Análise de Características Textuais na Automatização da Regulação Médica

Kauan Vaz do Nascimento[1], **Raimundo Santos Moura**[1]

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

{kauanvazc,rsm}@ufpi.edu.br

Resumo. *No Brasil, as Operadoras de Planos de Saúde (OPS) privadas enfrentam desafios financeiros, devido à crescente demanda por procedimentos clínicos e fraudes ou abusos na utilização dos serviços. A adoção da regulação médica foi importante, mas a manutenção de equipes especializadas para essa tarefa ainda é dispendiosa, gerando a necessidade de automatizar tal processo. Neste trabalho, investigamos oito modelos de Aprendizado de Máquina (AM) clássicos e profundos com os dados do Corpus MIMIC-CXR traduzidos para o Português. Os resultados mostram 95% de acurácia com uma RNN, além de identificar características importantes, usando o modelo LinearSVC.*

Abstract. *In Brazil, private health plans face many financial problem, due to increasing demand for medical services, and fraud or abuse in the use of services. The adoption of prior authorization was important, but the maintenance of specialized teams for this task is still expensive, generating the need to automate this process. In this work, we investigate eight Machine Learning (ML) models (classical and deep learning) with the MIMIC-CXR data translated into Portuguese. The results show 95% accuracy with an RNN, in addition to identify important features using the LinearSVC model.*

1. Introdução

Muitas pessoas têm recorrido à saúde suplementar, atividade que envolve a operação de planos e seguros privados de assistência médica à saúde. Dados da Agência Nacional de Saúde (ANS)[1], agência reguladora do mercado de planos privados de saúde, apontam que a taxa de cobertura de planos era de 25,8% em 2022, um aumento em relação ao ano anterior, que contava com uma taxa de 25,1% . Ainda assim, o setor tem piorado, com um recuo de 5,4 bilhões no resultado líquido no mesmo período.

Muitas Operadoras de Planos de Saúde (OPS) enfrentam dificuldades financeiras devido a fraudes ou abusos na utilização de seus serviços e, com a finalidade de diminuir gastos, elas começaram a utilizar um mecanismo chamado regulação, processo responsável por detectar, classificar e triar a demanda de requisições. Essa tarefa requer uma grande quantidade de profissionais para ser realizada, pois é feita manualmente e possui uma grande demanda, o que a torna custosa e impraticável para a realidade das OPS.

Uma informação importante para as operadoras na regulação é o resultado do exame ou procedimento clínico realizado, o chamado laudo clínico, documento que

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

contém informações detalhadas sobre o estado de saúde do paciente, resultados de exames laboratoriais, diagnósticos e outras observações relevantes, influenciando diretamente nas decisões das operadoras sobre a cobertura de procedimentos e tratamentos.

A análise de laudos pode ser especialmente difícil em casos de pacientes com sintomas vagos e inespecíficos, que pode indicar uma variedade de condições médicas. Nesses casos, o médico responsável precisa interpretar os resultados para chegar a um diagnóstico preciso. Esta tarefa pode ser complexa e demorada, exigindo experiência por parte do profissional de saúde.

Neste cenário, a motivação principal para explorar a área de regulação é a redução de despesas abusivas de uma OPS, especialmente no processo de regulação. De maneira geral, este estudo está focado na área de Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) que trata informações existentes em textos livres (não estruturados) e transforma-os em dados estruturados, permitindo uma manipulação inteligente das informações. O objetivo deste trabalho é aprofundar no estudo dos laudos clínicos analisando *features* linguísticas e utilizando modelos de AM clássicos e profundos, na tarefa de classificação binária de dados textuais (laudos clínicos). A intenção é classificar os laudos em duas classes: **normal** e **alterado**.

O restante do artigo está organizado da seguinte maneira: a Seção 2 apresenta os principais trabalhos da literatura que exploram o uso de características textuais no processo de regulação médica. A Seção 3 detalha a metodologia utilizada para investigar modelos de AM na tarefa de classificação de laudo clínicos. Na parte inicial da seção, destacamos o processo de coleta e anotação dos dados utilizados nos experimentos. A Seção 4 descreve os modelos de AM utilizados e apresenta e discute os resultados nos experimentos. Por fim, a Seção 5 conclui o artigo e aponta alguns trabalhos futuros.

