Extração de Entidades Médicas em Textos Clínicos

Vinicius Cantanhede dos Santos

Resumo do Projeto para NeuralMed

Março de 2025

1 Visão Geral

Este projeto foi desenvolvido como parte de um desafio para uma vaga de cientista de dados júnior na NeuralMed, empresa especializada em IA aplicada à saúde. O objetivo principal foi criar um sistema capaz de extrair informações estruturadas de documentos médicos não estruturados, facilitando a classificação de pacientes por doenças e condições médicas. O projeto utiliza técnicas de Processamento de Linguagem Natural (NLP) e Reconhecimento de Entidades Nomeadas (NER) para identificar e categorizar elementos importantes em textos médicos, como medicamentos, procedimentos, dispositivos, condições do paciente e resultados de exames.

2 Objetivos

Desenvolver um pipeline completo para processamento de documentação médica

- Implementar um modelo NER específico para terminologia médica em português.
- Avaliar a eficácia do modelo na extrações de informações médicas relevantes.
- Demonstrar a aplicabilidade prática do sistema para casos reais.

3 Metodologia

O projeto segue uma abordagem estruturada em várias etapas:

1. Pré-processamento especializado:

- Limpeza de texto em múltiplos formatos (HTML, RTF, texto puro).
- Normalização adaptada ao contexto médico (remoção de acentos, padronização).
- Tokenização especializada com foco em terminologia médica.
- Identifica padrões de horário e segmenta o texto cronologicamente.
- Função unificada que detecta o formato do texto e aplica a limpeza apropriada.

2. Anotação automática de entidades:

- Desenvolvimento de dicionários médicos abrangentes para cinco catetgorias principais: medicamentos, procedimentos, dipositivos e condições
- Implementação de regras contextuais para detecção de padrões específicos.
- Esquema de anotação no formato BIO (Beginning-Inside-Outside).

3. Desenvolvimento do modelo:

- Prepara os exemplos anotados para o formato da biblioteca Hugging Face, dividindo em treino (80%) e validação (20%).
- Alinha as etiquetas com os tokens gerados pelo tokenizer do modelo.
- Fine-Tunning de modelo transformer pré-treinado em português (neuralmind/bert-base-portuguese-cased).
- Adaptação para tarefas de NER com camada de classificação de tokens.
- Treinamento com dataset anotado de documentos médicos reais.

4. Avaliação e Predição:

- Avaliação do modelo em cenários clínicos representativos.
- Visualização intuitiva das entidades detectadas.
- Estruturação das informações extraídas para uso em aplicações médicas.

5. Sistema de reconhecimento de entidades:
Identificação de 5 categorias principais de entidades médicas:

- MEDICAMENTO: substâncias farmacêuticas (ex: "ceftriaxona", "insulina").
- PROCEDIMENTO: intervenções e ações clínicas (ex: "teste glicêmico", "coleta de exames").
- DISPOSITIVO: equipamentos e materiais médicos (ex: "cateter O2", "acesso venoso").
- CONDIÇÃO-PACIENTE: estado do paciente (ex: "calmo", "orientado", "dispneico").
- RESULTADO-EXAME: valores de exames clínicos (ex: "145 mg/dl", "36.8°C")

6. Pipeline de Treinamento Integrado:
Foi implementado um pipeline completo (train_medical_ner_pipeline) que integra todas as etapas, desde o processamento até a avaliação final. O pipeline foi executado com uma amostra de 3.000 documentos, resultando em:

- 2.991 documentos processados.
- 2.287 documentos com entidades anotadas.
- 11.094 entidades identificadas, distribuídas em: PROCEDIMENTO: 40.9%, DISPOSITIVO: 22.3%, CONDIÇÃO_PACIENTE: 19.9%, MEDICAMENTO: 10.1%, RESULTADO_EXAME: 6.8%.

O modelo foi treinado por 3 épocas, com perda de validação final de 0.018560.
7. Uso do Modelo em Aplicações Reais:
Foi implementada a função extract_medical_info para extrair e organizar informações médicas por categoria, demonstrando a aplicabilidade do modelo em cenários reais.
8. Implementação de Extrações via Prompt:
Como alternativa ao modelo NER, foram implementadas funções para extrair informações médicas usando prompts estruturados para modelos de linguagem (LLMs):

- **create_medical_extraction_prompt:** Cria prompts específicos para diferentes tipos de extração (medicamentos, exames, diagnósticos).
- **extract_via_prompt:** Extrai informações usando um serviço de LLM (simulação com respostas predefinidas).
- **compare_extraction_methods:** Compara os resultados obtidos via NER e via prompt.

9. Resultados e comparações:
Os resultados mostraram que:

- O modelo NER é eficiente para identificar entidades básicas como medicamentos, dispositivos e condições do paciente.
- As extrações via prompt (LLM) são mais detalhadas, extraindo informações estruturadas como dosagens, vias de administração e frequências de medicamentos.
- A abordagem combinada permite extrair informações completas e estruturadas de textos médicos não estruturados.

4 Diferenciais do Projeto

- **Especialização para português médico brasileiro:** Diferente da maioria das soluções disponíveis, que são focadas em inglês
- **Abordagem híbrida:** Combinação de técnicas baseadas em regras e aprendizado profundo.
- **Demonstração prática:** Exemplos concretos de uso em cenários clínicos reais.
- **Código modular e extensível:** Facilidade para adicionar novas categorias de entidades e regras

5 Conclusão

Este projeto demonstra uma solução prática e eficiente para um dos desafios centrais na aplicação de IA na saúde: a extração de informações estruturadas a partir de texto clínico livre. A abordagem desenvolvida combina conhecimento de domínio médico com técnicas modernas de NLP, resultando em um sistema que poderia ser facilmente adaptado e integrado aos produtos da NeuralMed, potencializando sua capacidade de processar e analisar documentação médica em larga escala.

A combinação de um modelo NER treinado para identificação básica de entidades com prompts estruturados para LLMs oferece uma solução robusta para extração de informações médicas, aproveitando os pontos fortes de ambas as abordagens.

Este projeto demonstra minha capacidade de implementar um pipeline completo de NLP para o domínio médico, aplicando conhecimentos técnicos de processamento de linguagem natural a um problema real e relevante. A solução é escalável, interpretável e pronta para ser refinada e expandida conforme as necessidades específicas da NeuralMed.

Estou entusiasmado com a possibilidade de contribuir para o desenvolvimento de soluções de IA que possam ter impacto real na qualidade do atendimento ao paciente e na eficiência dos processos de saúde.