Análise da Influência de Características Textuais no Processo de Automatização da Regulação Médica

Benjamim de Pinho Sabino[1], **Rafael T. Anchiêta**[2], **Raimundo S. Moura**[1]

[1]Departamento de Computação – Universidade Federal do Piauí (UFPI)
Teresina – PI – Brasil

[2]Instituto Federal do Piauí (IFPI)
Picos – PI – Brasil

{benjamim.sabino,rsm}@ufpi.edu.br, rta@ifpi.edu.br

Abstract. *This paper describes an investigation of the influence of textual features on Machine Learning (ML) models for predicting response to medical requests from a Health Insurance/Plan Companies. We used NLP techniques in the pre-processing stage to clean and normalize the clinical data, in addition to retrieving acronyms and technical terms in the area. We investigated two supervised to classify exam requests into two classes: Approved and Rejected. Preliminary results show moderate accuracy for Naive Bayes (versions: MultinomialNB and BernoulliNB) and LinearSVC algorithms, being 71%, 70%, and 72%, respectively.*

Resumo. *Este artigo descreve uma investigação da influência de características textuais em modelos de Apredizagem de Máquina (AM) para a predição da resposta de solicitações médicas de uma Operadora de Planos de Saúde (OPS). Usou-se técnicas de Processamento de Linguagem Natural (PLN) na etapa de pré-processamento para limpeza e normalização dos dados clínicos, além de recuperação de siglas e termos técnicos da área. Neste trabalho, investigou-se, especificamente, dois algoritmos de AM supervisionados para classificar as solicitações de exames em duas classes: Aprovada e Recusada. Como resultado, obteve-se 71%, 70% e 72% de acurácia para os algoritmos Naive Bayes (versões: MultinomialNB e BernoulliNB) e LinearSVC, respectivamente.*

1. Introdução

A Constituição Federal [BRASIL 1988], além de garantir o direito dos cidadãos à saúde como uma atribuição do Estado, também assegurou a oferta de serviços de assistência à saúde pela iniciativa privada, sob o controle do Estado. Porém, somente em 1998, a lei 9.656 [Brasil 1998] definiu as regras para o funcionamento do setor de saúde suplementar, e deu algumas garantias aos usuários, como proibir a rescisão unilateral de contratos e submeter ao governo os índices de reajuste anuais. Em 2000, Agência Nacional de Saúde Suplementar (ANS) foi criada, com o objetivo de colaborar com a regulamentação do setor [Brasil 2000].

Segundo dados da ANS[1], órgão governamental que regula o setor de planos de saúde privados no Brasil, em 2022, o setor de planos de saúde alcançou resultados significativos em números de beneficiários. De acordo com o levantamento da ANS, no mês de

[1]https://www.gov.br/ans/pt-br

dezembro, o setor totalizou 50.493.061 usuários em planos de assistência médica, maior número desde dezembro de 2014, conforme mostra a Figura 1.

