses resultados é o fato das entidades melhor classificadas aparecem mais contextualizadas nos documentos jurídicos. Por exemplo, às citações a juízes ocorrem com muita frequência nos finais das peças processuais, também é comum os seus nomes serem precedidos por pronomes pessoais de tratamento ou mesmo por sua profissão. Algo semelhante ocorre para às autoridades policiais, promotores e advogados citados nas peças. Por sua vez, às menções às vítimas, testemunhas e outras pessoas citadas ocorrem em diferentes partes dos documentos sendo comum as substituições dos nomes próprios por referências usando pronomes.

Prova: O REN não obteve bom desempenho para esta entidade (Medida-F 0,47). A Precisão foi excelente (0,87), com poucos FP, que ocorreram mais nos *tokens* "O". A Revocação ficou baixa, em 0,33, com FN em entidades de Normas, Prova, Pessoa e, principalmente, nos *tokens* "O".

Sentença: O modelo REN com o SPACY foi o único a conseguir identificar essa entidade. Ainda assim, obteve baixíssima Revocação (0,18), sendo a maioria dos FN de *tokens* tipo "O". O modelo fez poucas predições, mas teve bom resultado (0,67). A Medida-F final ficou em 0,29. Os resultados, possivelmente, foram prejudicados pela pequena quantidade de exemplos disponíveis.

Tabela 6 – Resultados da Medida-F para o REN na CDJUR-BR e LENER-BR (C2 a C5) utilizando o modelo BERT

Entidade	Cenário de Experimento			
	C2	C3	C4	C5
JURISPRUDÊNCIA	0.89	0.96	0.79	0.48
LEGISLAÇÃO	0.92	0.97	0.92	0.86
LOCAL	0.77	0.77	0.32	0.15
PESSOA	0.83	0.97	0.69	0.76
MMIMF	0.85	0.96	0.81	0.60
MMAMF	0.85	**0.92**	0.68	0.56
MPMF	0.85	0.96	0.79	0.74

4.3.2 Análise dos Resultados para os Cenários Comparativos com LENER-BR (C2, C3, C4 e C5).

Os experimentos destes cenários foram realizados com o modelo baseado no BERT. Em uma comparação direta de C2 com C3, obseva-se que os resultados de C2 são inferiores aos obtidos em C3. Dito com outras palavras, os resultados do REN treinado e testado com a CDJUR-BR são inferiores aos resultados quando se treina e se testa com o LENER-BR. A maior diversidade de documentos que compõem a CDJUR-BR, o desbalanceamento de algumas categorias (Endereço, Prova, Pena e Sentença) e algumas entidades com poucos exemplos, e ainda, a ambiguidade entre entidades podem ter contribuído para o seu desempenho inferior. Quando se compara o desempenho nos cenários C4 e C5, em que o modelo é treinado na CDJUR-BR e testado com o LENER-BR, atinge um MMAMF de 0,68, enquanto no cenário C5, treinado com o LENER-BR e testado com a CDJUR-BR, o modelo alcança um MMAMF de apenas 0,56. Esse resultado pode indicar que a CDJUR-BR tem maior capacidade de adaptabilidade para reconhecer entidades de outras coleções de documentos legais. A Tabela 6 apresenta os resultados obtidos para os cenários de 2 a 5 (C2 a C5).

5 CONCLUSÃO

Nesta dissertação, foi apresentada uma metodologia própria de anotação manual de documentos jurídicos, contendo diretrizes e exemplos de anotações, que serviu para criar a Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas (CDJUR-BR). A coleção é formada pelas classes semânticas Pessoa, Prova, Pena, Endereço, Sentença e Norma, dispondo de 44.526 anotações realizadas para 21 entidades nomeadas distintas. A avaliação da concordância entre anotadores alcançou medida Kappa de 0,69 para 73% dos documentos, e os demais documentos passaram por revisões por especialistas e etapas de refinamento. Foram realizados experimentos na tarefa de reconhecimento de entidades nomeadas com os modelos SPACY, BI-LSTM + CRF e BERT. Os resultados apontaram superioridade do modelo BERT com a Média Macro da Medida-F geral de 0,58 e testes comparativos entre CDJUR-BR e LENER-BR indiciaram que o modelo REN treinado com a CDJUR-BR é superior em reconhecer entidades de outra coleção.

