Universidade de Brasília
Departamento de Estatística

Processamento de Linguagem Natural aplicado a dados do poder judiciário

Análise comparativa de modelos de aprendizado de máquina e redes neurais para reconhecimento de entidades nomeadas e classificação de portarias jurídicas

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

Processamento de Linguagem Natural aplicado a dados do poder judiciário

Análise comparativa de modelos de aprendizado de máquina e redes neurais para reconhecimento de entidades nomeadas e classificação de portarias jurídicas

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 de 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. Nesse cenário, instituições jurídicas produzem uma vasta quantidade de documentos oficiais, como portarias, que representam fontes ricas de informações para análise e tomada de decisão. Contudo, a classificação desses textos e a extração automática de informações estruturadas, especialmente por meio de técnicas como o Reconhecimento de Entidades Nomeadas (NER), continuam desafiadoras devido às especificidades da linguagem jurídica e à heterogeneidade dos documentos conforme descrito por (Li et al, 2018, apud COSTA, 2023)

Nesse contexto, modelos de aprendizado de máquina e redes neurais profundas surgem como soluções promissoras para enfrentar esses desafios. Dito isso, este trabalho tem como objetivo geral comparar o desempenho dessas abordagens na tarefa de reconhecimento de entidades nomeadas e classificação de textos jurídicos, utilizando como fonte as portarias do Gabinete da Presidência do Tribunal de Justiça do Distrito Federal e dos Territórios (TJDFT). Essas técnicas têm o potencial de transformar dados não estruturados em conhecimento estruturado, promovendo maior eficiência no acesso às informações e agilidade nos processos administrativos e jurídicos.

Assim, espera-se contribuir para a modernização e eficiência dos processos jurídicos no Brasil, oferecendo subsídios para a adoção de soluções baseadas em inteligência artificial no sistema judiciário. *MELHORAR*

2 Objetivos

2.1 Objetivo Geral

Este trabalho tem como objetivo desenvolver e avaliar modelos baseados em aprendizado de máquina e redes neurais profundas para realizar tarefas de Reconhecimento de Entidades Nomeadas e classificação de textos presentes nas portarias recentes do Gabinete da Presidência do TJDFT. A proposta busca transformar dados não estruturados em informações organizadas e de fácil acesso, contribuindo para a aplicação prática de técnicas de Processamento de Linguagem Natural no contexto jurídico brasileiro. Espera-se que os resultados avancem a eficiência no gerenciamento de documentos oficiais e na extração de compreensões relevantes para o sistema judiciário.

2.2 Objetivos Específicos

- Levantamento dos Dados
  - Acessar o site oficial do TJDFT e extrair os dados relevantes das portarias publicadas do Gabinete da Presidência do TJDFT
- Construção e Pré-processamento das Representações Textuais
  - Aplicar técnicas de Processamento de Linguagem Natural (PLN)(e.g.,tokenização, remoção de stopwords...)
  - Gerar representações textuais utilizando embeddings de palavras (e.g., Word2Vec, GloVe) e embeddings contextuais (e.g., BERT).
- Desenvolvimento dos Modelos de Reconhecimento de Entidades Nomeadas (NER)
  - Ajustar redes neurais profundas para Reconhecimento de Entidades Nomeadas.(e.g.,LSTM-CRF, transformers ...)
- Desenvolvimento dos Modelos de Classificação de Textos
  - Treinar modelos não supervisionados para a identificar similaridade entre as portarias.
  - Treinar modelos supervisionados para a classificação de portarias em categorias específicas.
  - Implementar redes neurais profundas para comparação de desempenho.(e.g., BERT)

- Avaliação de Desempenho dos Modelos
  - Comparar os modelos desenvolvidos utilizando métricas (e.g. Precisão, Acurácia, F1-score, tempo de processamento).
  - Identificar as limitações das técnicas aplicadas e propor melhorias ou extensões para trabalhos futuros.
- Produção de Contribuições Finais
  - Gerar insights sobre o uso de modelos de aprendizado de máquina e redes neurais.

3 Metodologia

3.1 Conjunto 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 cabeçalho, número da portaria, data de emissão e o conteúdo principal. Essa abordagem permitiu a coleta precisa dos elementos necessários, excluindo partes irrelevantes, como rodapés e metadados, e assegurando que os dados estivessem alinhados com os objetivos do estudo.

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

1. **Tokenização:** Divisão dos textos em unidades discretas e individuais nomeadas tokens, utilizando bibliotecas python especializadas como *spaCy* ou *NLTK*.

2. **Remoção de Stop Words:**

3.2.2 Representação de Palavras: Word Embeddings

Explicação Geral

- **TF-IDF (Term Frequency-Inverse Document Frequency):**

  Uma abordagem eficiente para atribuir maior relevância às palavras mais representativas de um corpus é o uso do algoritmo TF-IDF como descrito por (FREITAS, 2023)

  A fórmula geral do TF-IDF é expressa como:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t)$$

  onde:

  - t: Termo específico.
  - d: Documento em análise.