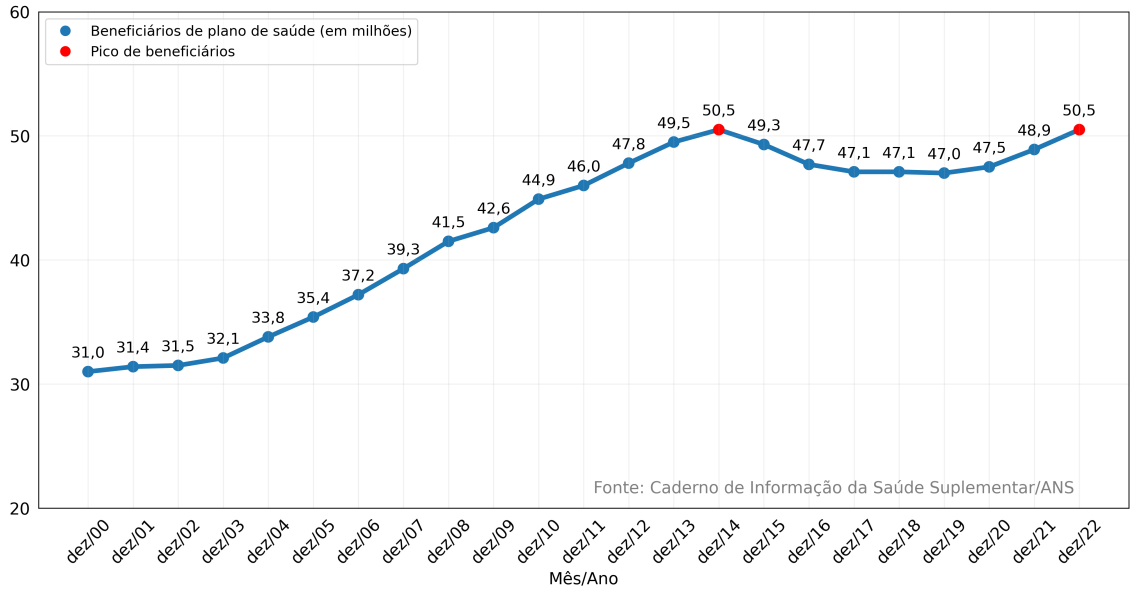

Figura 1. Beneficiários de planos de saúde privados no Brasil (2000-2022).

Mesmo com altos indicativos de movimentação financeira na área, muitas empresas operadoras de plano de saúde (OPS) enfrentam dificuldades financeiras devido a procedimentos desnecessários, fraudes ou abusos na utilização dos serviços de saúde. Isto gera um imenso problema para o funcionamento dessas empresas. Entende-se por fraude a produção intencional de informações falsas por uma entidade ou indivíduo, sabendo-se que essas informações falsas resultarão em algum benefício para essa entidade, indivíduo ou terceiros. Já abusos, no âmbito da assistência à saúde, podem ser entendidos como práticas que são inconsistentes com os critérios médicos e administrativos pré-estabelecidos [Kose et al. 2015].

Com a finalidade de reduzir gastos desnecessários, um dos mecanismos utilizados pelas OPS foi a regulação médica, onde uma análise prévia de cada solicitação recebida é feita antes de respondê-la. Dessa forma, as empresas obtêm maior controle sobre os procedimentos solicitados, quais foram aprovados ou recusados e qual a justificativa para a tomada de decisão. No contexto da regulação, técnicas de Mineração de Dados [da Silva et al. 2016] e Aprendizado de Máquina [Mitchell 1991] têm sido exploradas para automatizar esse processo.

Este artigo tem como objetivo principal investigar a influência de dados textuais em modelos de predição da resposta de solicitações médicas, se aprovada ou recusada. Para atingir esse objetivo, os modelos foram analisados, utilizando dados textuais (CID-10 com a descrição completa da doença e informações do quadro clínico).

Além dessa seção introdutória, o restante do artigo está organizado da seguinte maneira. A Seção 2 apresenta uma investigação prévia da literatura sobre trabalhos que possuem semelhança com este artigo. A Seção 3 descreve de maneira detalhada a metodologia proposta para ser empregada durante a execução da pesquisa. Em seguida, na Seção 4, é feito um detalhamento sobre a execução dos experimentos em conjunto com

a descrição dos resultados apresentados. Por fim, a Seção 5 apresenta as conclusões obtidas dos experimentos sob a luz do objetivo proposto e dos resultados gerados, fazendo considerações sobre a contribuição desse artigo para trabalhos futuros.

2. Trabalhos Relacionados

A área de análise e mineração de dados textuais está em constante evolução e diversos trabalhos têm sido publicados na literatura especializada. PLN está presentes em diversos contextos, mostrando avanços no estudo da eficácia das técnicas de extração e abrindo novas possibilidades para seu uso na área médica. No trabalho recente de [Pires et al. 2022], os autores utilizam documentos legais e de registro para extrair informações contidas neles. O trabalho avaliou a pertinência de substituir a extração utilizando métodos a nível de *token* por modelos *seq2seq*, que realizaram a extração através de (*Question Answering - QA*), concluindo que o uso do modelo *seq2seq* pode substituir parte do processo clássico de extração de informações usando *tokens*.

No trabalho de [Benicio 2020], o autor criou uma ferramenta para estruturação de textos da área da saúde que se provou estatisticamente bem sucedida, dando forte indicação que os dados textuais são relevantes para a extração de informações. Porém, o autor não fez a utilização das informações extraídas para a realização de atividades de auxílio de decisão médica.