2. Trabalhos Relacionados

Na literatura especializada, muitos trabalhos têm explorado o uso de aprendizagem de máquina supervisionada para a tarefa de classificação. Porém, poucos discutem o processo de regulação médica e uma quantidade ainda menor analisa as características textuais presentes em documentos médicos. Muitos artigos estão relacionados às admissões automáticas em serviços de emergência e tiveram grande influência nesta pesquisa, bem como artigos sobre AM que não necessariamente consideram atributos textuais. Nesta seção apresenta-se alguns dos principais trabalhos relacionados.

Em [Magalhães Junior et al. 2019], os autores exploraram os efeitos de características textuais na regulação médica automática através do uso de métodos de AM e PLN, em que características do paciente e informações textuais do quadro clínico foram utilizadas como entrada para que tais dados fossem classificados em duas classes: aprovada ou recusada. Os algoritmos escolhidos para experimentos foram *K-Nearest Neighbor* (KNN), J48, Naive Bayes (NB), *Random Forest* (RF) e *Support Vector Machine* (SVM). A melhor abordagem resultou da combinação da representação *Bag of Words (BoW)*, usando *Term Frequency–Inverse Document Frequency* (TF-IDF) com o modelo SVM em relação aos resultados dos modelos sem o uso de características textuais.

O trabalho [Graham et al. 2018] buscou uma solução para a automação de admissões em serviços de emergência, ambiente onde um fluxo controlado de pessoas é de

extrema importância e a superlotação pode causar impactos negativos para os pacientes. Os modelos foram construídos usando os algoritmos *Decision Tree* (DT), *Gradient Boosted Machine* (GBM) e *Logistic Regression* (LR). O último modelo permitiu extrair fatores relacionados às internações hospitalares como o local do hospital, idade, modo de chegada, categoria de triagem, grupo de atendimento, internação anterior no último mês e no último ano. Dentre os modelos analisados, o de melhor performance foi o GBM (precisão = 80,31% e *AUC* = 0,859).

Em [Roquette et al. 2020], os autores também lidam com os problemas da superlotação em serviços de emergência. Foram aplicadas redes neurais profundas sobre dados textuais (estruturados e não estruturados) disponíveis no momento da triagem para a previsão correta e precoce de admissões nesses serviços. Os dados consistem em descrições de exames anteriores, anotações de triagem, medicamentos utilizados e a queixa principal. Os testes foram analisados observando a medida AUC e o modelo que melhor performou consiste em uma rede neural profunda seguida de um classificador de aumento de gradiente, o *CatBoost* (AUC = 0,892).

No mesmo âmbito da medicina, [Wang et al. 2023] propõem um modelo híbrido para a análise automática de dados heterogêneos, visando agilidade na triagem entre atenção primária e secundária. A abordagem inclui três submodelos: Resultado de Exame de Sangue (RES), Documento de Encaminhamento do Médico Geral (DEMG) e o modelo híbrido. O RES é o resultado da melhor combinação entre técnicas de imputação de dados faltantes, métodos de sub-amostragem e algoritmos de aprendizagem de máquina. O DEMG faz uso do melhor modelo dentre aqueles baseados em PLN no BERT e de otimizações de *threshold*. Já o híbrido junta as predições do RES e do DEMG se o paciente possuir os dados correspondentes ou usa o respectivo modelo com dados disponíveis. A ideia proposta alcançou uma precisão de 0,83, recall de 0,82, F1-Score de 0,83. Além disso, para a classificação de pacientes nas classes: c1: com condições não inflamatórias e c2: com artrite inflamatória a precisão foi de 0,82 e AUC igual a 0,90.

O diferencial deste trabalho está na avaliação aprofundada de diferentes métodos de treinamento e classificação (algoritmos clássicos, redes neurais e modelos de língua), visando melhorar as comparações e o entendimento de quais modelos podem ser utilizados e, também, para entender a influência das características textuais no processo de regulação de uma OPS.

