

Processamento de Linguagem Natural para Aplicação de Técnicas de Aprendizado de Máquina e Reconhecimento de Entidades Nomeadas em Portarias Jurídicas

Davi Esmeraldo da Silva Albuquerque

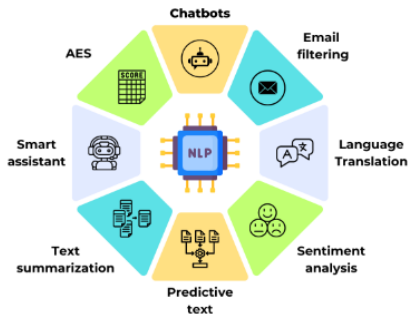
Orientador: Eduardo Monteiro de Castro

29 de julho de 2025

- 1 Introdução
- 2 Metodologia
- 3 Resultados
- 4 Considerações Finais
- 5 Referências

Definição

Processamento de Linguagem Natural (PLN) é uma área da **Inteligência Artificial (IA)** que busca desenvolver métodos e sistemas capazes de processar (interpretar, compreender e gerar) linguagem humana de forma computacional. (CASELI; NUNES, 2024)



Definição

Entidades Nomeadas são palavras ou frases que representam elementos específicos e bem definidos em um texto.

- Exemplos incluem:
 - **Pessoas:** Albert Einstein, Ada Lovelace
 - **Locais:** Brasília, Monte Everest
 - **Organizações:** ONU, Google
 - **Datas e Horários:** 22 de janeiro de 2025, 15h30

Definição

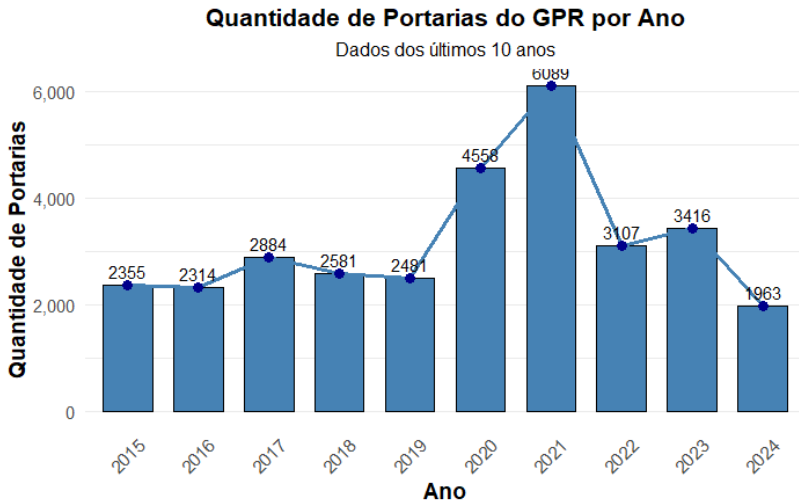
Reconhecimento de Entidades Nomeadas (NER - Named Entity Recognition) é uma tarefa de PLN que visa identificar e classificar automaticamente entidades nomeadas.

- Aplicações incluem:
 - Extração de informações de notícias e documentos.
 - Melhoria em sistemas de busca.
 - Mineração de opiniões.
 - Análise de currículos.

Modelos de **Aprendizado de Máquina** e de **Aprendizado Profundo**, quando combinados com técnicas de PLN, destacam-se como abordagens promissoras em razão do bom desempenho demonstrado a partir de **grandes volumes de dados textuais** (OLIVEIRA; NASCIMENTO, 2021).

Introdução

- Instituições jurídicas geram uma quantidade significativa de documentos oficiais.

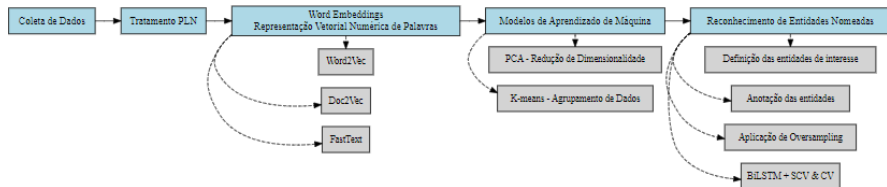


Fonte: Gabinete da Presidência do TJDF

- O objetivo deste trabalho é aplicar modelos baseados em aprendizado de máquina e aprendizado profundo para realizar as tarefas de mensuração de similaridade semântica, identificação de agrupamentos temáticos e Reconhecimento de Entidades Nomeadas (REN) de portarias emitidas em 2024 pelo Gabinete da Presidência do Tribunal de Justiça do Distrito Federal e Territórios (TJDFT).
- Propõe-se também o desenvolvimento de um aplicativo web.

