Universidade de Brasília
Departamento de Estatística

Aprendizado de Máquina e Redes Neurais para Reconhecimento de Entidades Nomeadas em Portarias Jurídicas via Processamento de Linguagem Natural

Davi Esmeraldo da Silva Albuquerque

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Davi Esmeraldo da Silva Albuquerque

Aprendizado de Máquina e Redes Neurais para Reconhecimento de Entidades Nomeadas em Portarias Jurídicas via Processamento de Linguagem Natural

Orientador: Prof. Eduardo Monteiro de Castro

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Sumário

1 Introdução

O avanço das técnicas de Processamento de Linguagem Natural (PLN) tem transformado significativamente a maneira como grandes volumes de textos são analisados em diversos domínios, incluindo o setor jurídico. Isso ocorre porque ferramentas baseadas em PLN permitem a extração de informações valiosas de documentos complexos, promovendo maior automação e eficiência nos processos administrativos. Contudo, a análise e classificação de textos jurídicos, bem como a extração automática de informações estruturadas, continuam sendo desafiadoras devido às especificidades da linguagem jurídica e à heterogeneidade dos documentos oficiais (GOCHHAIT, 2024).

Nesse contexto, modelos de aprendizado de máquina e redes neurais profundas emergem como soluções promissoras, demonstrando alto desempenho em tarefas complexas relacionadas à compreensão de linguagem natural. Essas técnicas têm o potencial de transformar dados não estruturados em conhecimento estruturado, o que facilita o acesso a informações relevantes e aumenta a agilidade nos processos administrativos e jurídicos (OLIVEIRA; NASCIMENTO, 2021).

Além disso, sabe-se que instituições jurídicas geram uma quantidade significativa de documentos oficiais, os quais representam fontes ricas para análises estratégicas e fundamentação de decisões. Nesse cenário, ao perceber essa oportunidade de atuação, o objetivo deste trabalho é aplicar e avaliar modelos baseados em aprendizado de máquina e redes neurais profundas para realizar as tarefas de Reconhecimento de Entidades Nomeadas (REN), agrupamento e classificação de portarias do Tribunal de Justiça do Distrito Federal e Territórios (TJDFT).

A aplicação de técnicas de agrupamento ou clusterização, nesse sentido, permite a criação de novas representações dos dados, o que facilita sua interpretação, tanto por humanos quanto por outros algoritmos de aprendizagem de máquina, visto que tais representações são mais interpretáveis quando comparadas às originais (Müller e Guido, 2016, apud FREITAS, 2023). Essas técnicas possibilitam segmentar documentos com base em similaridades temáticas ou características intrínsecas, otimizando o gerenciamento da informação e permitindo a identificação de padrões ocultos.

A classificação de documentos textuais, por sua vez, é outra abordagem amplamente utilizada nesse campo. Ao empregar modelos treinados, essa técnica viabiliza a categorização de documentos em classes específicas. Sua flexibilidade e alta precisão tornam-na uma ferramenta de grande relevância para a automação de processos documentais e a priorização de análises jurídicas.

Já o Reconhecimento de Entidades Nomeadas (REN) se destaca por sua capacidade de identificar e categorizar informações específicas, como nomes de pessoas, orga-

nizações e datas, presentes nos textos jurídicos. Esta técnica tem grande aplicação prática na extração automatizada de, por exemplo, cláusulas contratuais e jurisprudências, promovendo análises mais rápidas e precisas. No entanto, sua eficácia está intimamente ligada à adaptação de modelos às particularidades da linguagem jurídica, que frequentemente contém terminologias especializadas e nuances semânticas. Além disso, é válido destacar que iniciativas para a criação de corpora jurídicos anotados têm sido fundamentais para o aprimoramento desses modelos supervisionados, contribuindo para o aumento de sua precisão.

Diante do exposto, por meio da aplicação deste trabalho, espera-se que a automação de processos textuais contribua diretamente para a tomada de decisões estratégicas no setor jurídico. A utilização de técnicas como Reconhecimento de Entidades Nomeadas (REN), agrupamento e classificação para a análise de portarias visa promover maior eficiência, acessibilidade e transparência nos processos judiciais e administrativos. Esses avanços não apenas ampliam a produtividade dos profissionais da área, mas também fortalecem a confiabilidade e a robustez das decisões, ao transformar dados não estruturados em informações estruturadas e de fácil acesso.

2 Objetivos

2.1 Objetivo Geral

Este trabalho tem como objetivo principal aplicar e avaliar modelos baseados em aprendizado de máquina e redes neurais profundas para a realização das tarefas de Reconhecimento de Entidades Nomeadas (REN), agrupamento e classificação de documentos textuais. O foco está na análise de portarias recentes emitidas pelo Gabinete da Presidência do Tribunal de Justiça do Distrito Federal e Territórios (TJDFT).

Ademais, busca-se transformar dados não estruturados em informações organizadas e acessíveis, por meio da utilização de técnicas de Processamento de Linguagem Natural adaptadas ao contexto jurídico brasileiro.

Desse modo, espera-se que os resultados contribuam para o avanço da eficiência no gerenciamento de documentos oficiais, aprimorando a extração de informações relevantes e estruturadas. Esses desenvolvimentos viabilizam a otimização de processos administrativos e judiciários, fortalecendo a tomada de decisões informadas no sistema judiciário.

2.2 Objetivos Específicos

Referente aos objetivos específicos, para o cumprimento deste trabalho, será realizado um levantamento detalhado dos dados necessários. Esta etapa inicial incluirá o acesso ao site oficial do TJDFT, onde será realizada a extração de informações relevantes das portarias, bem como uma análise exploratória preliminar com o objetivo de identificar padrões e estruturas características dos documentos.

Finalizada a coleta de dados, objetiva-se submetê-los a um pré-processamento por meio de etapas essenciais de organização e padronização textual, garantindo que as informações estejam adequadas para a captura das nuances semânticas presentes nos textos jurídicos, assegurando uma base sólida para a aplicação dos modelos. Em sequência, busca-se aplicar a conversão de dados textuais em representações numéricas. Essas representações vetoriais capturam as relações semânticas e contextuais, facilitando a análise dos textos jurídicos e melhorando a precisão dos modelos aplicados.

Posteriormente, o desenvolvimento de modelos para o Reconhecimento de Entidades Nomeadas (REN), utilizando redes neurais profundas, visa identificar categorias-chave presentes nas portarias, como datas, nomes de partes envolvidas e tópicos jurídicos relevantes. Dito isso, tem-se essa identificação como fundamental para tornar o processo de análise e extração de informações mais eficiente e detalhado, facilitando o acesso e a

compreensão dos documentos.

Para auxiliar e enriquecer a etapa de Reconhecimento de Entidades Nomeadas (REN), será aplicada uma abordagem de agrupamento não supervisionado das portarias, utilizando algoritmos de aprendizado de máquina não supervisionados. Essa etapa visa não apenas segmentar os documentos em grupos com características semânticas semelhantes, mas também identificar possíveis categorias emergentes que poderão servir de base para o aprendizado supervisionado posterior. Além disso, esse processo de agrupamento auxiliará na detecção preliminar de padrões e possíveis entidades nomeadas que mereçam destaque na fase de reconhecimento automático.

Tem-se também como propósito aplicar modelos supervisionados de classificação das portarias. As técnicas de clusterização realizadas a priori permitirão uma análise inicial das portarias, enquanto os modelos supervisionados serão treinados para categorizar os documentos em classes pré-definidas.

Por fim, neste trabalho, propõe-se a realização de uma avaliação do desempenho dos modelos aplicados. Além disso, busca-se identificar as limitações das abordagens empregadas, com o objetivo de sugerir melhorias e indicar direções para pesquisas futuras, ampliando o impacto e a aplicabilidade das soluções no contexto jurídico brasileiro.

Essa abordagem integrada e detalhada visa garantir que o estudo produza resultados consistentes, contribuindo para o avanço das técnicas de Processamento de Linguagem Natural no gerenciamento de portarias jurídicas oficiais.

3 Revisão Bibliográfica

Neste capítulo, serão descritos os métodos estatísticos e computacionais empregados neste trabalho. Esses métodos foram selecionados devido à sua relevância no contexto dos dados analisados, bem como à sua capacidade de contribuir para alcançar os objetivos estabelecidos na seção anterior. A fundamentação teórica e a escolha das técnicas aplicadas baseiam-se, em grande parte, nos resultados de revisões sistemáticas, as quais serviram como base para a revisão do estado da arte e para a seleção das abordagens adotadas.