A metodologia de anotação proposta apresenta contribuições significativas que a tornam adaptável a diferentes tarefas de anotação de corpus. Uma particularidade importante é que em suas etapas há a previsão de definição de critérios que possibilitam a adequação das atividades aos objetivos específicos do esforço de anotação. Essa flexibilidade intrínseca permite que a metodologia seja adaptada para projetos que empregam a abordagem de *crowdsourcing* (SABOU *et al.*, 2014), a qual é amplamente utilizada na anotação de grandes corpus para o processamento de linguagem natural. Ademais, a metodologia prevê definições claras e inequívocas das categorias e subcategorias de entidades nomeadas, juntamente com instruções precisas sobre o que deve e o que não deve ser anotado para as diversas entidades definidas. A inclusão de exemplos de anotações também facilita a compreensão dos anotadores, o que é fundamental para garantir a consistência e a qualidade das anotações. Embora inicialmente aplicada no domínio jurídico, as características da metodologia demonstram sua relevância e utilidade em outros domínios, tais como os campos da medicina, biologia, literatura, ciências sociais, entre outros, como evidenciado nos trabalhos de (KRALLINGER *et al.*, 2015; JOVANOVIĆ; BAGHERI, 2017; CAMPOS; MATOS; OLIVEIRA, 2012; NEVES; LESER, 2014; BAMMAN; POPAT; SHEN, 2019).

5.1 LIMITAÇÕES E TRABALHOS FUTUROS

Dada as características íntrisicas de cada documento que compuseram o corpus da CDJUR-BR, algumas entidades tiveram poucos exemplos anotados. Através da análises dos experimentos realizados, percebe-se que isto impactou consideralvemente no desempenho dos modelos desenvolvidos de REN. Isso sugere melhorias para aumentar a quantidade de exemplos anotados de algumas entidades, como END-AUTOR, END-VITIMA, PENA e SENTENÇA, propiciando balancear estas categorias e reduzir o impacto dos dados nos modelos REN.

Os resultados também mostram que não existe um modelo REN universal que reconheça todas as entidades da melhor maneira. Portanto, um classificador formado por vários modelos poderia ser construído para alcançar melhores resultados em categorias ou entidades nomeadas específicas.

Adicionalmente, percebe-se que comparar a CDJUR-BR somente com o LENER-BR apresenta limitações significativas. Essa abordagem pode levar a conclusões restritas e enviesadas, pois ambos os *corpora* podem compartilhar um viés específico e limitar a representatividade dos dados. A falta de diversidade nos *corpora* prejudica a análise da variação linguística em textos jurídicos, diminuindo a precisão das conclusões. Além disso, a ausência de referências adicionais impede a obtenção de *insights* e abordagens alternativas. Para obter resultados e generalizações dos dados mais confiáveis, em trabalhos futuros será incluída uma comparação mais ampla e diversificada com outros *corpora* relevantes e abordagens metodológicas adicionais.

Com o advento dos LLMs (Large Language Models) (BROWN *et al.*, 2020) e alto poder gerativo das IAs como o ChaGPT, há que se avaliar o impacto destes modelos na tarefa de REN, analisando a necessidade e escopo de coleções douradas para domínios específicos como o jurídico. Embora os LLMs tenham a capacidade de gerar textos coerentes e relevantes, é fundamental considerar a sua confiabilidade e acurácia no reconhecimento de entidades em domínios especializados como o jurídico, pois seu desempenho no REN ainda está significativamente abaixo dos trabalhos de referências com modelos supervisionados (WANG *et al.*, 2023). A análise futura ajudará a determinar que entidades específicas deverão ser anotadas e em que quantidades necessárias, uma vez que os LLMs reduzem de forma considerável a necessidade de dados rotulados. Coleções douradas específicas, como a CDJUR-BR, poderão contribuir para aprimorar modelos REN criados com LLMs.

5.2 PUBLICAÇÃO DECORRENTE DESTA PESQUISA

- BRITO, Maurício; PINHEIRO, Vládia; FURTADO, Vasco; MONTEIRO NETO, João Araújo; BOMFIM, Francisco das Chagas Jucá; DA COSTA, André Câmara Ferreira; SILVEIRA, Raquel. CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 177-186. DOI: https://doi.org/10.5753/stil.2023.234217.

REFERÊNCIAS

ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G.; SOUZA, E.; SILVA, N. F. da; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A. *et al.* Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In: SPRINGER. **Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings**. [S.l.], 2022. p. 3–14.

ANGELIDIS, I.; CHALKIDIS, I.; KOUBARAKIS, M. Named entity recognition, linking and generation for greek legislation. In: **JURIX**. [S.l.: s.n.], 2018. p. 1–10.

ARAUJO, P. H. L. de; CAMPOS, T. E. de; OLIVEIRA, R. R. de; STAUFFER, M.; COUTO, S.; BERMEJO, P. Lener-br: a dataset for named entity recognition in brazilian legal text. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 313–323.

ATDAĞ, S.; LABATUT, V. A comparison of named entity recognition tools applied to biographical texts. In: IEEE. **2nd International conference on systems and computer science**. [S.l.], 2013. p. 228–233.