- Implementação dos processos computacionais em **Python** (versão 3.10), utilizando o ambiente **Google Colab**.

Metodologia - Síntese



Avaliação Intrínseca: A capacidade dos *embeddings* em capturar relações sintáticas ou semânticas entre termos textuais é analisada diretamente, avaliando sua coerência e representatividade (SCHNABEL et al., 2015).

- **Avaliação Intrínseca:** Verificação empírica da coerência das similaridades entre a portaria 1963 e as dez portarias mais semelhantes a ela, mensuradas pela similaridade cosseno.

Coesão e proximidade dos *embeddings*:

- Coesão Global: Média das similaridades entre todos os pares possíveis de portarias.
- Coesão Local: Média das similaridades entre cada portaria e suas 10 portarias mais similares.

Avaliação Extrínseca: os *embeddings* são utilizados como características de entrada em tarefas externas. O desempenho nessas tarefas funciona como um indicador da qualidade dos *embeddings* (SCHNABEL et al., 2015).

- **Avaliação Extrínseca:** Verificação da consistência dos resultados obtidos nos processos de agrupamento e reconhecimento de entidades nomeadas, utilizando os *embeddings* gerados.

Resultado - Dados Processados

- Total de 1.707 portarias publicadas ao longo do ano de 2024.

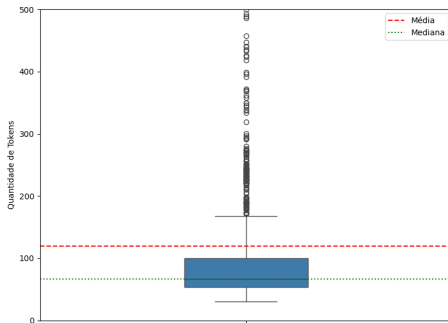
Resultado - Dados Processados

- Total de 1.707 portarias publicadas ao longo do ano de 2024.

Tabela: Estatísticas descritivas da quantidade de tokens por portaria

Estatística	Valor
Média	119,71
Mediana	66
Mínimo	30
Máximo	4161
Desvio padrão	216,13

Figura: Boxplot - Tokens por Portaria



Resultado - Word Embeddings

Tabela: Modelos avaliados, hiperparâmetros ótimos e métricas de coesão

Modelo	Hiperparâmetros	Global	Local
Word2Vec	Skip-gram	0,8650	0,9918
Doc2Vec	DBOW	0,8604	0,9907
FastText	Skip-gram	0,8575	0,9913

Resultado - Word Embeddings

Tabela: Modelos avaliados, hiperparâmetros ótimos e métricas de coesão

Modelo	Hiperparâmetros	Global	Local
Word2Vec	Skip-gram	0,8650	0,9918
Doc2Vec	DBOW	0,8604	0,9907
FastText	Skip-gram	0,8575	0,9913

- Após a avaliação intrínseca das similaridades dos Embeddings *Doc2Vec*, constatou-se que portarias semanticamente distintas foram consideradas como as mais similares à portaria de número 1963.



Resultado - Word Embeddings

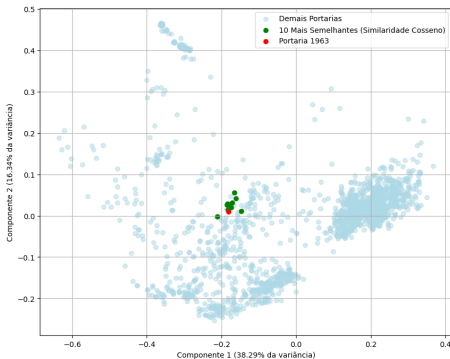
Tabela: Top 10 portarias mais semelhantes à portaria de referência segundo Word2Vec e FastText

Posição	Word2Vec	Similaridade	FastText	Similaridade
1	Portaria 489	0,9995	Portaria 489	0,9994
2	Portaria 1598	0,9991	Portaria 1598	0,9990
3	Portaria 1951	0,9982	Portaria 1951	0,9974
4	Portaria 1472	0,9964	Portaria 1472	0,9967
5	Portaria 1597	0,9958	Portaria 1766	0,9966
6	Portaria 1766	0,9956	Portaria 1597	0,9962
7	Portaria 1033	0,9943	Portaria 1033	0,9956
8	Portaria 1709	0,9941	Portaria 1709	0,9948
9	Portaria 1524	0,9917	Portaria 1524	0,9924
10	Portaria 1767	0,9893	Portaria 1407	0,9907