[Lucini et al. 2017] usaram técnicas de mineração de textos com o objetivo de tentar prevê a necessidade de um paciente de ser internado, usando como base seu histórico de análises médicas. Eles usaram as representações de palavras: binária, TF [1] e TF-IDF [2]. Analisaram também o uso de unigrama, bigrama e trigramas para a formação de *features*. O módulo de predição testou oito algoritmos clássicos de AM e a melhor performance obtida foi 77,7% de *F1-score* com o algoritmo *Nu-Support Vector Machine*.

Quanto aos estudos relacionados à capacidade de previsão baseada em dados médicos, [Bertozzo 2022] faz um estudo que propõe a tentativa de prever a falta de uma paciente à consulta médica previamente agendada usando o *dataset* de consultas do Sistema Único de Saúde (SUS), porém sem fazer uso de dados textuais. Usando métodos de classificação, o autor obteve um grau de certeza de 80% na previsão, indicando a pertinência da previsão na área médica.

No trabalho [Hasan et al. 2020], há esforços na tentativa de encontrar um modelo de redes neurais capaz de classificar relações entre conceitos médicos contidos em campos textuais. Os autores fizeram uso de múltiplos modelos de redes neurais do estado da arte.

Os trabalhos [de Araújo 2014] e [Magalhães Jr 2019] representam esforços na tentativa de descobrir conhecimentos em bases de dados para a regulação médica/odontológica em operadoras de planos de saúde. Porém, o primeiro não considerou nenhuma informação dos quadros clínicos escritos pelos médicos, deixando esse gap como trabalhos futuros. O segundo investigou a influência de características textuais, considerando modelos clássicos de AM supervisionada (*Naive Bayes*, *J48* e *Random*

[1]Term Frequency: indica a frequência que determinado termo aparece no documento.

[2]Term Frequency-Inverse Document Frequency: frequência relativa obtida pela razão entre a frequência com que determinado termo aparece no documento pela frequência inversa desse mesmo termo no conjunto de documentos.

Forest) e uma base de dados disponibilizada por uma operadora de planos de saúde. Os autores concluem que o uso das características textuais influenciou positivamente os resultados, melhorando a tarefa de classificação de laudos clínicos.

Dos trabalhos descritos anteriormente, os dois últimos estão fortemente relacionados com a nossa pesquisa. Porém, o *dataset* utilizado nos experimentos é diferente, sendo mais atual que os avaliados. Além disso, no nosso trabalho incluímos a descrição geral da doença (CID-10), como *feature* adicional para ser analisada, o que não foi feito nos trabalhos anteriores.

3. Método Proposto

Para avaliarmos a influência dos dados textuais na tarefa de predição da resposta de solicitações médicas, investigou-se os dados textuais das descrições das doenças (CID-10) concatenado com as informações do quadro clínico do paciente. Para esse processo foi usado a linguagem python com o auxílio da biblioteca pandas. Os textos foram representados usando a técnica de *Bag-of-Words (BoW)*, com TF e TF-IDF. Nas representações, algumas técnicas de PLN para pré-processamento e extração de informações foram utilizados. A Figura 2 mostra uma visão geral do método proposto.