BAMMAN, D.; POPAT, S.; SHEN, S. An annotated dataset of literary entities. 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)**. [S.l.: s.n.], 2019. p. 2138–2144.

BOELLA, G.; CARO, L. D.; LEONE, V. Semi-automatic knowledge population in a legal document management system. **Artificial intelligence and Law**, Springer, v. 27, n. 2, p. 227–251, 2019.

BORDO, M. D.; ROCKOFF, H. The gold standard as a "good housekeeping seal of approval". **The Journal of Economic History**, Cambridge University Press, v. 56, n. 2, p. 389–428, 1996.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. *et al.* Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

CAMPOS, D.; MATOS, S.; OLIVEIRA, J. L. Biomedical named entity recognition: a survey of machine-learning tools. **Theory and applications for advanced text mining**, InTech Rijeka, Croatia, v. 11, p. 175–195, 2012.

CASTRO, P. V. Q. d. *et al.* Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico. Universidade Federal de Goiás, 2019.

CEJUELA, J. M.; MCQUILTON, P.; PONTING, L.; MARYGOLD, S. J.; STEFANCSIK, R.; MILLBURN, G. H.; ROST, B.; CONSORTIUM, F. *et al.* tagtog: interactive and text-mining-assisted annotation of gene mentions in plos full-text articles. **Database**, Oxford Academic, v. 2014, 2014.

CHOWDHURY, G. G. Natural language processing. **Annual review of information science and technology**, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

CNJ, C. N. de J. **Relatório Justiça em Números**. [S.l.]: Conselho Nacional de Justiça-Brasília,2023, 2023. 326 p.

COAN, E. I. Atributos da linguagem jurídica. Revista dos Tribunais, 2003.

COATES-STEPHENS, S. The analysis and acquisition of proper names for the understanding of free text. **Computers and the Humanities**, Springer, v. 26, p. 441–456, 1992.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

FIGUEIREDO, G. S. Projeto athos: um estudo de caso sobre a inserção do superior tribunal de justiça na era da inteligência artificial. 2022.

FLEISCHMAN, M. Automated subcategorization of named entities. In: CITESEER. **ACL (Companion Volume)**. [S.l.], 2001. p. 25–30.

FLEISCHMAN, M.; HOVY, E. Fine grained classification of named entities. In: **COLING 2002: The 19th International Conference on Computational Linguistics**. [S.l.: s.n.], 2002.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. In: **XXIII Congresso da Sociedade Brasileira de Computaçao**. [S.l.: s.n.], 2003. v. 3, p. 347–395.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural networks**, Elsevier, v. 18, n. 5-6, p. 602–610, 2005.

GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**. [S.l.: s.n.], 1996.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017.

HOVY, E.; LAVID, J. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. **International journal of translation**, v. 22, n. 1, p. 13–36, 2010.

HUANG, A. *et al.* Similarity measures for text document clustering. In: **Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand**. [S.l.: s.n.], 2008. v. 4, p. 9–56.

HUANG, W.; HU, D.; DENG, Z.; NIE, J. Named entity recognition for chinese judgment documents based on bilstm and crf. **EURASIP Journal on Image and Video Processing**, SpringerOpen, v. 2020, n. 1, p. 1–14, 2020.

JIANG, J. Information extraction from text. In: **Mining text data**. [S.l.]: Springer, 2012. p. 11–41.

JIANG, R.; BANCHS, R. E.; LI, H. Evaluating and combining name entity recognition systems. In: **Proceedings of the Sixth Named Entity Workshop**. [S.l.: s.n.], 2016. p. 21–27.

JOVANOVIĆ, J.; BAGHERI, E. Semantic annotation in biomedicine: the current landscape. **Journal of biomedical semantics**, BioMed Central, v. 8, n. 1, p. 1–18, 2017.

KANAPALA, A.; PAL, S.; PAMULA, R. Text summarization from legal documents: a survey. **Artificial Intelligence Review**, Springer, v. 51, n. 3, p. 371–402, 2019.

KLIE, J.-C.; BUGERT, M.; BOULLOSA, B.; CASTILHO, R. E. de; GUREVYCH, I. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: **Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations**. [S.l.: s.n.], 2018. p. 5–9.

KRALLINGER, M.; RABAL, O.; LEITNER, F.; VAZQUEZ, M.; SALGADO, D.; LU, Z.; LEAMAN, R.; LU, Y.; JI, D.; LOWE, D. M. *et al.* The chemdner corpus of chemicals and drugs and its annotation principles. **Journal of cheminformatics**, BioMed Central, v. 7, n. 1, p. 1–17, 2015.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

LEE, S.; LEE, G. G. Heuristic methods for reducing errors of geographic named entities learned by bootstrapping. In: SPRINGER. **International Conference on Natural Language Processing**. [S.l.], 2005. p. 658–669.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. A dataset of german legal documents for named entity recognition. **arXiv preprint arXiv:2003.13016**, 2020.