A revisão sistemática realizada por (BARROS et al., 2024), por exemplo, incluiu consultas a diversas bases científicas renomadas e seus objetivos centrais foram baseados em quatro pilares fundamentais: identificar as publicações científicas mais relevantes sobre a aplicação de PLN em documentação jurídica nos últimos anos, mapear as técnicas e ferramentas de PLN utilizadas no tratamento de documentos no domínio jurídico, avaliar o desempenho dessas aplicações em novos documentos jurídicos brasileiros e verificar quais bases de dados jurídicas existentes no Brasil possuem algum pré-processamento que facilite a aplicação de PLN.

Com base nas informações descritas em revisões sistemáticas, neste trabalho são priorizadas técnicas e ferramentas de PLN amplamente validadas na literatura científica recente e ajustadas ao contexto dos dados jurídicos brasileiros. Paralelamente, foram realizadas pesquisas nas temáticas relacionadas aos objetivos do estudo, aprofundando a compreensão e identificando abordagens inovadoras que complementassem as análises. Essa priorização e a realização de pesquisas visaram garantir a robustez metodológica e a relevância dos resultados obtidos, assegurando a aplicabilidade e eficácia das análises realizadas neste estudo.

3.1 Aprendizado de Máquina

A aprendizagem de máquina (*Machine Learning*), ramo da inteligência artificial, concentra-se no desenvolvimento de algoritmos capazes de aprender com experiências anteriores e tomar decisões baseadas em dados, minimizando a necessidade de interferência humana. Essa abordagem, situada na interseção entre computação científica, inteligência artificial (*IA*) e estatística, tem-se mostrado eficaz na solução de problemas complexos, como classificação de dados, reconhecimento de padrões e previsões (FREITAS, 2023).

Tipos de Aprendizado

Os métodos de aprendizado de máquina podem ser classificados em três categorias principais, de acordo com o tipo de supervisão empregada no treinamento dos modelos.

3.1.1 Aprendizado Supervisionado

O primeiro tipo é o *Aprendizado Supervisionado*, no qual o modelo é treinado com um conjunto de dados rotulados, ou seja, para cada entrada existe uma saída esperada. De maneira geral, como descrito por (FREITAS, 2023) o algoritmo precisa ser capaz de gerar respostas para novas entradas sem intervenção humana, baseando-se unicamente na experiência adquirida e nas regras desenvolvidas a partir de um conjunto inicial de dados de treinamento.

Além disso, é reforçado por (FREITAS, 2023) que os problemas supervisionados de aprendizado de máquina podem ser classificados em dois tipos principais, que são a classificação e a regressão. Na classificação, o objetivo é prever a categoria de uma observação entre várias opções, enquanto a regressão tem como objetivo a previsão de valores contínuos.

Fluxo de processamento de aprendizado de máquina supervisionado

Um algoritmo de aprendizagem de máquina segundo (Lantz,2013, apud Freitas et al, 2024) segue uma série de etapas estruturadas, começando pela coleta de dados, que envolve a obtenção, organização e preparação das informações necessárias para o treinamento do modelo.

Em seguida, realiza-se a análise exploratória, etapa crucial para avaliar a qualidade dos dados. Isso inclui o tratamento de dados faltantes, a verificação de correlações entre variáveis e a visualização de padrões iniciais, como agrupamentos e outliers.

Após essa fase, ocorre a crucial etapa de treinamento dos modelos, a qual consiste em ajustar os algoritmos com base nos dados de treinamento. Para garantir a generalização dos modelos, adota-se uma estratégia de divisão de dados que inclui subconjuntos para treino, validação e teste. O conjunto de treino é utilizado para ajustar os pesos e parâmetros dos modelos. O conjunto de validação auxilia no ajuste de hiperparâmetros, como a escolha da taxa de aprendizado ou do número de neurônios em uma rede neural, enquanto o conjunto de teste é reservado para avaliar o desempenho final do modelo em dados não vistos, simulando o comportamento em um ambiente real.

Após o treinamento, a avaliação do modelo é realizada, com a aplicação de

métricas específicas para medir seu desempenho. Essa avaliação envolve testar o modelo fora do conjunto de treinamento e calcular os resultados de interesse. Comumente, os modelos são comparados utilizando métricas como precisão (*precision*), abrangência ou revocação (*recall*) e *F1-Score*. Essas métricas são definidas como:

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{3.1.1}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3.1.2}$$

$$\textbf{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.1.3}$$

O conceito de Verdadeiro Positivo (TP, True Positive) refere-se à situação em que o modelo identifica corretamente que a classe é positiva (1) e, ao verificar o rótulo real, confirma que a classe é, de fato, positiva. Por outro lado, o Verdadeiro Negativo (TN, True Negative) ocorre quando o modelo classifica corretamente a classe como negativa, e ao comparar com o rótulo real, verifica-se que a classe é, de fato, negativa. O Falso Positivo (FP, False Positive) é identificado quando o modelo erroneamente classifica a classe como positiva, mas ao verificar a resposta real, constata-se que a classe é negativa. Finalmente, o Falso Negativo (FN, False Negative) ocorre quando o modelo prediz a classe como negativa, mas ao verificar a resposta, descobre-se que a classe era positiva como descrito por (FREITAS, 2023).

Quando todas essas etapas são concluídas, o modelo está pronto para ser utilizado, seja para a descrição ou previsão de novos dados, fornecendo informações claras e úteis aos usuários finais. (Lantz,2013, apud Freitas et al, 2024)

3.1.2 Aprendizado Não Supervisionado

O aprendizado não supervisionado é uma abordagem do aprendizado de máquina em que os modelos são treinados utilizando dados não rotulados, ou seja, sem informações explícitas sobre os resultados esperados. Diferentemente do aprendizado supervisionado, o objetivo principal é explorar e identificar padrões, estruturas ou agrupamentos (*clusters*) que estão intrinsecamente presentes nos dados. Essa técnica é amplamente utilizada em cenários onde os rótulos não estão disponíveis ou sua obtenção seria inviável devido ao alto custo ou tempo necessário para rotular grandes volumes de dados.

Por sua natureza exploratória, o aprendizado não supervisionado desempenha um papel crucial na descoberta de conhecimento e na compreensão inicial de bases de dados, fornecendo insights que muitas vezes guiam decisões e estratégias em diversas áreas de aplicação.

K-Means

O algoritmo *K-Means* é uma técnica amplamente utilizada em aprendizado de máquina não supervisionado para o agrupamento (*clusterização*) de dados em k grupos distintos. Seu funcionamento baseia-se na minimização da soma das distâncias quadráticas entre cada ponto de dado e o centróide do grupo ao qual pertence.

O processo de operação do *K-Means* segue uma sequência de etapas bem definidas. Inicialmente, são escolhidos k centróides, que podem ser selecionados de maneira aleatória. Em seguida, cada ponto do conjunto de dados é atribuído ao centróide mais próximo, geralmente considerando a distância euclidiana como métrica de proximidade.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.1.4}$$

Após essa atribuição, os centróides são atualizados calculando-se a média dos pontos associados a cada grupo, conforme descrito na equação:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \tag{3.1.5}$$

Nessa equação, μ_j representa o centróide do grupo j, C_j é o conjunto de pontos pertencentes ao grupo j, e x_i são os pontos que compõem o grupo.

As etapas de atribuição e atualização são repetidas iterativamente até que ocorra a convergência. Isso pode ser identificado quando os centróides permanecem praticamente inalterados entre iterações ou quando é atingido um número máximo de iterações predefinido. Essa abordagem permite a identificação de agrupamentos nos dados, otimizando a separação entre os k grupos.

Método do Cotovelo (Elbow Method)

Uma abordagem popularmente utilizada para determinar o número ideal de grupos (k) no algoritmo *K-Means* é o **Método do Cotovelo** (I.; OBUNADIKE, 2022). Este método baseia-se na análise da soma das distâncias quadráticas dentro dos grupos

(*Within-Cluster Sum of Squares*, ou WCSS) para diferentes valores de k. O WCSS mede a compactação dos grupos formados, representando a soma das distâncias entre cada ponto de dado e o centróide de seu respectivo grupo.

A ideia principal do *Método do Cotovelo* é encontrar o ponto em que o aumento no número de grupos não resulta em uma redução significativa no WCSS. Para isso, calcula-se o WCSS para um intervalo de valores de k, imprimindo-se esses valores em um gráfico, com o eixo x representando os valores de k e o eixo y representando o WCSS correspondente.