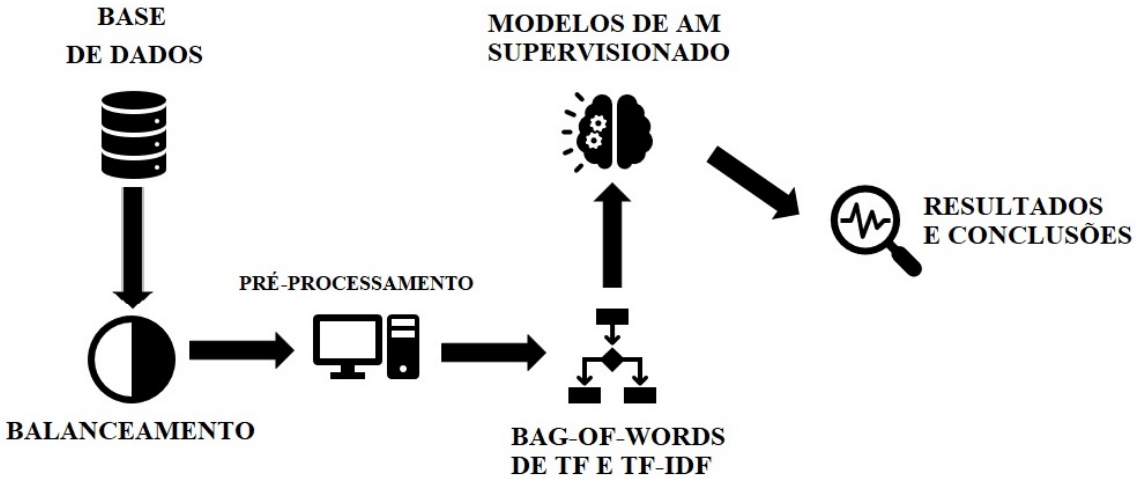

Figura 2. Etapas do método proposto.

A base de dados foi disponibilizada por uma operadora de plano de saúde atuante no estado do Piauí, Brasil, sendo fornecidos 19.231 registros de solicitações, distribuídos em três arquivos de texto em formato csv, cada um deles com campos de características próprios e com quantidade de registros diferentes. Esses arquivos foram padronizados, isto é, apenas os campos de características que estavam presentes em todos os arquivos foram mantidos, e então unificados em um único arquivo em formato csv, contendo as seguintes informações: CID-10[1], quadro clínico, sexo, idade e a resposta (aprovado ou recusado). Destaca-se que o CID-10 (descrição da doença) e o quadro clínico são as informações textuais que serão analisadas.

No processo de padronização dos dados, foi necessário utilizar uma fonte externa para fornecer informações sobre o CID de maneira estruturada. Para isso foi utilizada a ta-

[1]Classificação Internacional de Doenças, denominada, após a 10ª revisão, como Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde

bela da CID-10 do sistema DATASUS[2]. Além disso, foi usada a tabela de siglas contida no manual do Hospital de Clínicas da Universidade Federal Do Triângulo Mineiro (UFTM)[3], para a substituição de siglas e abreviações presentes nos quadros clínicos. Portanto, a descrição padronizada do CID-10 foi concatenada com o quadro clínico do paciente para formar a descrição textual de um registro.

Durante o processo de unificação, cada arquivo de registros foi aberto usando a biblioteca pandas e seus dados foram acomodados em uma estrutura de dados do pandas chamada DataFrame, semelhante a uma tabela bidimensional e, após selecionadas as colunas de características oportunas, as informações contidas nas colunas de características de todos os registros individuais foram inseridas em um novo DataFrame, de forma a padronizar os dados dos registros. Posteriormente esse DataFrame foi usado para gerar o arquivo csv unificado. Esse processo é ilustrado na Figura 3. Após a unificação dos dados, a base de dados ficou com 8.379 solicitações recusadas e 10.852 aprovadas.

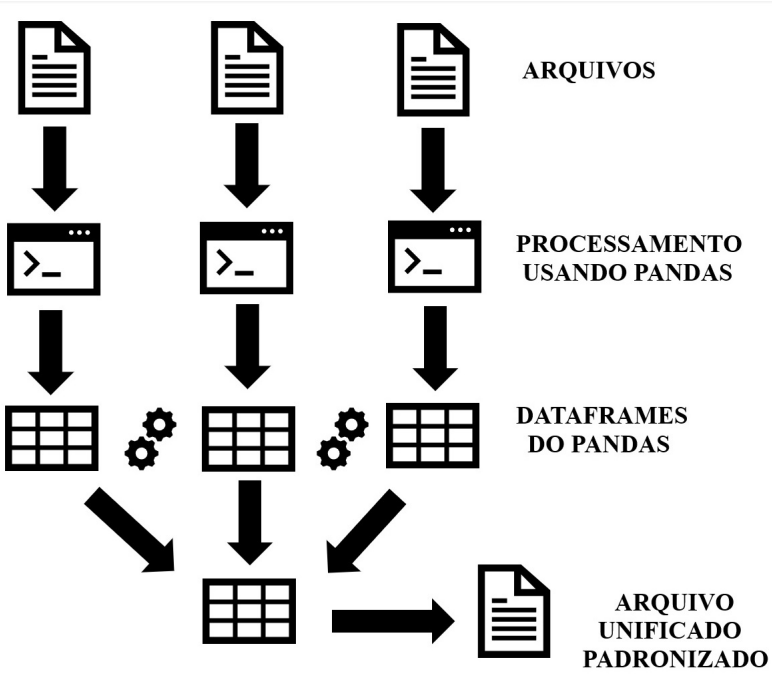

Figura 3. Processo de unificação e padronização.