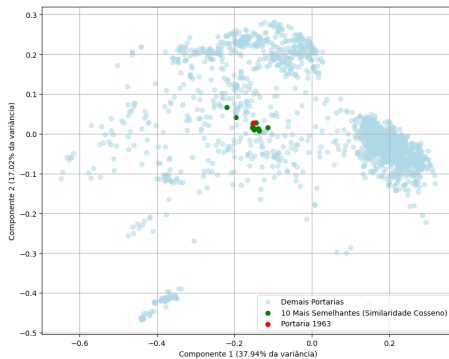
Resultado - Redução de Dimensionalidade

Figura: PCA - Word2Vec



■ 54,63% da variância total

Figura: PCA - FastText



■ 54,96% da variância total

Resultado - Agrupamento das Portarias

■ Determinação do número ótimo de clusters

Figura: Word2Vec

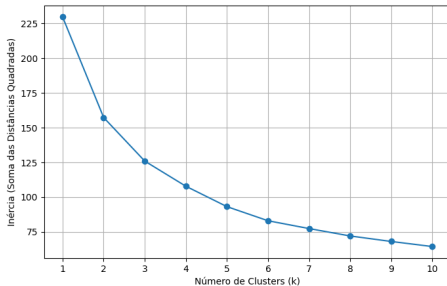
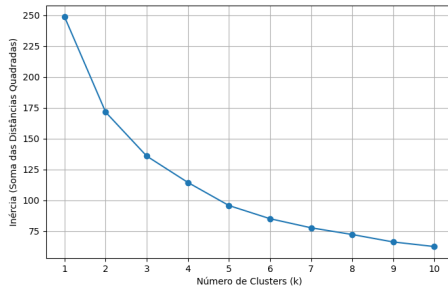


Figura: FastText



Resultado - Agrupamento das Portarias

Figura: Clusterização das portarias - Word2Vec

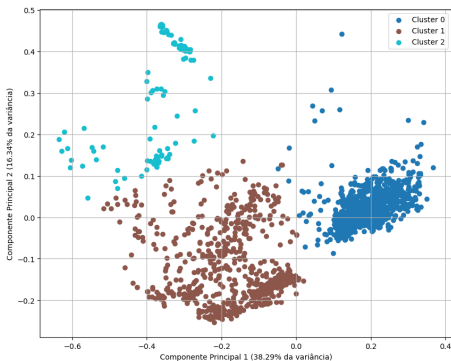
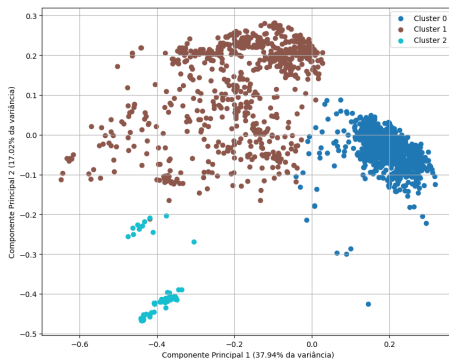


Figura: Clusterização das portarias - FastText



Resultado - Agrupamento das Portarias

Figura: Clusterização das portarias - Word2Vec

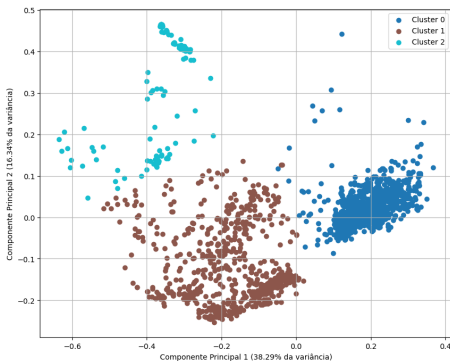


Tabela: Distribuição de portarias por cluster

Cluster	Quantidade	Percentual (%)
Cluster 0	890	52,13%
Cluster 1	660	38,66%
Cluster 2	157	9,19%
Total	1707	100%

- Foram anotadas manualmente 735 portarias no total.