A curva gerada pelo gráfico geralmente apresenta uma queda acentuada no início, devido à melhora significativa na compactação dos grupos. No entanto, em determinado ponto, a redução no WCSS começa a diminuir de forma menos expressiva. Este ponto de inflexão, que se assemelha a um "cotovelo"na curva, indica o número ótimo de grupos, pois representa um equilíbrio entre a compactação dos grupos e a complexidade do modelo.

Matematicamente, o WCSS para um número k de grupos pode ser definido como:

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \tag{3.1.6}$$

onde C_j representa o conjunto de pontos no grupo j, x_i é um ponto pertencente a C_j, e μ_j é o centróide do grupo j. A escolha do valor de k no ponto do cotovelo garante um bom compromisso entre a qualidade do agrupamento e a simplicidade do modelo, evitando o superdimensionamento do número de grupos.

Similaridade Cosseno (Cosine Similarity)

Para superar a falha de mensuração de distância pela distância euclidiana em dados textuais, em (OLIVEIRA; NASCIMENTO, 2021), a distância nos agrupamentos, por exemplo, é avaliada por meio da similaridade cosseno, uma medida que calcula o cosseno do ângulo entre dois vetores projetados em um plano multidimensional. O valor resultante dessa medida varia entre 0 e 1, sendo que 1 indica que os vetores são totalmente similares e 0 indica que são completamente diferentes. Dada a definição de dois vetores, X e Y, a similaridade cosseno pode ser expressa pelo produto escalar, conforme apresentado na Equação:

$$\text{similaridade} = \cos(\theta) = \frac{X \cdot Y}{|X| \cdot |Y|} \tag{3.1.7}$$

Correlato a isso, em (FREITAS, 2023) a similaridade por cosseno é empregada para classificar processos similares ao comparar representações vetoriais de textos. No caso

específico, a abordagem consistiu em identificar, para cada texto não rotulado, o texto etiquetado mais similar com base na similaridade por cosseno, replicando as etiquetas correspondentes.

3.1.3 Aprendizado Profundo

O aprendizado profundo, uma subárea do aprendizado de máquina, refere-se a métodos de modelagem computacional que utilizam redes neurais artificiais com múltiplas camadas (também chamadas de camadas ocultas) para aprender representações de dados de forma hierárquica. O aprendizado profundo se tornou fundamental em várias áreas, como visão computacional, processamento de linguagem natural, reconhecimento de fala e muitas outras, devido ao seu desempenho superior em tarefas de grande escala.

Perceptron

O perceptron é um dos primeiros modelos de rede neural e foi proposto como um classificador binário. Ele é composto por uma única camada de unidades de processamento, chamadas de neurônios, que estão conectados entre si de forma semelhante a uma rede neural. O perceptron recebe entradas, aplica uma função de ativação e gera uma saída. Cada entrada é multiplicada por um peso, que é ajustado durante o treinamento, permitindo que o perceptron aprenda a classificar dados. Embora simples, o perceptron foi a base para o desenvolvimento de redes neurais mais complexas.

Principais Arquiteturas de Redes Profundas

O aprendizado profundo se tornou eficaz graças à evolução de várias arquiteturas de redes neurais. A seguir, são exploradas as principais arquiteturas que dominaram a área de aprendizado profundo.

Redes Convolucionais (CNN's)

As redes neurais convolucionais (CNNs) são projetadas para processar dados com uma estrutura em grid, como imagens. Elas são compostas por camadas convolucionais que aplicam filtros (ou kernels) aos dados de entrada para extrair características como bordas, texturas e padrões complexos. De acordo com (FREITAS, 2023), as redes convolucionais possuem uma arquitetura feedforward, com fluxos de informações em uma única direção, e são conhecidas por sua capacidade de serem invariantes a deslocamento ou invariantes no espaço. Essas redes operam em pequenos lotes de informações, utilizando

pesos compartilhados de camada para camada, o que permite uma redução no número de parâmetros necessários para o processamento. Além disso, camadas de pooling simplificam as informações das camadas anteriores, o que torna as CNNs eficientes e rápidas para o reconhecimento de imagens. Elas têm se mostrado extremamente eficazes em tarefas de visão computacional, como reconhecimento de objetos e classificação de imagens.

Redes Recorrentes (RNN's)

As redes neurais recorrentes (RNNs) são projetadas para processar dados sequenciais, onde a ordem das entradas é importante, como em textos, séries temporais e discursos. Elas possuem conexões cíclicas que permitem que informações sejam passadas de uma etapa para outra, tornando-as adequadas para modelar dependências temporais e sequenciais. De acordo com (FREITAS, 2023), as RNNs se diferenciam das redes feed-forward por incluírem loops de realimentação, o que cria uma memória entre as camadas da rede, tornando os modelos mais dinâmicos. Esse mecanismo de realimentação pode ocorrer até que uma regra de treinamento seja atingida, o que é particularmente útil em tarefas que requerem apenas informações recentes. No entanto, à medida que a rede cresce e as informações necessárias para a atualização dos neurônios estão distantes da iteração atual, as RNNs podem perder a capacidade de conectar essas informações de maneira eficaz, o que pode reduzir a capacidade das funções de ativação de destacar informações relevantes.

3.2 Processamento de Linguagem Natural (PLN)

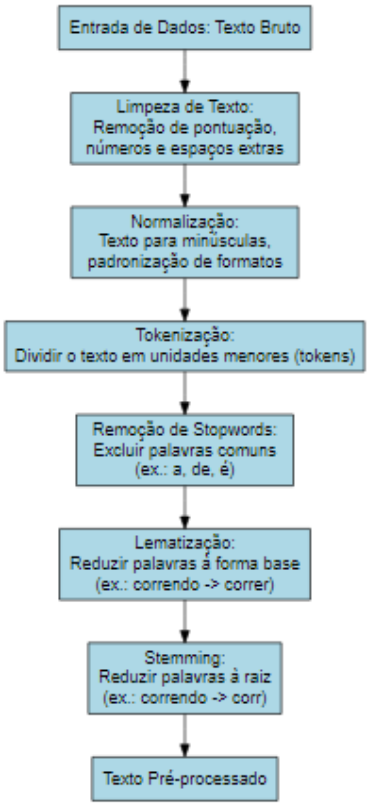
O Processamento de Linguagem Natural (PLN), ou Natural Language Processing (NLP), é uma área de pesquisa vinculada à Inteligência Artificial (IA) que busca desenvolver métodos e sistemas capazes de processar a linguagem humana de forma computacional. Esse campo concentra-se em compreender, interpretar e gerar linguagem natural, destacando-se por seu foco nas línguas faladas pelos humanos (CASELI; NUNES, 2024).

3.2.1 Pré Processamento

O pré-processamento de dados textuais é uma etapa essencial na aplicação de técnicas de inteligência artificial, como o *Processamento de Linguagem Natural (PLN)* (OLIVEIRA; NASCIMENTO, 2021). Seu objetivo principal é transformar dados brutos em um formato mais limpo e estruturado, eliminando ruídos e facilitando a modelagem computacional ou estatística. Esse processo pode ser dividido em várias etapas, cada uma contribuindo para melhorar a qualidade dos textos analisados.

O fluxograma a seguir sintetiza as etapas de pré-processamento de dados textuais, proporcionando uma visão clara do processo e das operações envolvidas.

Figura 1: Fluxograma de Processamento de Linguagem Natural



Fonte: Confecção pelo autor.

A primeira etapa envolve a remoção de caracteres especiais e números que não possuem relevância para a análise. Elementos como pontuação, símbolos ou dígitos isolados são descartados para que o texto fique mais limpo e uniforme. Isso reduz o impacto de informações irrelevantes que poderiam influenciar negativamente os resultados de modelos aplicados posteriormente.

A etapa fundamental subsequente é a *normalização*, que visa garantir maior consistência nos dados textuais. Isso inclui a transformação de todas as palavras para letras minúsculas (*lowercasing*), assegurando que variações de capitalização, como "Casa"e "casa", sejam tratadas como equivalentes.

A próxima etapa, conhecida como *tokenização*, consiste em dividir o texto em unidades menores chamadas de *tokens*. Esses tokens podem ser palavras, frases ou até caracteres, dependendo da necessidade do projeto. Por exemplo, um parágrafo pode ser transformado em uma lista de palavras individuais, facilitando a análise de seu conteúdo.

