Projeto de Formatura – 2025

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema: **Desenvolvimento de um Modelo de Segmentação Automática para Documentos Jurídicos**

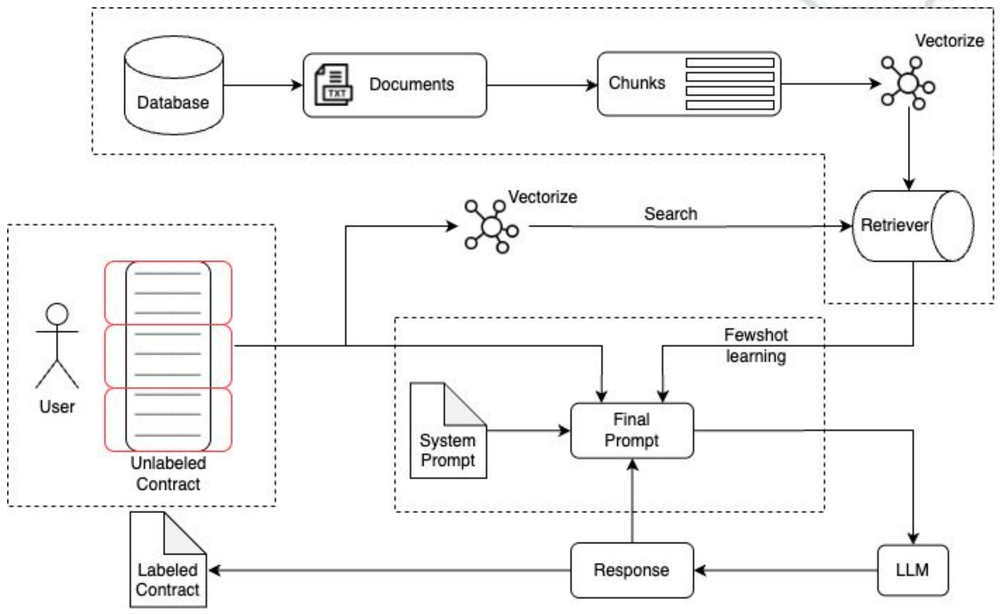
INTRODUÇÃO

O volume e a complexidade de contratos jurídicos digitais cresceram de forma acelerada, tornando a leitura e a revisão manual lentas, caras e sujeitas a falhas. Abordagens baseadas apenas em regras ou padrões de layout, como expressões regulares, não se adaptam bem à diversidade de formatos e idiomas. Este trabalho desenvolve um modelo de Processamento de Linguagem Natural capaz de segmentar e rotular automaticamente contratos a partir do texto bruto, estruturando o documento para apoiar análise contratual, revisão e extração de informações.

OBJETIVOS

O objetivo é desenvolver um sistema capaz de, a partir do texto bruto, detectar com precisão as fronteiras de seções, como artigos, anexos e blocos de assinatura, em nível de linha e atribuir rótulos semânticos consistentes a cada segmento. O modelo deve lidar com classes raras e com a sobreposição de segmentos hierárquicos, mantendo bom desempenho em métricas de classificação e segmentação mesmo com um conjunto limitado de contratos anotados manualmente.

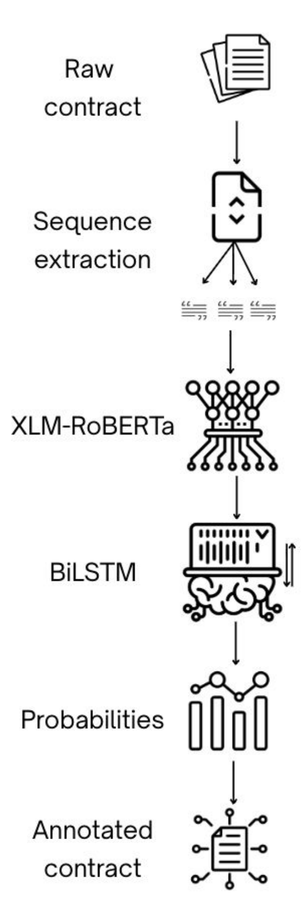
DATA AUGMENTATION



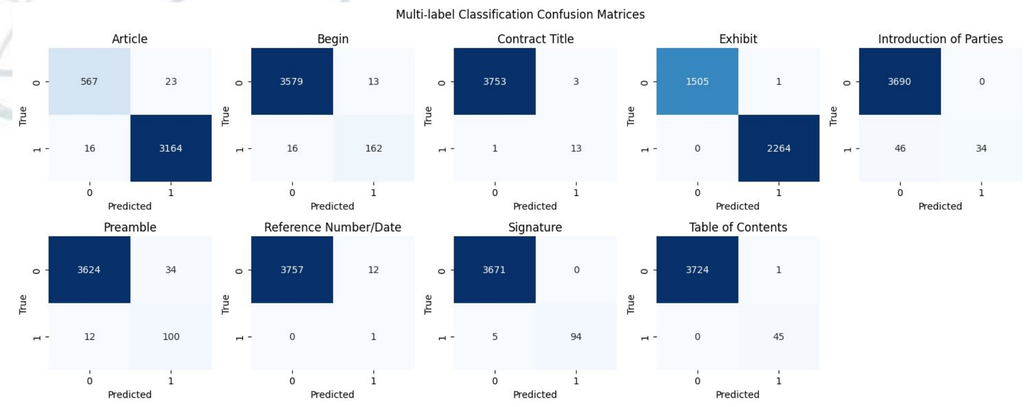
Para contornar a escassez de contratos anotados, foi criada uma etapa de aumento de dados baseada em LLM. Um modelo GPT-4o recebe trechos de contratos não rotulados, exemplos gold e um esquema rígido de entrada e saída, e produz rótulos *silver* para cada linha. Esses rótulos automáticos são validados e combinados ao conjunto gold, ampliando o treino sem custo proporcional de anotação humana e transferindo o conhecimento do LLM para o bi-encoder menor e mais eficiente usado em produção.

ARQUITETURA DO MODELO

A solução é um bi-encoder leve em duas etapas. Primeiro, um modelo XLM-RoBERTa multilíngue gera embeddings de cada linha do contrato, capturando o contexto semântico local. Em seguida, um codificador BiLSTM modela a sequência completa de sentenças e produz, para cada linha, probabilidades de início de segmento e de rótulos como Título do Contrato, Preâmbulo, Artigo, Assinatura e Anexo. A saída é *multilabel*, com ativações independentes para cada categoria, permitindo que uma mesma linha acumule diferentes funções no documento. Dessa forma, preserva-se a estrutura sobreposta típica de contratos extensos e o modelo fica apto a alimentar etapas posteriores de busca, revisão e extração de cláusulas.



RESULTADOS



Os experimentos em um corpus bilíngue de contratos em inglês e francês mostram que as anotações *silver* geradas pelo LLM aumentam claramente o desempenho do modelo. O F2 ponderado de segmentação sobe de 0,16 (apenas dados *gold*) para 0,51 com o conjunto aumentado. Na classificação semântica, o modelo atinge F2 ponderado de 0,95, mantendo fronteiras coerentes e rótulos consistentes ao longo do contrato.

CONCLUSÕES

O trabalho mostra que é possível segmentar e rotular automaticamente contratos jurídicos extensos a partir apenas do texto simples, em diferentes idiomas. A combinação de LLM para geração de rótulos *silver* com o bi-encoder de segmentação produz contratos estruturados, prontos para apoiar análise, revisão e extração de cláusulas em sistemas jurídicos automatizados.

Integrante: Rafael Sandrini Guaracho

Professor(a) Orientador(a): Profª Tereza Cristina Melo de Brito Carvalho