Tabela: Distribuição das entidades anotadas

Entidade	Quantidade	Percentual (%)
ACAO	1107	55,60
SUJEITO	522	26,22
LOCAL	220	11,05
DATA	142	7,13
Total	1991	100,00

Resultado - Dados Anotados Agrupados

Tabela: Distribuição das entidades anotadas por cluster

Entidade	Cluster 0 (%)	Cluster 1 (%)	Cluster 2 (%)
ACAO	682 (65,58)	317 (50,16)	108 (33,86)
SUJEITO	349 (33,56)	164 (25,95)	9 (2,82)
LOCAL	9 (0,87)	99 (15,66)	112 (35,11)
DATA	0 (0,00)	52 (8,23)	90 (28,21)
Total	1040 (100,00)	632 (100,00)	319 (100,00)

Resultados - Reconhecimento de Entidades Nomeadas

■ Avaliação do Modelo com Validação Cruzada Estratificada

Tabela: Desempenho médio do modelo - Validação Cruzada Estratificada

Média	F1-score
Macro	$0,7973 \pm 0,0471$
Micro	$0,8789 \pm 0,0205$

Tabela: F1-score médio por entidades - Validação Cruzada Estratificada

Entidade	F1-score
DATA	$0,7504 \pm 0,1172$
ACAO	$0,8698 \pm 0,0334$
LOCAL	$0,6474 \pm 0,0559$
SUJEITO	$0,9217 \pm 0,0155$



Resultados - Reconhecimento de Entidades Nomeadas

■ Avaliação do Modelo com Validação Cruzada Simples

Tabela: Desempenho médio do modelo - Validação Cruzada Simples

Média	F1-score
Macro	$0,8036 \pm 0,0365$
Micro	$0,8766 \pm 0,0180$

Tabela: F1-score médio por entidades - Validação Cruzada Simples

Entidade	F1-score
DATA	$0,7351 \pm 0,0939$
ACAO	$0,8691 \pm 0,0296$
LOCAL	$0,6972 \pm 0,0465$
SUJEITO	$0,9129 \pm 0,0128$



Resultado - Dados Anotados após Sobreamostragem

Tabela: Distribuição das entidades anotadas após sobreamostragem

Entidade	Quantidade	Percentual (%)
ACAO	1371	51,34
SUJEITO	576	21,56
LOCAL	440	16,47
DATA	284	10,63
Total	2671	100,00

Resultados - Reconhecimento de Entidades Nomeadas

- Avaliação do Modelo com Validação Cruzada Simples após sobre amostragem

Tabela: Desempenho médio do modelo após sobre amostragem - Validação Cruzada Simples

Média	F1-score
Macro	0,8957 \pm 0,0200
Micro	0,9146 \pm 0,0178

Tabela: F1-score médio por entidades após sobre amostragem - Validação Cruzada Simples

Entidade	F1-score
DATA	0,9376 \pm 0,0364
ACAO	0,9053 \pm 0,0165
LOCAL	0,7968 \pm 0,0554
SUJEITO	0,9432 \pm 0,0211





Resultado - Aplicativo


Considerações Finais


- Modernização da análise de portarias no setor jurídico.
- Ganhos de eficiência.
- Suporte à tomada de decisões administrativas e judiciais.
- Transparência dos processos internos presentes nas portarias contempladas.

- Skipgram e CBOW apresentaram resultados mais coerentes em comparação com as demais arquiteturas para o conjunto de dados analisado.
- Mensuração de similaridade semântica, identificação de agrupamentos temáticos e Reconhecimento de Entidades Nomeadas (REN) apresentaram resultados pertinentes.
- Embeddings gerados validados por meio de avaliação intrínseca e extrínseca.


 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.


 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>>.

 GARCIA, G. C. *Reconhecimento de Entidades Nomeadas na base de notificações de eventos adversos e queixas técnicas de dispositivos médicos no Brasil*. Dissertação (Dissertação (Mestrado Profissional em Computação Aplicada)) — Universidade de Brasília, Brasília, ago 2021. Data de defesa: 31 de agosto de 2021. Disponível em: <<http://repositorio.unb.br/handle/10482/42718>>.



 GÉRON, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2. ed. Sebastopol, CA: O'Reilly Media, Inc., 2019. Disponível em: <<https://www.rasa-ai.com/wp-content/uploads/2022/02/Aurélien-Géron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-TensorFlow-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O'Reilly.pdf>>. Acesso em: 14 jun. 2025. ISBN 978-1-492-03264-9.

 OLIVEIRA, R.; NASCIMENTO, E. G. S. Clustering by similarity of brazilian legal documents using natural language processing approaches. In: _____. [s.n.], 2021. ISBN 978-1-83969-887-3. Disponível em: <https://www.researchgate.net/publication/354579623_Clustering_by_Similarity_of_Brazilian_Legal_Documents_Using_Natural_Language_Processing_Approaches>.

 SCHNABEL, T. et al. Evaluation methods for unsupervised word embeddings. In: MÀRQUEZ, L.; CALLISON-BURCH, C.; SU, J. (Ed.). *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 298–307. Disponível em: <<https://aclanthology.org/D15-1036/>>.