Após a tokenização, realiza-se a *remoção de stopwords*, que elimina palavras de pouco valor semântico, como preposições e artigos ("de", "para", "o", "a"). Embora comuns, essas palavras não contribuem significativamente para a análise e podem gerar ruído. Removê-las resulta em um texto mais focado e relevante para o objetivo da aplicação.

A lematização desempenha um papel importante ao reduzir as palavras à sua forma base, ou *lema*. Por exemplo, palavras como "correr", "corre"e "correndo"são todas transformadas na forma "correr". Essa simplificação é útil para evitar redundâncias no modelo.

A etapa de *stemming* também é essencial, complementando a lematização ao reduzir as palavras às suas raízes. Diferente da lematização, o stemming pode gerar formas incompletas ou truncadas das palavras, mas igualmente pode contribuir para a redução da dimensionalidade e para a melhoria da performance dos modelos.

Com a aplicação dessas etapas, o texto passa de uma forma desestruturada para um formato que favorece a extração de informações relevantes. Esse processo de pré-processamento é essencial para o sucesso de modelos de inteligência artificial, pois garante que os dados utilizados sejam de alta qualidade e representativos das análises pretendidas.

3.2.2 Representação Numérica de Palavras:

No *Processamento de Linguagem Natural (PLN)*, a representação de palavras é um passo crucial, pois permite transformar dados textuais em formas numéricas que algoritmos podem processar como vetores ou matrizes de números (FREITAS, 2023).

TF-IDF (Term Frequency-Inverse Document Frequency)

O *TF-IDF* (Term Frequency-Inverse Document Frequency) é uma técnica amplamente utilizada para vetorização de textos, a qual transforma documentos em representações numéricas que podem ser usadas por algoritmos de aprendizado de máquina. Ele combina duas abordagens: a frequência de um termo dentro de um documento e a inversa da frequência de documentos em que o termo aparece. A ideia é que termos que ocorrem com frequência em um documento, mas são raros no restante da coleção de documentos, são considerados mais informativos.

A fórmula para o cálculo do *TF-IDF* é dada por:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right) \tag{3.2.1}$$

onde:

- $\mathrm{TF}(t, d)$ é a frequência do termo t no documento d, que indica o número de vezes que o termo aparece no documento. - $\mathrm{DF}(t)$ é a frequência de documentos contendo o termo t. - N é o número total de documentos.

O *TF-IDF* tem como principal vantagem identificar termos que são únicos para um determinado documento, o que o torna útil para a representação de textos em tarefas como classificação e clusterização. Em contrapartida, como abardado por (FREITAS, 2023) uma limitação da técnica TF-IDF é que ela pode atribuir vetores similares a palavras com distribuições semelhantes no corpus, mesmo que essas palavras não possuam proximidade semântica. Da mesma forma, palavras com significados próximos podem não estar próximas no espaço vetorial devido à natureza esparsa dos vetores gerados.

Word Embeddings

Uma das abordagens mais avançadas e amplamente utilizadas é o uso de *Word Embeddings*, que projetam palavras em espaços vetoriais contínuos, preservando suas relações semânticas e sintáticas.

Os *Word Embeddings* mapeiam palavras para vetores densos em um espaço de menor dimensão, comumente em $\mathbb{R}^d$, onde d é o número de dimensões do vetor. Essa representação captura similaridades semânticas entre palavras.

Um exemplo amplamente utilizado é o **Word2Vec**, desenvolvido por (MIKOLOV et al., 2013), que oferece duas arquiteturas principais para a modelagem de palavras: o **Continuous Bag of Words (CBOW)** e o **Skip-gram**. O CBOW prevê uma palavra com base no contexto de palavras vizinhas, utilizando a seguinte fórmula:

$$P(w_t \mid w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}) \tag{3.2.2}$$

Por outro lado, o *Skip-gram* prevê o contexto de uma palavra dada a palavra central, representado pela equação:

$$P(w_{t-c}, \ldots, w_{t+c} \mid w_t) \tag{3.2.3}$$

Já o **Wang2Vec**, extensão do *Word2Vec*, introduz informações de ordem na sequência de palavras, adicionando pesos para dependências sequenciais. Essa característica permite que o modelo capture melhor as relações posicionais entre as palavras

em um contexto dado. A função objetivo do *Wang2Vec* é definida como:

$$\mathcal{L} = -\sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log P(w_{t+j} | w_t, \text{ordem}(j)) \tag{3.2.4}$$

onde ordem(j) adiciona pesos que representam a influência da posição relativa j nas dependências sequenciais entre palavras.

Outra abordagem popular é o **GloVe (Global Vectors for Word Representation)**, proposto por (PENNINGTON; SOCHER; MANNING, 2014), que utiliza uma técnica baseada em matrizes de coocorrência. De forma resumida, o *Glove* cria representações vetoriais de palavras ao considerar a matriz global de coocorrência entre termos em um corpus, conforme discutido em (CâNDIDO, 2020).

A função de perda minimizada pelo *GloVe* é:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(\mathbf{v}_i^\top \mathbf{u}_j + b_i + b_j - \log(X_{ij}) \right)^2 \tag{3.2.5}$$

Além disso, o **FastText**, introduzido por (JOULIN et al., 2016), é uma extensão do *Word2Vec* que melhora a qualidade das representações de palavras, especialmente para palavras raras ou morfologicamente complexas. Enquanto o *Word2Vec* trata palavras como unidades indivisíveis, o *FastText* leva em consideração subpalavras (ou n-grams de caracteres) para gerar as representações vetoriais. As principais vantagens do *FastText* incluem a capacidade de lidar com palavras desconhecidas e a captura aprimorada da morfologia das palavras, o que se revela particularmente útil em línguas com alta inflexão, como o português, ou quando se trabalha com vocabulários técnicos e jargões. Essa abordagem permite uma modelagem mais robusta e flexível, especialmente em contextos linguísticos complexos.

A representação vetorial de uma palavra w no *FastText* é dada pela soma dos vetores das suas subpalavras n-gram:

$$\mathbf{v}_w = \sum_{n \in N(w)} \mathbf{v}_n \tag{3.2.6}$$

onde $N(w)$ é o conjunto de n-grams de caracteres para a palavra w, e $\mathbf{v}_n$ é o vetor correspondente ao n-gram n.

Em suma, a escolha da técnica mais adequada para a geração de embeddings de palavras exige um processo cuidadoso de investigação, experimentação e comparação de diferentes modelos. O uso desses modelos tem-se mostrado essencial para melhorar os resultados em tarefas de inteligência artificial, como detecção de padrões e classificação de dados, especialmente quando se trabalha com grandes volumes de dados não estruturados (OLIVEIRA; NASCIMENTO, 2021).

3.3 Reconhecimento de Entidades Nomeadas (REN

O *Reconhecimento de Entidades Nomeadas (REN)*, ou Named Entity Recognition (NER), é uma técnica fundamental dentro do campo de Processamento de Linguagem Natural (PLN), voltada para a identificação e classificação automática de entidades mencionadas em textos. Essas entidades podem pertencer a diversas categorias, como pessoas, organizações, locais, datas, valores monetários, entre outras. No contexto jurídico, por exemplo, o REN se destaca ao identificar entidades como nomes de juízes, advogados e outros profissionais do direito em documentos legais, atribuindo um nível adicional de semântica aos dados extraídos. Este tipo de análise torna-se essencial para a compreensão e organização de grandes volumes de informações jurídicas, otimizando o processo de extração e interpretação de dados significativos (CASTRO, 2019).

3.3.1 Corpus e Corpora

No contexto do Reconhecimento de Entidades Nomeadas (REN), o termo *corpus* refere-se a um conjunto estruturado de textos utilizado como base para o treinamento ou avaliação de modelos de Processamento de Linguagem Natural (PLN). No plural, *corpora* designa múltiplos conjuntos de textos, que são essenciais para o desenvolvimento e a validação desses modelos. Esses dados oferecem exemplos rotulados de entidades, permitindo que o modelo aprenda a identificar padrões e contextos relevantes.