Para o bom funcionamento dos modelos de predição, foi necessário remover registros que apresentavam-se repetidos com as respostas da solicitação conflitantes, dando-se preferência ao rótulo de solicitação 'aprovada'. Além disso, alguns registros aleatórios das solicitações 'aprovadas' foram retirados, realizando a subamostragem dos dados. O objetivo principal foi balancear a quantidade de registros de cada classe, o que produziu uma base de dados com 14.804 solicitações, sendo 7.402 de cada classe.

Como os dados textuais das solicitações são compostos por texto de escrita livre, foi preciso realizar um pré-processamento para estruturar as informações. O pré-

[2]Disponível em http://www2.datasus.gov.br/cid10/V2008/cid10.htm

[3]Tabela disponível em https://www.gov.br/ebserh/pt-br/hospitais-universitarios/regiao-sudeste/hc-uftm/documentos/manuais

processamento proposto baseou-se na substituição de siglas e abreviações do campo medicinal, seguido da retirada de *stopwords* e pelos processos de tokenização e consequente lematização dos *tokens*, gerando uma *bag-of-words* com os elementos léxicos de maior relevância no texto. A relevância será avaliada a partir de medidas estatísticas como TF e TF-IDF, respectivamente, *Term Frequency* e *Term Frequency-Inverse Document Frequency*. Na Tabela 1 são apresentados exemplos dos dados textuais antes e depois do pré-processamento, compostos pelo CID-10 padronizado, concatenado com o quadro clínico do paciente.

Tabela 1. Exemplo de registros de dados: original e pré-processado

Classe	Registro	Registro pré-processado
APROVADA	"M751 - Síndrome do manguito rotador; paciente com historico de queixa algica no ombro esquerdo sem melhora com tratamento conservador"	"M751 síndrome manguito rotador paciente queixar algico dor ombro esquerdo sem melhora tratamento conservador"
RECUSADA	"S83 - Luxação, entorse e distensão das articulações e dos ligamentos do joelho; paciente apresenta suspeita de lesao do menisco lateral e necessita de ressonância magnética para melhor conduta"	"S83 luxação entorse distensão articulação ligamento joelho paciente apresentar suspeito lesao menisco lateral necessitar ressonância magnético para bom conduta"

Por fim, dois algoritmos clássicos de AM supervisionada foram analisados: *Naive Bayes*, nas versões *MultinomialNB*, *BernoulliNB*, e *LinearSVC*. Os experimentos e resultados obtidos são descritos e discutidos na próxima seção.

4. Experimentos

Para a execução dos experimentos, os dados balanceados, pré-processados e representados em uma *bag-of-words* foram submetidos para a realização do treinamento e teste dos modelos de aprendizado de máquina. Foram utilizados 3 modelos de AM supervisionada: *Naive Bayes* nas versões *MultinomialNB* e *BernoulliNB* e *LinearSVM*. Os modelos foram treinados, usando validação cruzada *(k-fold cross-validation)*, com 5 e 10-folds. A validação cruzada usando *k-fold* é um método para avaliar a curva de aprendizagem do treinamento dos modelos. Ela consiste na alternância da seleção do conjunto de validação entre um dos *k-folds* (conjuntos de tamanhos iguais), sendo que para cada seleção, os outros *k-1 folds* são usados para treinamento do modelo.

Com relação à representação das *Bag-of-Words*, usou-se as medidas TF e TF-IDF como entradas para cada um dos modelos escolhidos. As Tabelas 2 e 3 mostram a acurácia média de cada um dos modelos na fase de desenvolvimento, para 5 e 10-folds. Destaca-se que todos os dados foram usados durante o desenvolvimento dos modelos. Assim, um novo conjunto de dados (*test set*) será utilizado para avaliação geral dos modelos.

