Processamento de linguagem natural (NLP) não supervisionado na identificação de pacientes privados suspeitos de infeção por covid-19 para promoção a saúde

Silva, Rildo Pinto da; Pollettini, J.T.; Pazin-Filho, A; Metrus Instituto de Seguridade Social – São Paulo – SP

OBJETIVOS: O tratamento das informações em saúde é um desafio pelo seu volume crescente, conter dados faltantes, multidimensionais, não estruturados e imprecisos. Existe oportunidade para análise de informações não estruturadas na operadora. As autorizações prévias (APs) tem campo textual com o motivo da solicitação do pedido e oferecem uma oportunidade de seleção precoce dos pacientes. A pandemia de Covid-19 trouxe desafios importantes entre eles a identificação precoce de pacientes com Síndrome Pós_Covid – forma prolongada da doença que acomete 10% a 30% dos pacientes e que precisam tratamento multidisciplinar. Desenvolvemos um método de NLP para identificação precoce de pacientes com síndrome pós-covid que se beneficiassem de um programa de gestão de casos crônicos.

MÉTODOS: É um estudo observacional descritivo, baseado em dados secundários (APs). Foram selecionadas APs de 01/09/2019 a 30/06/2022 e excluídas as não preenchidas totalizando 184.371 APs incluídas neste estudo. Sobre a variável de interesse foi aplicado o modelo BERTopic. Trata-se de um algoritmo não-supervisionado para modelagem de tópicos. Esse algoritmo, resumidamente, faz a conversão vetorial do texto de cada AP usando o modelo Bidirectional Encoder Representations from Transformers, reduz a dimensionalidade dos vetores, aplica clusterização nesses vetores e, para nomear os tópicos, a técnica frequência do termo/frequência inversa dos documentos (TF-IDF). O resultado são grupos de APs classificados em tópicos – o chamado topic modeling. Foram utilizados diferentes parâmetros de fine tuning. Descrevemos o resultado do modelo com mais de 1.000 APs por tópico.

RESULTADOS: Somente 587 (0,3%) das APs tinham a identificação de infecção por Covid-19 pelo CID. 90,7% corresponderam a tratamentos clínicos, sendo 81,0% ambulatoriais. Os atendimentos clínicos em internação corresponderam a 15.741 autorizações (8,5%). O modelo foi capaz de identificar 1.987 APs (1,9%) de pacientes com suspeita de infecção por covid-19 com gasto total de R$ 20,3 milhões (5,4%), ou seja, custo médio de R$ 10.205 por AP. A análise manual de 100 casos principais mostrou correção da classificação de 70% - outros 20% dos casos foram infecções respiratórias graves que poderiam estar relacionadas a doença. Um grupo relevante de APs (661) não seriam identificadas por métodos tradicionais de pesquisa. O modelo foi ainda capaz de identificar, embora não fosse o interesse principal, casos graves de pacientes em tratamento de câncer (1.500 APs e R$ 6 milhões gasto) e doenças ortopédicas (4.531 AP e R$ 13,7 milhões de gasto).

DISCUSSÃO: As operadoras têm o desafio de atender seus beneficiários com custo-efetividade. Não há possiblidade de acréscimo nos custos administrativos. Assim métodos de identificação de clusteres de pacientes que se beneficiem ao máximo de programas de gestão de saúde tem um papel crucial no uso eficiente e correta alocação dos recursos e na disponibilização do melhor tratamento para quem realmente precisa (right care).