Nos últimos anos, diversos corpora foram desenvolvidos especificamente para treinar e avaliar modelos de REN, cada um com características e focos distintos, voltados para diferentes áreas de aplicação. Um exemplo relevante é o **LeNER-Br** (ARAUJO et al., 2018), um corpus voltado para o português brasileiro, especialmente projetado para trabalhar com textos jurídicos e administrativos. Outro corpus importante é o **WikiNER** (NOTHMAN et al., 2013), um corpus multilíngue extraído da Wikipédia, amplamente utilizado para treinar modelos de REN mais generalistas, aplicáveis a uma vasta gama de textos. Além disso, o **Paramopama** é um corpus especializado em domínios técnicos e científicos, com foco particular em entidades mais complexas. O **HAREM**, por sua vez, concentra-se no português europeu e é amplamente utilizado como um *benchmark* para

avaliar a eficácia dos modelos de REN desenvolvidos para esse idioma. Esses corpora desempenham um papel crucial no avanço da precisão e aplicabilidade dos sistemas de reconhecimento de entidades nomeadas.

A construção de modelos eficientes para Reconhecimento de Entidades Nomeadas depende diretamente da qualidade e representatividade do corpus utilizado. O corpus, ao fornecer a base textual para o treinamento, precisa ser cuidadosamente preparado e anotado, assegurando que as entidades de interesse sejam claramente identificadas. Esse preparo meticuloso é essencial para que os modelos possam generalizar e alcançar um bom desempenho em diversos contextos. Na sequência, será abordada a utilização de redes neurais no REN, explorando sua capacidade de aprender representações complexas a partir de dados textuais, impulsionando avanços significativos nesta área.

3.3.2 BERT (Bidirectional Encoder Representations from Transformers)

O **BERT (Bidirectional Encoder Representations from Transformers)**, apresentado por (DEVLIN et al., 2019), é amplamente reconhecido por sua capacidade de capturar contextos bidirecionais de maneira eficaz, tornando-o particularmente útil em tarefas como o *Reconhecimento de Entidades Nomeadas (REN)*. Baseado na arquitetura de Transformers, o BERT considera o contexto completo de uma palavra dentro de uma sentença, analisando tanto os elementos que a precedem quanto os que a sucedem. Essa característica o diferencia de outros modelos e é essencial para resolver ambiguidades em textos complexos, proporcionando uma compreensão mais precisa do significado das palavras em diferentes contextos.

Durante o treinamento, o BERT combina duas tarefas principais. A primeira é a Máscara de Palavras (*Masked Language Modeling - MLM*), que treina o modelo a prever palavras ocultas com base no restante da sentença. A segunda é a Predição de Sentenças (*Next Sentence Prediction - NSP*), que avalia a conexão lógica entre pares de sentenças. De acordo com (COSTA, 2023), no contexto do REN, o processo de *fine-tuning* envolve a adaptação dos pesos do modelo para que ele seja capaz de identificar corretamente as entidades nomeadas em textos específicos, tornando-o uma ferramenta poderosa para áreas como os setores jurídico, médico ou financeiro.

O modelo também é altamente robusto ao lidar com palavras raras ou compostas, graças às representações baseadas em subpalavras, o que garante um desempenho superior em cenários com vocabulários diversificados ou pouco frequentes. Além disso, sua capacidade de adaptação por meio de *fine-tuning* amplia sua aplicabilidade, permitindo atender a demandas específicas de áreas com terminologias particulares.

Outra vantagem significativa do BERT é sua versatilidade em tarefas de *Proces-*

samento de Linguagem Natural em geral, especialmente ao lidar com dados de setores especializados. Isso o torna uma ferramenta poderosa tanto para aplicações gerais quanto para domínios específicos, consolidando sua relevância no campo de aprendizado profundo.

3.3.3 RoBERTa (Robustly Optimized BERT Pretraining Approach)

O **RoBERTa (Robustly Optimized BERT Pretraining Approach)** é uma versão aprimorada do BERT, proposta por (LIU et al., 2019). Essa melhoria se concentra no processo de pré-treinamento, resultando em um desempenho superior em diversas tarefas de Processamento de Linguagem Natural (PLN), incluindo o *Reconhecimento de Entidades Nomeadas (REN)*, análise de sentimento e resposta a perguntas (COSTA, 2023).

Uma das alterações mais notáveis do RoBERTa é o uso extensivo de dados durante o treinamento, utilizando um volume muito maior de corpora, como *Common Crawl News* e *OpenWebText*, o que expande a capacidade do modelo de capturar padrões linguísticos complexos. Outra inovação importante é o ajuste na aplicação da máscara, que é agora feita de forma dinâmica durante a execução. Isso assegura maior variação nas entradas mascaradas, promovendo um aprendizado mais eficiente. Por fim, o RoBERTa adota tamanhos de lote maiores e taxas de aprendizado ajustadas, permitindo um treinamento mais robusto e eficiente em comparação ao BERT original.

Entre suas principais vantagens, destaca-se o desempenho superior em domínios específicos, resultado do uso de dados mais diversos e abrangentes durante o treinamento. Além disso, o modelo demonstra maior robustez em contextos complexos, sendo capaz de lidar de forma eficiente com textos longos. Outra característica relevante é sua excelente capacidade de generalização, o que permite sua aplicação eficaz em novas tarefas e domínios, mantendo um desempenho de alta qualidade.

Assim como o BERT, o RoBERTa pode ser ajustado (*fine-tuned*) para tarefas de REN por meio da adição de uma camada de classificação. Essa camada prevê rótulos para cada token na sentença de entrada. Devido ao seu pré-treinamento mais robusto, o RoBERTa geralmente alcança melhor desempenho em tarefas como identificação de entidades, incluindo pessoas, locais e organizações.

3.3.4 Long Short-Term Memory (LSTM)

O modelo de *Long Short-Term Memory (LSTM)* é uma evolução das redes neurais recorrentes (*Recurrent Neural Networks - RNNs*), introduzido por (HOCHREITER; SCHMIDHUBER, 1997). Ele foi projetado para superar limitações das RNNs tradicionais, especialmente em relação à dificuldade de aprender dependências de longo prazo em sequências de dados. Essa capacidade torna o LSTM uma escolha apropriada para uma

variedade de tarefas que exigem processamento sequencial, como o *Reconhecimento de Entidades Nomeadas (REN)*.

A arquitetura do LSTM se destaca por suas *células de memória*, que controlam o fluxo de informações por meio de três portas principais. A porta de esquecimento (f_t) define quais informações da célula de memória anterior devem ser descartadas, a porta de entrada (i_t) decide quais novas informações serão incorporadas, e a porta de saída (o_t) determina quais informações serão usadas para gerar a saída. Esse formato permite que o LSTM mantenha informações relevantes ao longo de longas sequências de texto, melhorando sua eficácia em tarefas que requerem análise contextual.

No contexto de REN, as LSTMs processam sequências de texto representadas por vetores, como *word embeddings*, que servem como entrada para a rede. A saída gerada pela LSTM é frequentemente conectada a uma camada de *softmax*, que prevê os rótulos das entidades para cada token da sequência. Essa abordagem possibilita a identificação de padrões e dependências contextuais cruciais para a tarefa.

Embora as LSTMs ofereçam vantagens, como a capacidade de capturar dependências de longo alcance e sua compatibilidade com representações pré-treinadas, como *word embeddings*, elas apresentam desafios. Entre eles, destacam-se o custo computacional elevado em corpora grandes e a competição com modelos mais recentes, baseados em *transformers*, como BERT e RoBERTa, que muitas vezes oferecem desempenho superior em tarefas mais complexas.

3.3.5 Bidirectional Long Short-Term Memory - BiLSTM

As Redes Neurais Recorrentes Bidirecionais (*Bidirectional Long Short-Term Memory - BiLSTM*) são uma extensão das LSTMs, projetadas para processar sequências de dados em ambas as direções, capturando informações contextuais tanto do passado quanto do futuro para determinar a etiqueta no tempo t (COSTA, 2023). Por essa razão, essa abordagem é especialmente útil em tarefas como o *Reconhecimento de Entidades Nomeadas (REN)*, onde o contexto completo de uma palavra é essencial para a correta determinação da entidade.

No BiLSTM, duas redes LSTM independentes são treinadas: a LSTM direta ($\overrightarrow{h_t}$), que processa a sequência do início para o fim, e a LSTM reversa ($\overleftarrow{h_t}$), que processa a sequência na direção oposta, do fim para o início. A saída do BiLSTM para cada posição t é obtida pela concatenação dos estados ocultos das duas direções, expressa como:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t}$$

Uma vantagem adicional do BiLSTM é que sua arquitetura se combina facilmente com técnicas complementares, como *word embeddings* e *Conditional Random Fields* (CRFs). Outro ponto positivo é de que dado sua habilidade de lembrar a ordem das informações apresentadas, o BiLSTM é capaz de filtrar informações irrelevantes ou redundantes, mantendo apenas os elementos mais importantes, o que contribui para reduzir ruídos e aprimorar a qualidade dos resultados em tarefas como classificação de texto (GOCHHAIT, 2024).

