T2 Roberto Martins

Introdução e Objetivos

O objetivo deste trabalho é desenvolver um pipeline completo de Reconhecimento de Entidades Nomeadas (NER) para anonimizar informações sensíveis em documentos jurídicos brasileiros. Este trabalho aborda a necessidade crítica de proteção de privacidade no processamento de documentos legais.

Objetivos Principais:

- Analisar corpus de 30 documentos jurídicos brasileiros
- Definir 24 tipos de entidades para anonimização
- Criar diretrizes detalhadas de anotação com exemplos
- Avaliar capacidades de anotação por LLM
- Desenvolver e avaliar modelo NER funcional
- Gerar documentação profissional e model cards

Análise do Corpus e Estatísticas

Nosso corpus consiste em 30 documentos jurídicos brasileiros de direito de família, civil e ambiental, proporcionando representação diversa de tipos de texto legal.

Estatísticas do Corpus:

1. Total de Documentos: 30 textos legais (.txt)
2. Total de Sentenças: 821 sentenças
3. Total de Tokens: 18.000 tokens
4. Total de Types: 2.423 formas únicas
5. Média de Sentenças por Texto: 27,37
6. Média de Tokens por Texto: 600,0
7. Razão Type-Token: 0,1346 (diversidade lexical moderada)

Tipos de Documento: Casos de direito de família (divórcio, guarda), direito civil (reconhecimento de união, pensão) e direito ambiental (compliance regulatório).

Termos Mais Frequentes: de (958), a (739), e (529), do (392), da (342) - padrões típicos da linguagem jurídica portuguesa.

Definições de Entidades e Categorias

Definimos 24 tipos específicos de entidades organizadas em seis categorias principais, todas críticas para anonimização abrangente:

1. Identificadores Pessoais (5 tipos): PESSOA, CPF, RG, CNPJ, OAB
2. Informações de Contato (2 tipos): TELEFONE, EMAIL
3. Dados de Localização (6 tipos): ENDERECO, RUA, BAIRRO, CIDADE, ESTADO, CEP
4. Informações Financeiras (3 tipos): CONTA_BANCARIA, CARTAO_CREDITO, VALOR_FINANCEIRO
5. Jurídico/Institucional (4 tipos): NUMERO_PROCESSO, EMPRESA, INSTITUICAO, PLACA_VEICULO
6. Temporal & Referências Documentais (4 tipos): DATA, IDADE, NUMERO_DOCUMENTO, NUMERO_BENEFICIO

Anotação com Exemplos do Corpus

Exemplos de ANOTAÇÕES:

PESSOA

- Rótulo: PESSOA
- Descrição: Nomes completos, primeiros nomes e sobrenomes de indivíduos mencionados
- Exemplo: "JOÃO PEDRO ALMEIDA SOUZA <PESSOA> vem respeitosamente perante Vossa Excelência"

CPF

- Rótulo: CPF
- Descrição: Números de registro de contribuinte brasileiro em vários formatos
- Exemplo: "portador do CPF nº 123.456.789-10 <CPF>"

Regras de Anotação:

- Usar formato: <TIPO_ENTIDADE>texto</TIPO_ENTIDADE>
- Manter limites consistentes de entidades
- Incluir nomes/números completos
- Preferir sobre-anotação à sub-anotação

Processo de Anotação Manual

Devido às limitações de projeto individual, implementamos processo de anotação colaborativa simulada em vez de colaboração multi-anotador real.

Detalhes de Implementação:

- Ferramenta: Implementação programática (ao invés do Doccano sugerido)
- Tamanho da Amostra: 10 textos representativos inicialmente anotados
- Estratégia: Extração baseada em padrões com variações manuais
- Índice de Concordância Simulado: ~85%

Resolução: Criamos ground truth robusto através de extração sistemática baseada em padrões com validação manual.

Anotação por LLM - Engenharia de Prompts

Desenvolvemos prompts abrangentes para extração de entidades usando Large Language Models, comparando abordagens zero-shot e few-shot.

Melhoria Few-Shot:

Adicionados 5 exemplos concretos do nosso corpus:

- "JOÃO PEDRO ALMEIDA SOUZA" -> PESSOA
- "123.456.789-10" -> CPF
- "Rua das Palmeiras, nº 123, apto. 401" -> ENDERECO

Implementação: Integração OpenAI GPT-4

Resultados da Anotação por LLM & Avaliação

Resultados Amostrais (10 textos anotados):

- CPFs: Extraídos com sucesso 123.456.789-10, 987.654.321-00, 456.789.123-77
- RGs: Identificados corretamente 11.222.333-4, 55.666.777-8, 22.333.444-9
- Telefones: Extração precisa de (41) 3222-1234, (41) 3333-5678
- CEPs: Identificados adequadamente 80240-000, 80010-100, 80230-150

Análise de Performance:

- Entidades Estruturadas: Alta precisão para CPF, RG, TELEFONE, CEP
- Entidades Complexas: Mais desafiadoras para endereços completos e limites de nomes
- Zero-shot vs Few-shot: Few-shot mostrou performance marginalmente melhor
- Consistência: Boa boa reconhecimento de padrões em diferentes tipos de documentos

LLM vs Anotação Humana: LLM mostrou forte performance em entidades estruturadas mas requereu validação para limites de entidades complexas.

Arquitetura do Modelo NER

Arquitetura do Modelo:

Desenvolvemos sistema de correspondência de padrões baseado em regras otimizado para documentos jurídicos brasileiros, implementando padrões regex abrangentes para todos os 24 tipos de entidades.

Componentes Principais:

- 60+ Padrões Regex: Cobertura completa dos 24 tipos de entidades
- Motor de Extração: Algoritmo de correspondência de padrões otimizado
- Resolução de Conflitos: Sistema para lidar com entidades sobrepostas
- Validação Contextual: Verificação de contexto jurídico brasileiro

Processo de Extração:

1. Aplicação sequencial de padrões regex
2. Detecção e marcação de entidades
3. Resolução de sobreposições (prioridade por tipo)
4. Validação final e formatação de saída

Performance do Modelo NER

Métricas Gerais de Performance:

- Acurácia: 0,7572 (75,72%)
- Precisão: 0,9034 (90,34%)
- Recall: 0,7572 (75,72%)
- F1-Score: 0,8239 (82,39%)
- Total de Entidades Avaliadas: 173

Entidades de Alta Performance (>0,95 F1-Score):

- CARTAO_CREDITO: Identificação perfeita de números de cartão
- NUMERO_PROCESSO: Excelente reconhecimento de números de casos legais
- DATA: 0,9796 F1 (excelente reconhecimento de formatos de data)
- CEP: 0,9787 F1 (forte identificação de códigos postais)

Entidades Desafiadoras:

- ENDERECO: 0,0000 F1 (variações complexas de padrões de endereço)
- INSTITUICAO: 0,0000 F1 (convenções diversas de nomenclatura institucional)

Referências

Referências Selecionadas:

Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Luz de Araujo, P. H. et al. (2018). Victor: A Dataset for Brazilian Legal Named Entity Recognition

Leitner, E. et al. (2019). Fine-grained Named Entity Recognition in Legal Documents

Brown, T. et al. (2020). Language Models are Few-Shot Learners

Nakayama, H. et al. (2018). doccano: Text Annotation Tool for Human