Técnicas de Processamento de Linguagem Natural como ferramenta para aprimorar o recrutamento de pacientes oncológicos em ensaios clínicos

Lucas Emanuel Silva e Oliveira[1], Luiz Henrique Pereira Niero[1], João Vitor Andrioli de Souza[1], Nicolas Henrique Borges[1], Lara Zimmermann[1], Gustavo Caetano Giavarini[1], José Mendes da Silva Neto[2], Josiane Mourão Dias[2], Dayana Mendes Ribeiro[2], Daniel D'Almeida Preto[2]

1. Comsentimento, NLP Lab, São Paulo, SP, Brasil
2. Hospital de Câncer de Barretos, Barretos, SP, Brasil

Introdução

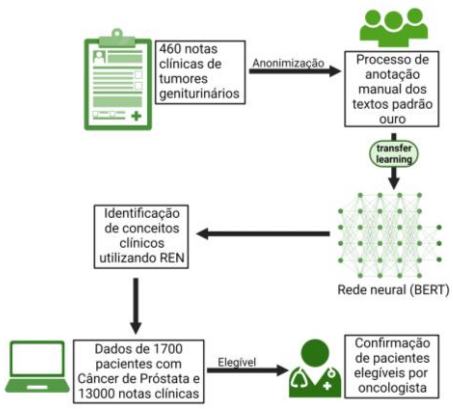
Ensaio clínico (EC) intervencionista é o principal meio de avaliar a segurança e eficácia de uma terapia na prática clínica. Uma das etapas fundamentais cerca-se da correta identificação de indivíduos elegíveis ao recrutamento, processo oneroso e dependente de profissionais para efetuar busca manual ativa nos prontuários, e que é considerado um dos principais gargalos no desenvolvimento de pesquisas clínicas (CHAUDHARI et al., 2020). Nesse contexto, técnicas de **Processamento de Linguagem Natural** (PLN) e **Machine Learning** (ML) podem apoiar a extração de dados contidos no Registro Eletrônico de Saúde (RES) (ASSALE et al., 2019; JUHN; LIU, 2020; SPASIC; NENADIC, 2020), e com isso, aprimorar o processo de identificação de pacientes elegíveis.

Neste trabalho, desenvolvemos um método baseado em PLN e ML para: (i) estruturar dados textuais do RES de pacientes oncológicos, e (ii) identificar potenciais participantes para estudos clínicos em processo de recrutamento.

Casuística e Métodos

Foram selecionadas 460 notas clínicas do setor de tumores geniturinários de um hospital oncológico, que foram submetidas a um processo de anonimização, para remoção de informações que remetem a identificação dos pacientes. Visando à identificação de conceitos clínicos no texto, através de um algoritmo conhecido por **Reconhecimento de Entidades Nomeadas** (REN) (SOUZA et al., 2020), foi realizado um processo de anotação manual dos textos e geração de um padrão-ouro (seguindo as categorias apresentadas na Tabela 1). Então, foi treinado um **modelo de rede neural supervisionado**, baseado na arquitetura BERT (DEVLIN et al., 2019), que permite a execução da técnica de ***transfer learning*** e consequente ajuste fino de modelos pré-treinados para o contexto clínico *(SCHNEIDER et al., 2020)*. O modelo, acrescido com algumas camadas adicionais de extração de dados, gerou uma representação estruturada de um conjunto de dados composto por 1700 pacientes com câncer de próstata (CaP), contendo aproximadamente 13000 notas clínicas. O protocolo experimental visou avaliar tanto o processo de estruturação dos dados (REN), levantado através das métricas de precisão, revocação e F1; quanto à classificação de elegibilidade para um EC de CaP avançado, realizada por um oncologista que analisou os pacientes sugeridos pelo sistema como potencialmente elegíveis.