Os resultados mostraram uma acurácia bem próxima entre os modelos analisados, considerando as representações TF e TF-IDF e em relação ao número de *folds* na

Tabela 2. Acurácia dos modelos de AM: representação TF

Modelo	K-fold	Acurácia média	Desvio padrão da acurácia
MultinomialNB	k = 5	0,7054	0,0092
BernoulliNB	k = 5	0,7043	0,0094
LinearSVC	k = 5	0,7200	0,0063
MultinomialNB	k = 10	0,7058	0,0101
BernoulliNB	k = 10	0,7064	0,0116
LinearSVC	k = 10	0,7080	0,0088

Tabela 3. Acurácia dos modelos de AM: representação TF-IDF

Modelo	K-fold	Acurácia média	Desvio padrão da acurácia
MultinomialNB	k = 5	0,7087	0,0079
BernoulliNB	k = 5	0,7043	0,0094
LinearSVC	k = 5	0,7061	0,0085
MultinomialNB	k = 10	0,7103	0,0112
BernoulliNB	k = 10	0,7064	0,0116
LinearSVC	k = 10	0,7212	0,0089

validação cruzada (5 e 10-*folds*). Portanto, não foi possível identificar o método mais eficaz entre os usadas nos experimentos. Novos experimentos devem ser realizados no futuro próximo.

5. Conclusão e Trabalhos Futuros

Este artigo investigou a influência de características textuais em modelos de AM para a predição da resposta de solicitações médicas de uma Operadora de Planos de Saúde.

Apesar de existir muitos algoritmos de AM Supervisionada disponíveis, apenas dois algoritmos clássicos foram analisados neste trabalho. A nossa intenção é estender os resultados para modelos de *Deep Learning*.

Os resultados preliminares mostraram uma acurácia de 71%, 70% e 72% para os algoritmos *Naive Bayes (versões: MultinomialNB e BernoulliBB)* e *LinearSVC*, respectivamente. Como trabalhos futuros, incluem:

- Avaliar modelos de *Deep Learning* e as representações *Word Embeddings*;
- Estender o conjunto de dados (*Corpus* de solicitações), com novos registros;
- Fazer uma análise de erros do modelos investigados;
- Incluir outros atributos nas representações das solicitações, tais como: sexo, idade, histórico de doenças, entre outros;
- Aplicar técnicas de explicabilidade para tornar os resultados mais claros, seguros e auditáveis.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Piauí (FAPEPI) – Edital 008/2018 - PRONEM.

Referências

Benicio, D. H. P. (2020). Aplicação de mineração de texto e processamento de linguagem natural em prontuários eletrônicos de pacientes para extração e transformação de texto em dado estruturado. Master's thesis, Universidade Federal do Rio Grande do Norte (UFRN).

Bertozzo, R. J. (2022). Aplicação de machine learning em dataset de consultas médicas do SUS. Monografia, Universidade Federal de Santa Catarina (UFSC).

BRASIL (1988). Constituição da república federativa do brasil de 1988. Brasília, DF: Presidente da República.

Brasil (1998). Lei nº 9.656, de 3 de junho de 1998. *Diário Oficial [da] República Federativa do Brasil*.

Brasil (2000). Lei nº 9.961, de 28 de janeiro de 2000. *Diário Oficial [da] República Federativa do Brasil*.

da Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à mineração de dados com aplicações em R*. Rio de Janeiro: Elsevier.

de Araújo, F. H. D. (2014). Descoberta de conhecimento em base de dados para o aprendizado da regulação médica/odontológica em operadora de plano de saúde. Master's thesis, Universidade Federal do Piauí (UFPI).

Hasan, F., Roy, A., and Pan, S. (2020). Integrating text embedding with traditional nlp features for clinical relation extraction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425.

Kose, I., Gokturk, M., and Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing*, 36:283–299.

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J. L., Anzanello, M. J., Kuchenbecker, R. S., and Schaan, B. D. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics*, 100:1–8.

Magalhães Jr, G. V. (2019). Estudo da influência de características textuais no processo de automatização da regulação médica. Master's thesis, Universidade Federal do Piauí (UFPI).

Mitchell, T. M. (1991). *Key Ideas in Machine Learning*. John Wiley & Sons ltd.

Pires, R., de Souza, F. C., Rosa, G., Lotufo, R. A., and Nogueira, R. (2022). Sequence-to-sequence models for extracting information from registration and legal documents. In Uchida, S., Barney, E., and Eglin, V., editors, *Document Analysis Systems*, pages 83–95, Cham. Springer International Publishing.