3.3.6 Conditional Random Fields (CRF)

O *Conditional Random Fields (CRF)* (LAFFERTY; MCCALLUM; PEREIRA, 2001) é um modelo amplamente utilizado para tarefas de rotulagem sequencial, como o *Reconhecimento de Entidades Nomeadas (REN)*, e frequentemente complementa o BiLSTM para melhorar a precisão ao considerar a estrutura sequencial dos rótulos. O CRF modela a probabilidade condicional de uma sequência de rótulos $Y = \{y_1, y_2, \ldots, y_n\}$ dada uma sequência de entradas $X = \{x_1, x_2, \ldots, x_n\}$, conforme:

$$P(Y|X) = \frac{\exp\left(\sum_{t=1}^{n} \psi(y_t, y_{t-1}, X)\right)}{\sum_{Y'} \exp\left(\sum_{t=1}^{n} \psi(y'_t, y'_{t-1}, X)\right)} \tag{3.3.1}$$

onde $\psi(y_t, y_{t-1}, X)$ é uma função de pontuação que avalia as dependências entre rótulos adjacentes.

Ao ser conectado à saída de um BiLSTM, o CRF ajusta as pontuações fornecidas pela rede, levando em conta as relações estruturais entre os rótulos. A pontuação total de uma sequência Y é expressa como:

$$S(X, Y) = \sum_{t=1}^{n} \psi(y_t, y_{t-1}, X) + \sum_{t=1}^{n} s_t(y_t) \tag{3.3.2}$$

onde $s_t(y_t)$ são as pontuações obtidas na saída do BiLSTM. O CRF busca maximizar $S(X, Y)$, escolhendo a sequência de rótulos mais provável.

A combinação de BiLSTM e CRF constitui uma abordagem robusta para tarefas de REN, capturando tanto dependências contextuais quanto estruturais nas sequências de texto. Apesar do crescente domínio de modelos baseados em *transformers*, como BERT e RoBERTa, a combinação BiLSTM-CRF continua sendo uma solução competitiva em muitos cenários (COSTA, 2023).

4 Trabalhos Relacionados

O trabalho de (GOCHHAIT, 2024) propôs a aplicação de um modelo baseado em BiLSTM para a categorização de dados textuais e previsão das taxas de sobrevivência dos passageiros do Titanic, utilizando o Titanic Disaster Dataset. A pesquisa comparou o desempenho do BiLSTM com algoritmos tradicionais, como Naive Bayes (NB), Gradient Boosting (GB) e Support Vector Machine (SVM). Os resultados demonstraram que o BiLSTM obteve uma precisão de 98,5%, superando as abordagens tradicionais. Este resultado destaca o potencial do BiLSTM em tarefas de classificação, especialmente por sua capacidade de eliminar informações irrelevantes enquanto preserva a estrutura hierárquica dos dados textuais.

De maneira semelhante, o estudo de (CâNDIDO, 2020) analisou diferentes arquiteturas de redes neurais para a classificação de texto, incluindo CNNs, VDCNNs, HANs e BERT, comparando-as a métodos tradicionais como o SVM. A pesquisa indicou que, embora as técnicas de aprendizado profundo tenham vantagens em grandes conjuntos de dados, elas apresentam desempenho inferior em dados pequenos devido a dificuldades de generalização. O BERT demonstrou melhorias em classificação, embora não tenha superado significativamente o VDCNN. Além disso, o estudo evidenciou o trade-off entre eficácia e custo, destacando que as abordagens mais recentes ainda carecem de eficiência para aplicações industriais em larga escala.

Já o estudo de (FREITAS, 2023) abordou a expansão da base de rótulos dos Objetivos de Desenvolvimento Sustentável (ODS) da Agenda 2030 da ONU, utilizando métodos de agrupamento de textos jurídicos. A pesquisa explorou a clusterização para gerar rótulos sintéticos e melhorar a classificação de textos desbalanceados em relação aos rótulos dos ODS. A investigação evidenciou que a clusterização, como técnica de aumento de dados, foi mais eficaz que a anotação manual, equilibrando as classes e aprimorando a classificação, especialmente quando combinada com redes LSTM e modelos ensemble como o CatBoost.

No que tange ao reconhecimento de entidades nomeadas (REN), o estudo de (CASTRO, 2019) focou em modelos de REN baseados em redes neurais profundas, utilizando ELMo como modelo de linguagem contextual. O trabalho demonstrou que o ELMo, combinado com outras técnicas como CNN e Wang2Vec, superou os benchmarks existentes, alcançando um F-Score de 93,81% no domínio jurídico trabalhista brasileiro e 83,22% no português geral. O treinamento com corpora maiores, como o brWaC, também melhorou o desempenho, destacando a superioridade do ELMo em cenários específicos.

Além disso, (COSTA, 2023) ampliou o corpus UlyssesNER-Br para REN, incorporando textos informais, como comentários sobre projetos de lei em português do Brasil,

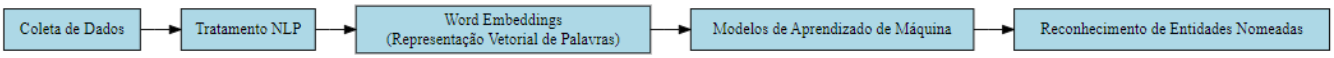
e textos formais. O estudo revelou que a combinação de textos formais e informais contribuiu para melhorar a identificação de entidades em textos informais. O modelo BERT, ajustado para o C-corpus, obteve o melhor desempenho, com um F1-score de 78,65%, validando a hipótese de que a integração de textos formais e informais melhora a performance dos sistemas REN. No entanto, embora o BERT tenha se mostrado superior, modelos tradicionais como CRF e BiLSTM-CRF também apresentaram resultados satisfatórios, destacando sua eficácia mesmo em comparação com abordagens baseadas em modelos de linguagem modernos. A hipótese de que modelos baseados em arquiteturas transformer, como o BERT, são superiores foi confirmada, embora o RoBERTa não tenha superado o BERT, provavelmente devido à diferença no corpus de pré-treinamento.

5 Metodologia

Esta seção apresenta as etapas necessárias tanto para alcançar os resultados obtidos neste estudo quanto para viabilizar a aplicação das técnicas descritas anteriormente. As implementações das rotinas desenvolvidas foram realizadas utilizando a linguagem de programação Python (versão 3.10), amplamente reconhecida por sua versatilidade e suporte a tarefas de processamento de linguagem natural. Entre as bibliotecas utilizadas, destacam-se Numpy e Pandas, para manipulação e análise de dados; Sklearn, para implementação de algoritmos de aprendizado de máquina; e Spacy e Nltk, para tratamento e análise de textos.

A seguir, é apresentado um fluxograma que sintetiza a metodologia adotada neste trabalho, ilustrando o processo de aplicação das técnicas descritas.

Figura 2: Síntese Metodologia



Fonte: Confecção pelo autor.

5.1 Coleta de Dados

O conjunto de dados utilizado neste trabalho compreende portarias emitidas em 2024 pelo Gabinete da Presidência do Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT). Essas portarias foram coletadas diretamente das publicações oficiais disponíveis no site do Tribunal, utilizando técnicas de web scraping.

A prática de web scraping tem sido amplamente adotada devido à sua capacidade de automatizar a coleta de informações de páginas da web, transformando esses dados em formatos estruturados, sem a necessidade de intervenção manual, o que a torna uma ferramenta prática e eficiente para extração de dados (Bhardwaj et al., 2021, apud Oliveira, 2023). Neste trabalho, foi implementado um processo de web scraping para acessar o site oficial do TJDFT e extrair os textos completos das portarias publicadas.

Para garantir a extração apenas dos conteúdos relevantes, expressões regulares foram utilizadas para identificar e delimitar informações específicas, como número da portaria, data de emissão e o conteúdo principal. Essa abordagem permitiu a coleta precisa dos elementos necessários e assegurou que os dados estivessem alinhados com os objetivos do estudo, excluindo partes irrelevantes, como, por exemplo, rodapés.

5.2 Pré processamento e limpeza dos dados

O tratamento linguístico dos textos seguiu o conjunto de procedimentos descritos na seção de revisão de bibliografia para preparar os dados para as análises subsequentes. Primeiramente, os textos foram convertidos para letras minúsculas para garantir a uniformidade e evitar a duplicidade de palavras com variações de maiúsculas e minúsculas.