LI, J.; SUN, A.; HAN, J.; LI, C. A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 34, n. 1, p. 50–70, 2020.

MANNING, C.; SCHUTZE, H. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MARTINS, A. D. M. Agrupamento automático de documentos jurídicos com uso de inteligência artificial. IDP/EDAB, 2018.

MCHUGH, M. L. Interrater reliability: the kappa statistic. **Biochemia medica**, Medicinska naklada, v. 22, n. 3, p. 276–282, 2012.

MIKHEEV, A.; MOENS, M.; GROVER, C. Named entity recognition without gazetteers. In: **Ninth Conference of the European Chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 1999. p. 1–8.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, John Benjamins, v. 30, n. 1, p. 3–26, 2007.

NEVES, M.; LESER, U. A survey on annotation tools for the biomedical literature. **Briefings in bioinformatics**, Oxford University Press, v. 15, n. 2, p. 327–340, 2014.

PEIXOTO, F. H. Projeto victor: relato do desenvolvimento da inteligência artificial na repercussão geral do supremo tribunal federal. **Revista Brasileira de Inteligência Artificial e Direito-RBIAD**, v. 1, n. 1, p. 1–22, 2020.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **arXiv preprint arXiv:2010.16061**, 2020.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. **Natural language processing using very large corpora**, Springer, p. 157–176, 1999.

ROSSET, S.; GALIBERT, O.; ADDA, G.; BILINSKI, E. The limsi participation in the qast track. In: SPRINGER. **Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 8**. [S.l.], 2008. p. 414–423.

SABOU, M.; BONTCHEVA, K.; DERCZYNSKI, L.; SCHARL, A. Corpus annotation through crowdsourcing: Towards best practice guidelines. In: CITESEER. **LREC**. [S.l.], 2014. p. 859–866.

SANTOS, D.; CARDOSO, N. A golden resource for named entity recognition in portuguese. In: SPRINGER. **International Workshop on Computational Processing of the Portuguese Language**. [S.l.], 2006. p. 69–79.

SCHMITT, X.; KUBLER, S.; ROBERT, J.; PAPADAKIS, M.; LETRAON, Y. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In: IEEE. **2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.], 2019. p. 338–343.

SEKINE, S.; NOBATA, C. Definition, dictionaries and tagger for extended named entity hierarchy. In: LISBON, PORTUGAL. **LREC**. [S.l.], 2004. p. 1977–1980.

SILVA, R. L. d.; HOCH, P. A.; RIGHI, L. M. Transparência pública e a atuação normativa do cnj. **Revista direito GV**, SciELO Brasil, v. 9, p. 489–514, 2013.

SINGH, S. Natural language processing for information extraction. **arXiv preprint arXiv:1807.02383**, 2018.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2020. p. 403–417.

SURDEN, H. Artificial intelligence and law: An overview. **Georgia State University Law Review**, v. 35, p. 19–22, 2019.

TOMANEK, K.; WERMTER, J.; HAHN, U. An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data. In: **Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)**. [S.l.: s.n.], 2007. p. 486–495.

WANG, S.; SUN, X.; LI, X.; OUYANG, R.; WU, F.; ZHANG, T.; LI, J.; WANG, G. Gpt-ner: Named entity recognition via large language models. **arXiv preprint arXiv:2304.10428**, 2023.

WISSLER, L.; ALMASHRAEE, M.; DÍAZ, D. M.; PASCHKE, A. The gold standard in corpus annotation. In: **IEEE GSC**. [S.l.: s.n.], 2014.

WYNER, A. Z.; PETERS, W.; KATZ, D. A case study on legal case annotation. In: **JURIX**. [S.l.: s.n.], 2013. p. 165–174.

YADAV, V.; BETHARD, S. A survey on recent advances in named entity recognition from deep learning models. **arXiv preprint arXiv:1910.11470**, 2019.

YAMADA, H.; TEUFEL, S.; TOKUNAGA, T. Building a corpus of legal argumentation in japanese judgement documents: towards structure-based summarisation. **Artificial Intelligence and Law**, Springer, v. 27, n. 2, p. 141–170, 2019.

ZHONG, H.; XIAO, C.; TU, C.; ZHANG, T.; LIU, Z.; SUN, M. How does nlp benefit legal system: A summary of legal artificial intelligence. **arXiv preprint arXiv:2004.12158**, 2020.