Uso de Modelos de Linguagem para a Extração de Features de Documentos Não Estruturados

Jéssica Gomes, Valesca Silva

Introdução

Atualmente, a maioria dos dados gerados por organizações e portais de notícias encontra-se em formato não estruturado, como relatórios em PDF e artigos de imprensa. A análise manual desse volume é lenta, custosa e sujeita a falhas, impedindo a extração de insights em tempo real.

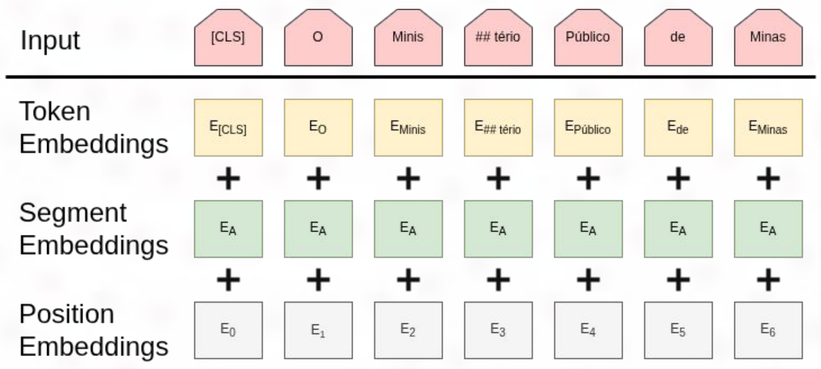
Para superar esse gargalo, a Extração de Informação (EI) utiliza o Processamento de Linguagem Natural (PLN) para converter textos não estruturados contendo textos em linguagem humana em dados estruturado. Nesse contexto, o Reconhecimento de Entidades Nomeadas (NER) destaca-se como uma técnica fundamental de rotulagem de sequências, cujo o objetivo é localizar todas as entidades nomeadas mencionadas em textos não estruturados e classificá-las em categorias predefinidas, como nomes de pessoas, organizações e locais.

Metodologia

Para a extração de features e reconhecimento de entidades (NER), utilizou-se o modelo Bidirectional Encoder Representations from Transformers (BERT). O BERT é um tipo de transformer que usa apenas a parte do encoder da arquitetura para gerar um modelo da linguagem.

A rede processa o texto através da soma de três camadas de embeddings, garantindo que o modelo compreenda não apenas o significado das palavras, mas também sua posição e relação estrutural:

- **Token Embeddings:** Segmentação do texto em unidades atômicas (tokens).
- **Segment Embeddings:** Diferenciação entre pares de sentenças.
- **Position Embeddings:** Codificação da ordem sequencial dos tokens.



O diferencial do BERT reside na sua capacidade de aprendizado bidirecional através de duas tarefas principais durante o pré-treinamento em larga escala:

- **Masked Language Modeling (MLM):** Ocultação aleatória de 15% dos tokens para que o modelo aprenda a predizer palavras com base no contexto global.
- **Next Sentence Prediction (NSP):** Treinamento binário para identificar se uma sentença sucede logicamente a outra, capturando relações de longo prazo.

Após o pré-treino, o ajuste fino (Fine-tuning) é realizado voltado especificamente para a tarefa de NER. Enquanto o pré-treinamento é computacionalmente exaustivo, o ajuste fino é pouco dispendioso e pode ser realizado em poucas horas numa TPU ou GPU.

Conjunto de Dados

O conjunto de dados utilizado foi o LeNER-Br, um dataset em língua portuguesa para tarefa de NER, composto por 70 documentos jurídicos brasileiros (66 de tribunais superiores e estaduais, e 4 de legislação, como a Lei Maria da Penha) anotados manualmente utilizando o WebAnno seguindo o esquema de marcação IOB:

Token	Tag
O	O
Ministério	B-ORGANIZACAO
Público	I-ORGANIZACAO
de	I-ORGANIZACAO
Minas	I-ORGANIZACAO
Gerais,	I-ORGANIZACAO

- **B (Beginning):** Indica que o token é o começo da entidade.
- **I (Inside):** Indica que o token pertence a entidade.
- **O (Outside):** Indica que o token não pertence a nenhuma entidade.

Resultados

O BERTimbau base foi avaliado sobre o conjunto de teste (1.390 sentenças, 47.630 tokens). A avaliação considera a entidade completa como unidade de medida — uma predição só é contada como correta se todos os tokens da entidade forem classificados corretamente. Os resultados são apresentados na Tabela 1.

Entidade	Precisão	Recall	F1
JURISPRUDENCIA	74%	87%	80%
LEGISLACAO	92%	97%	94%
LOCAL	67%	81%	73%
ORGANIZACAO	86%	90%	88%
PESSOA	93%	97%	95%
TEMPO	92%	91%	91%
micro avg	87%	92%	90%
macro avg	84%	91%	87%

Tabela 1. Precision, Recall e F1-score por entidade no conjunto de teste do leNER-Br.

O modelo alcançou F1 geral de 0,90, com desempenho acima de 0,90 em PESSOA (0,95), LEGISLACAO (0,94) e TEMPO (0,91). A categoria LOCAL obteve o menor F1 (0,73): como mostra a Tabela 2, ela possui apenas 132 palavras no teste contra 2.669 de LEGISLACAO — desequilíbrio já documentado pelos autores do dataset (Luz de Araujo et al., 2018).

Entidade	Treino	Validação	Teste
JURISPRUDENCIA	3.967	743	660
LEGISLACAO	13.039	2.609	2.669
LOCAL	1.417	244	132
ORGANIZACAO	6.671	1.608	1.367
PESSOA	4.612	894	735
TEMPO	2.343	543	260

Tabela 2. Contagem de palavras em entidades por categoria e por split do leNER-Br.

Exemplo de inferência do modelo em sentença jurídica contendo as seis categorias de entidade do LeNER-Br:

O Ministério Público de Minas Gerais, representado pelo promotor João Silva em Brasília, ajuizou ação com base na Lei Maria da Penha em 15/03/2022, conforme decidido no Habeas Corpus 110.260 do Superior Tribunal de Justiça.

— ORGANIZACAO — PESSOA — LOCAL — LEGISLACAO — TEMPO — JURISPRUDENCIA

Conclusões

A aplicação do modelo BERT demonstrou ser eficaz para o desafio de extrair informações de documentos jurídicos. O modelo alcançou um excelente equilíbrio em classificar as entidades corretamente, alcançando F1 score acima de 80% para a maioria das entidades.

Os resultados mostram que essa abordagem consegue lidar bem com documentos não estruturados, transformando grandes volumes de dados brutos em variáveis estruturadas.

Bibliografia

NAMYSŁ, Marcin. Robust Information Extraction From Unstructured Documents. 2023. Tese de Doutorado. Universitäts-und Landesbibliothek Bonn.

LUZ DE ARAUJO, Pedro Henrique et al. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In: International Conference on Computational Processing of the Portuguese Language. Cham: Springer International Publishing, 2018. p. 313-323.

BERTimbau - Portuguese BERT. Disponível em: https://huggingface.co/neuralmind/bert-base-portuguese-cased