A remoção de stopwords foi realizada utilizando o pacote NLTK do Python, que fornece uma lista de palavras comuns em português, como "o", "a", "de", "em", entre outras. Além dessas palavras foram excluídos caracteres especiais, como "§", "º", "°", "§§", "§º", "ª", "art", "caput", "cc".

Em seguida, foi realizada a remoção de pontuações, que inclui símbolos como pontos, vírgulas e outros sinais que não contribuem para a análise textual. Além disso, a remoção de números e números romanos (I, II, III, IV, V, etc.) também foi realizada,

As remoções dessas palavras, caracteres e números não agregam valor significativo à análise, pois são considerados de baixo valor semântico e aparecem com grande frequência no texto. Assim, ao removê-los, a análise torna-se mais eficiente, permitindo que o foco permaneça nas palavras e expressões que realmente contribuem para o entendimento do conteúdo, como termos mais específicos ou técnicos presentes nas portarias.

Por fim, a tokenização foi aplicada para dividir os textos em unidades menores, facilitando as etapas subsequentes de análise, como a vetorização numérica e extração de informações dos documentos.

6 Resultados

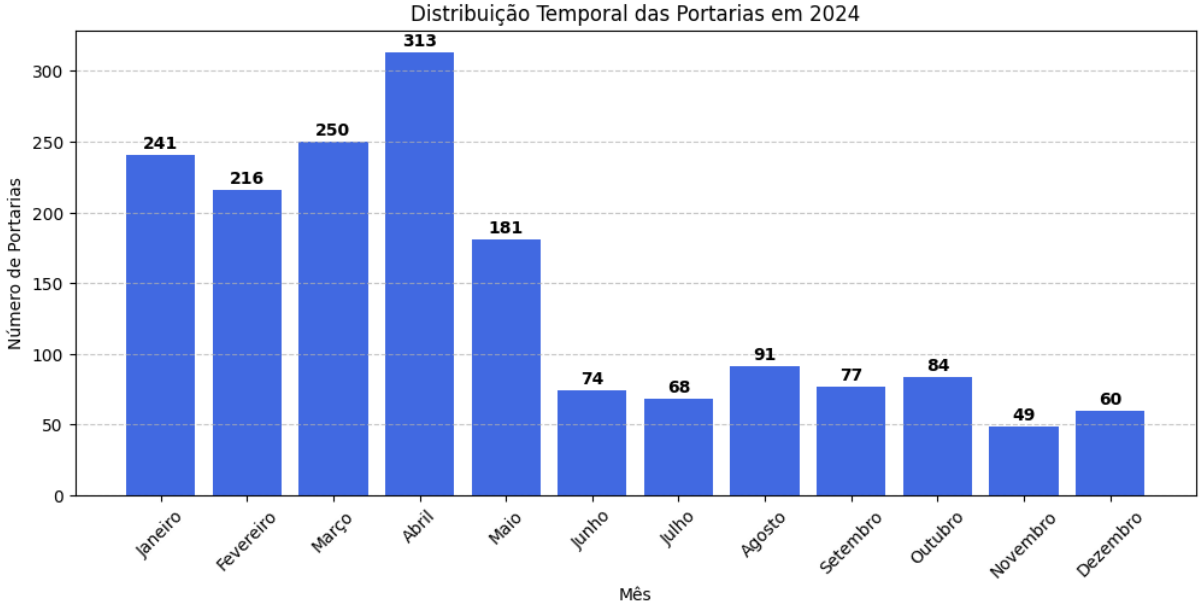
A seguir, serão discutidos os resultados obtidos em cada uma das etapas descritas anteriormente na metodologia, destacando-se as contribuições específicas de cada técnica e como elas interagem para alcançar os objetivos propostos no presente trabalho.

6.1 Análise Exploratória dos Dados

O banco de dados gerado a partir do processo de raspagem contém o total de 1.704 portarias publicadas ao longo do ano de 2024. A primeira portaria foi registrada na data de 2 de janeiro, enquanto a última foi publicada em 30 de dezembro.

A última portaria registrada foi identificada com o número 1963, o que sugere a existência de numerações ausentes ao longo do período analisado. Desse modo, verificou-se a ausência de 259 registros, indicando possíveis lacunas no processo de publicação das portarias. Vale ressaltar que esse fato pode estar relacionado a diferentes fatores, como erros de catalogação, portarias anuladas ou não disponibilizadas publicamente.

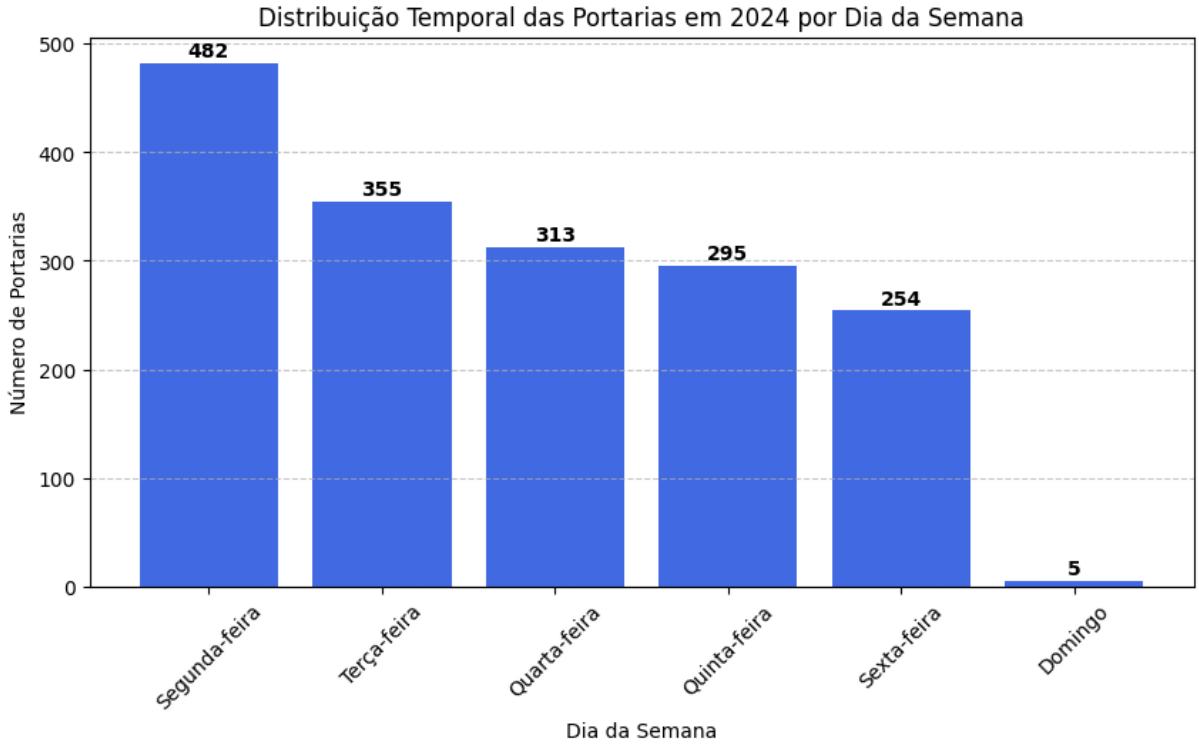
A distribuição das portarias ao longo do ano de 2024 apresenta variações significativas na frequência de publicações mensais. O gráfico a seguir ilustra essa distribuição, permitindo uma visualização das oscilações no volume de portarias emitidas por mês.



Nota-se que abril concentrou o maior número de publicações, totalizando 313 portarias, enquanto novembro registrou a menor quantidade, com apenas 49 publicações.

Essa discrepância pode estar associada a fatores sazonais, mudanças administrativas ou variações no volume de demandas regulatórias ao longo do ano.

Já a análise a seguir reflete a quantidade de portarias publicadas por dias da semana, fornecendo informações relevantes sobre o fluxo de trabalho e a intensidade das publicações ao longo dos dias.