Figura 1 – Visão geral do método desenvolvido



Fonte: os autores, 2022

Resultados

Para o modelo de estruturação de dados clínicos (REN), utilizando uma estratégia de exact match, obteve-se um F1 médio de 0.75, considerando todas as categorias de conceitos clínicos; e um F1 máximo de 0.89, no caso da extração de substâncias e medicamentos. Os resultados detalhados por tipo de dados são apresentados na Tabela 1. Já o sistema de recrutamento de pacientes, ilustrado na Figura 2, identificou 75 pacientes potencialmente elegíveis (aproximadamente 4.5% do total de pacientes do conjunto), sendo 60 destes corretamente identificados, representando uma acurácia de 80% na identificação de pacientes elegíveis. Considerando dados retrospectivos do processo de recrutamento para o estudo avaliado, **observamos que 34 dos pacientes elencados não haviam sido identificados pela equipe médica, quando aplicados os métodos tradicionais de busca por pacientes elegíveis**, ou seja, o método aqui utilizado conseguiu ampliar substancialmente a cobertura de pacientes potencialmente elegíveis ao estudo. Nove pacientes já haviam sido incluídos no EC quando feita análise manual dos prontuários, e 17 destes, apesar de apresentarem os critérios de elegibilidade compatíveis, não foram incluídos por outros motivos (e.g., 2º tumor primário, COVID, atendimento via teleconsulta).

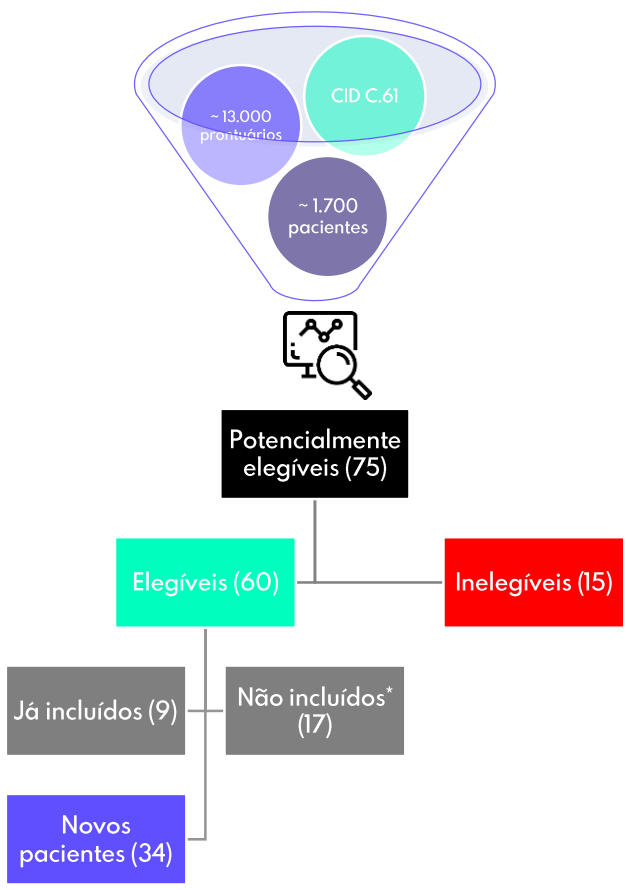
Resultados

Tabela 1 – F1-score obtido pelo algoritmo de REN por categoria de entidade clínica

Anatomic location	0,504
Condition	0,659
Drug substance	0,890
Measurement	0,873
Observation	0,838
Procedure/Device	0,775
Qualifier/Modifier	0,605
Temporal constraint	0,882

Fonte: os autores, 2022

Figura 2 – Visão geral do conjuntos de dados e pacientes selecionados



Fonte: os autores, 2022

Conclusões

As técnicas de PLN e ML para extração e estruturação de informações de textos clínicos embasou o desenvolvimento de um sistema para identificação de elegibilidade em um EC de CaP avançado. O emprego de um modelo híbrido de REN, apoiado por uma arquitetura que permite o ajuste fino de grandes modelos de linguagem para o contexto médico, aliado a um esforço moderado na criação de um padrão-ouro, podem **acelerar o processo de recrutamento para estudos clínicos**, pois conseguem filtrar grande parte dos pacientes passíveis de análise. Ademais, o emprego destas técnicas podem **incrementar substancialmente o volume de pacientes participantes**, que nos dias atuais é um dos maiores impeditivos para a realização de estudos clínicos.

Referências