  A frequência do termo é definida como:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

  onde:

  - $f_{t,d}$: Número de vezes que o termo t aparece no documento d.
  - $\sum_{t' \in d} f_{t',d}$: Soma de frequências de todos os termos t' no documento d.

  A frequência inversa do documento é calculada como:

$$\text{IDF}(t) = \log\left(\frac{N}{1 + \text{DF}(t)}\right)$$

  onde:

  - N: Número total de documentos no corpus.
  - $\text{DF}(t)$: Número de documentos em que o termo t aparece pelo menos uma vez.
  - O valor 1 é adicionado ao denominador para evitar divisão por zero.

  A combinação das duas fórmulas resulta no valor TF-IDF:

$$\text{TF-IDF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log\left(\frac{N}{1 + \text{DF}(t)}\right)$$

- **Word2Vec:**
- **GloVe (Global Vectors for Word Representation):**
- **FastText:**

BERT (Bidirectional Encoder Representations from Transformers)

3.2.3 Named Entity Recognition (NER)

Explicação Geral

Formato de Anotação :

BIO (Begin-Inside-Outside)
IOB2
IOBES

Arquitetura ELMo

- LeNER-Br
- WikiNER
- Paramopama
- HAREM
- Wang2Vec

3.3 Aprendizado de Máquina

Explicação Geral

Pipeline de Treinamento

Divisão dos dados (treino, validação e ajuste de hiperparâmetros, teste) .

3.3.1 Aprendizado Supervisionado

Explicação Geral

Implementar processo de identificação de temáticas gerais das portarias (Exemplo : Aposentadoria ...)

3.3.2 Aprendizado Não Supervisionado

Explicação Geral

- Kmeans
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

3.3.3 Avaliação

Métricas: precisão, recall, F1-score, tempo de processamento

Precisão (Precision):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Abrangência):

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

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

Legenda:

- TP: True Positives (Verdadeiros Positivos)
- FP: False Positives (Falsos Positivos)
- FN: False Negatives (Falsos Negativos)

3.4 Redes Neurais

Explicação Geral

- Bidirectional Encoder Representation from Transformers (BERT)
- Long Short-Term Memory (LSTM)

- Conditional Random Fields (CRF)
- LSTM-CRF

(GOCHHAIT, 2024) (BARROS et al., 2024) (CASELI; NUNES, 2024) (CASTRO, 2019) (COSTA, 2023) (OLIVEIRA; NASCIMENTO, 2023) (OLIVEIRA; NASCIMENTO, 2023) (FREITAS, 2023) (OLIVEIRA; NASCIMENTO, 2021) (NOGUTI; VELLASQUES; OLIVEIRA, 2020) (OLIVEIRA, 2023) (SILVA, 2020)

4 Cronograma

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

Tabela 1: Cronograma do TCC 1

Atividades	2/2024				
	Outubro	Novembro	Dezembro	Janeiro	Fevereiro
Escolha do tema a ser abordado					
Levantamento de bibliografias relacionadas ao tema					
Desenvolvimento da proposta de projeto					
Entrega da proposta de projeto ao Orientador					
Revisão da proposta					
Coleta e Raspagem dos Dados					
Elaboração da apresentação da proposta					
Apresentação oral da proposta					
Pré Processamento das Representações Textuais					

Tabela 2: Cronograma do TCC 2

Atividades	1/2025				
	Março	Abril	Maio	Junho	Julho
Aplicação de Modelos REN					
Aplicação de Modelos para Classificação de Textos					
Avaliação de Desempenho dos Modelos					
Elaboração do relatório final					
Entrega do relatório final ao Professor Orientador					
Revisão do relatório final					
Entrega do relatório final para a banca					
Elaboração da apresentação da proposta					
Apresentação oral da proposta					
Correção do relatório final					

Referências

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.

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.

NOGUTI, M. Y.; VELLASQUES, E.; OLIVEIRA, L. S. Legal document classification: An application to law area prediction of petitions to public prosecution service. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020. p. 1–8. Disponível em: ⟨http://dx.doi.org/10.1109/IJCNN48605.2020.9207211⟩.

OLIVEIRA, J. S. *Web scraping na extração e combinação sistemática de conteúdos: ferramenta auxiliar em processos de pesquisa, desenvolvimento e inovação (PD&I)*. Dissertação (Dissertação (Mestrado em Engenharia Biomédica)) — Universidade de Brasília, Faculdade UnB Gama, Brasília, 2023. Disponível em: ⟨http://repositorio.unb.br/handle/10482/49967⟩.

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.

OLIVEIRA, R. S. de; NASCIMENTO, E. G. S. *Analysing similarities between legal court documents using natural language processing approaches based on Transformers*. 2023. Disponível em: ⟨https://arxiv.org/abs/2204.07182⟩.

SILVA, A. V. e. *Um modelo de classificação para o Reconhecimento de Entidades Nomeadas*. Dissertação (Dissertação (Mestrado em Semiótica e Lingüística Geral)) — Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2020. Doi:10.11606/D.8.2020.tde-06042021-192617, Acesso em: 2024-12-12.