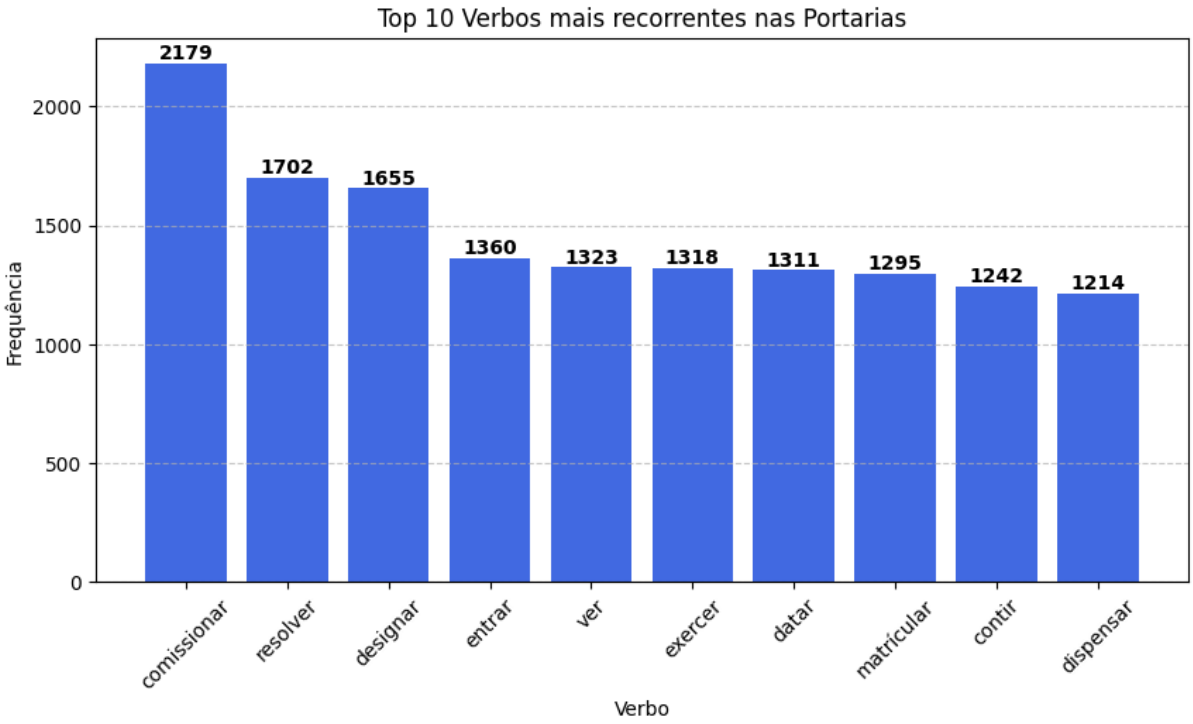


A análise da distribuição semanal das portarias evidencia um pico significativo nas publicações às segundas-feiras, seguido por uma queda gradual ao longo da semana. A ausência de publicações aos sábados e a quantidade mínima registrada aos domingos reforçam o caráter institucional, cuja divulgação se concentra nos dias úteis.

A nuvem de palavras a seguir destaca os termos mais recorrentes no conjunto de documentos, proporcionando uma visualização da frequência de cada palavra. Essa análise pode evidenciar possíveis temas predominantes nas publicações e auxiliar na identificação de entidades nomeadas a serem extraídas posteriormente.

A análise da nuvem de palavras evidencia que termos como "tribunal", "quadro", "área", "pessoal", "judiciário", "comissionada"e "distrito federal"se destacam entre as expressões mais recorrentes nas portarias. Essa predominância pode refletir uma forte ênfase nos aspectos relacionados à estrutura e à organização das instituições jurídicas, ressaltando temas institucionais e administrativos presentes nos documentos.

De maneira correlata, a análise dos verbos mais recorrentes nas portarias visa compreender a natureza das ações mais frequentemente documentadas ao longo do período analisado. Identificar esses termos auxilia na caracterização das principais finalidades administrativas das portarias, fornecendo percepções sobre possíveis temáticas predominantes nos atos publicados. O gráfico a seguir apresenta a distribuição da frequência desses verbos nas portarias analisadas.



Entre os verbos mais frequentes, destacam-se "Comissionar", "Resolver", "Designar", "Entrar", "Ver"e "Exercer". Esses termos indicam um forte enfoque em ações funcionais, decisões administrativas e delegações de responsabilidade, ressaltando o papel fundamental das portarias na formalização de atos institucionais.

7 Cronograma

As atividades a serem desenvolvidas durante o Trabalho de Conclusão de Curso são:

Tabela 1: Cronograma do TCC 1

Atividades	2/2024				
	Out	Nov	Dez	Jan	Fev
Escolha do tema					
Levantamento de bibliografias relacionadas					
Desenvolvimento da proposta de projeto					
Entrega da proposta ao Orientador					
Revisão da proposta					
Elaboração da apresentação da proposta					
Apresentação oral da proposta					
Raspagem dos Dados					
Pré-processamento Linguístico					
Representação Vetorial (Word Embeddings)					

Tabela 2: Cronograma do TCC 2

Atividades	1/2025				
	Mar	Abr	Mai	Jun	Jul
Aplicação de Modelos para Agrupamento					
Aplicação de Modelos REN					
Avaliação de Desempenho dos Modelos					
Elaboração do relatório final					
Entrega do relatório final ao Professor Orientador					
Revisão do relatório final					
Elaboração da apresentação da proposta					
Entrega do relatório final para a banca					
Apresentação oral da proposta					
Correção do relatório final					

8 Referências

Referências

ARAUJO, P. H. Luz de et al. Lener-br: A dataset for named entity recognition in brazilian legal text: 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings. In: _____. [S.l.: s.n.], 2018. p. 313–323. ISBN 978-3-319-99721-6.

BARROS, F. M. d. C. et al. Processamento de linguagem natural como ferramenta de suporte em documentos jurídicos: uma revisão sistemática. *Revista de Casos e Consultoria*, v. 15, n. 1, p. e36701, ago. 2024. Disponível em: ⟨https://periodicos.ufrn.br/casoseconsultoria/article/view/36701⟩.

CASELI, H. d. M.; NUNES, M. d. G. V. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 2ª Edição*. [S.l.]: BPLN, São Carlos, 2024. Disponível em: ⟨https://brasileiraspln.com/livro-pln/2a-edicao/⟩. ISBN 978-65-00-95750-1.

CASTRO, P. V. Q. *Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Goiás, Goiânia, 2019. 125 f.

COSTA, R. P. *Reconhecimento de entidades nomeadas em textos informais no domínio legislativo*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Goiás, Goiânia, 2023. 70 f.

CâNDIDO, E. C. R. *Um estudo comparativo de redes neurais profundas para classificação automática de texto*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Minas Gerais, Departamento de Ciência da Computação, Belo Horizonte, 2020. Disponível em: ⟨http://hdl.handle.net/1843/50545⟩. Acesso em: 23 dez. 2024.

DEVLIN, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. Disponível em: ⟨https://arxiv.org/abs/1810.04805⟩.

FREITAS, L. J. G. *Clusterização de textos aplicada ao tratamento de dados jurídicos desbalanceados*. Dissertação (Dissertação (Mestrado em Estatística)) — Universidade de Brasília, Departamento de Estatística, Brasília, 2023. Disponível em: ⟨http://repositorio.unb.br/handle/10482/48841⟩.

GOCHHAIT, D. S. Comparative analysis of machine and deep learning techniques for text classification with emphasis on data preprocessing. *Qeios*, 05 2024.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, v. 9, p. 1735–80, 12 1997.

I., D.; OBUNADIKE, G. Analysis and visualization of market segementation in banking sector using kmeans machine learning algorithm. *FUDMA JOURNAL OF SCIENCES*, v. 6, n. 1, p. 387 – 393, Apr. 2022. Disponível em: ⟨https://fjs.fudutsinma.edu.ng/index.php/fjs/article/view/910⟩.

JOULIN, A. et al. *FastText.zip: Compressing text classification models*. 2016. Disponível em: ⟨https://arxiv.org/abs/1612.03651⟩.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 282–289. ISBN 1-55860-778-1. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=645530.655813⟩.

LIU, Y. et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. Disponível em: ⟨https://arxiv.org/abs/1907.11692⟩.

MIKOLOV, T. et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. Disponível em: ⟨https://arxiv.org/abs/1301.3781⟩.

NOTHMAN, J. et al. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, v. 194, p. 151–175, 2013. ISSN 0004-3702. Artificial Intelligence, Wikipedia and Semi-Structured Resources. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0004370212000276⟩.

OLIVEIRA, R.; NASCIMENTO, E. G. S. Clustering by similarity of brazilian legal documents using natural language processing approaches. In: _____. [S.l.: s.n.], 2021. ISBN 978-1-83969-887-3.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: ⟨https://aclanthology.org/D14-1162⟩.