3. Metodologia

De maneira geral, soluções computacionais que utilizam técnicas de AM Supervisionada realizam o processo de classificação em duas etapas: treinamento (*training*) e predição (*prediction*). Os textos foram representados usando *Bag of Words (BoW)* com TF-IDF e *Words Embeddings*. Além disso, foram realizados ajustes nos hiperparâmetros, em busca do melhor modelo possível, conforme pipeline mostrado na Figura 1.

3.1. Coleta dos Dados

O *Corpus MIMIC-CXR v2.0.0*[2] é um extenso repositório de radiografias de tórax disponíveis no formato DICOM, arquivos que seguem um conjunto de normas para o arquivamento e comunicação de imagens obtidas em exames, padronizando imagens digitais

[2]https://physionet.org/content/mimic-cxr/2.0.0/

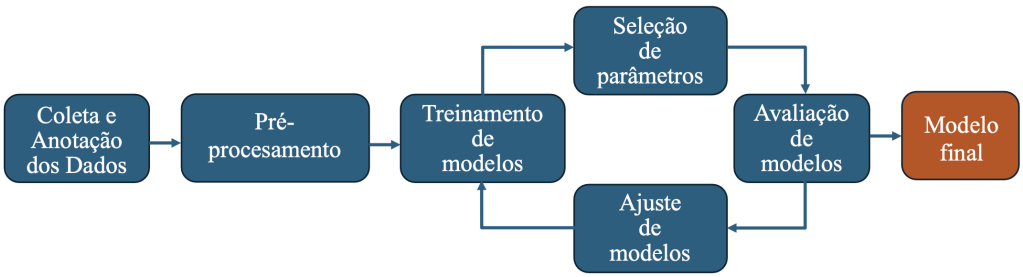

Figura 1. Pipeline: etapas do processo de treinamneto dos modelos de AM

geradas mesmo em equipamentos de diferentes fabricantes. O *Corpus* contém 377.110 imagens acompanhadas por relatórios escritos em texto livre e corresponde a 227.835 estudos radiográficos. Os dados foram extraídos do sistema de registros eletrônicos de saúde do *Beth Israel Deaconess Medical Center em Boston, MA* e foram devidamente desidentificados, garantindo a remoção de qualquer informação confidencial. No entanto, os registros não estão rotulados, o que gerou a necessidade do uso de uma outra base de dados. O *Corpus MIMIC Chest X-ray JPG v2.0.0*[3] é um estudo derivado do MIMIC-CXR v2.0.0 que disponibiliza as radiografias no formato JPG com rótulos estruturados derivados dos relatórios radiológicos, seguindo uma abordagem *multi-rótulo*, situação em que uma única imagem é associada a múltiplas classes simultaneamente.

Os rótulos foram determinados com o uso de duas ferramentas de código aberto: *NegBio*, baseada em regras para negação e detecção incerta em relatórios radiológicos, e *CheXpert*, também baseada em regras e construída em cima do NegBio. Portanto, dois arquivos CSV foram gerados, cada um contendo os resultados de cada ferramenta.

3.1.1. Anotação dos Dados

O tratamento se inicia com a transformação dos dados do formato *multi-rótulo* para o formato *multi-classe*, onde um registro possui apenas uma classe dentre várias classes existentes.

Destaca-se que os rótulos dos registros originais são constituídos de várias condições médicas (por exemplo, pneumonia e cardiomegalia) e a condição de 'nenhuma descoberta' (*no findings*). Cada rótulo contém um dos quatro valores possíveis:

- **1.0** - A condição foi mencionada positivamente e está presente em uma ou mais das imagens correspondentes;
- **0.0** - A condição foi mencionada negativamente e não está presente em nenhuma das imagens correspondentes;
- **-1.0** - A condição foi mencionada com incerteza no relatório e pode ou não estar presente na imagem correspondente, ou foi mencionada com linguagem ambígua no laudo e não está claro se a patologia existe ou não;
- **Ausente** - Nenhuma menção à condição foi feita no laudo.

Destaca-se que o valor 1.0 na condição *'nenhuma descoberta'* indica que nenhuma anormalidade foi encontrada.

[3]https://physionet.org/content/mimic-cxr-jpg/2.0.0/

Considerando a dificuldade de interpretar os rótulos atuais dos laudos, uma reestruturação foi proposta obedecendo o seguinte mapeamento:

- **1.0** - O valor 1.0 foi encontrado em alguma das condições médicas;
- **0.0** - O valor 1.0 estava presente na condição *'nenhuma descoberta'*;
- **-1.0** - O valor 1.0 não foi encontrado em nenhuma das condições médicas e nem na condição *'nenhuma descoberta'*.

Após a transformação, os laudos com rótulos comuns nos dois arquivos foram mantidos e os que divergiram foram descartados. Exclui-se também os registros com o novo rótulo igual a -1,0. No final, restaram 210.670 laudos anotados com duas classes: 0, para normal e 1, para alterado.

Em seguida, foi feita a extração do conteúdo relevante dos laudos, que está presente principalmente nas seções: *'Findings'* e *'Impression'*. Porém, na ausência de ambas, foi utilizado o texto completo do laudo com a remoção de cabeçalhos e de algumas seções previamente definidas com conteúdos irrelevantes, por exemplo, as seções com informações de datas e horas. Um exemplo de laudo e o processo de extração de conteúdo é mostrado na Figura 2

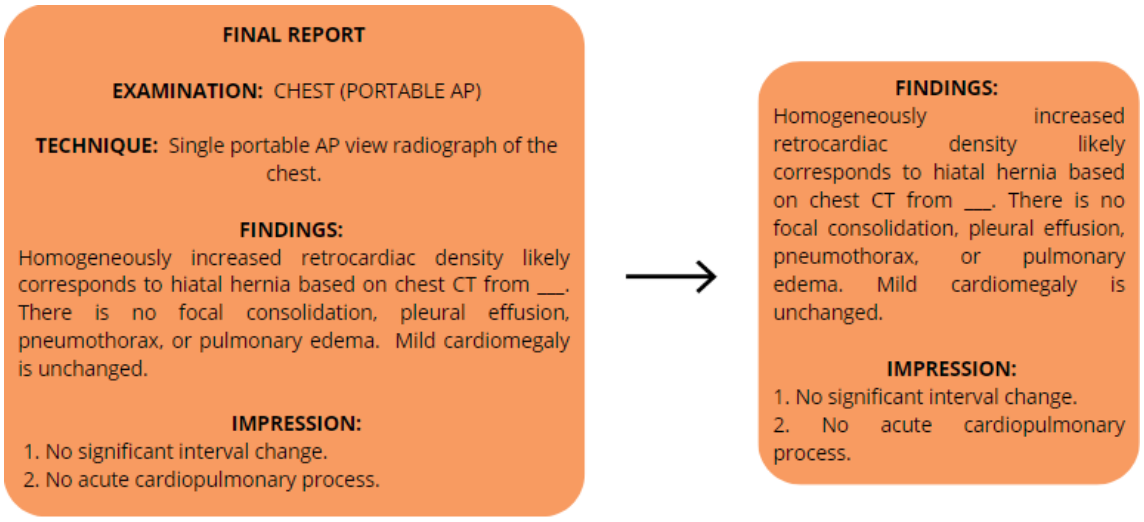

Figura 2. Extração do conteúdo relevante.

Por fim, foi aplicado o processo de *Backtranslation*, técnica que envolve a tradução de textos para outro idioma e a subsequente tradução de volta para o idioma original, introduzindo-se uma forma de se ter os dados em português enquanto possibilita a medição da qualidade da tradução através da comparação entre o conteúdo original e o conteúdo retraduzido.

Apenas 86 mil registros foram escolhidos aleatoriamente para tradução com a API do *Google Cloud Translation*[4], sendo 43 mil laudos de cada classe. Esse diminuição de dados se deu em razão de limitações técnicas no uso da ferramenta. A qualidade das traduções foi medida com a métrica *BERT Score*[5], que usa a similaridade semântica entre os textos. Como a abordagem obteve bons resultados (*média = 0.96* e *desvio padrão = 0.02*), todos os laudos foram considerados para compor o *dataset* final.

[4]https://cloud.google.com/translate
[5]https://huggingface.co/spaces/evaluate-metric/bertscore

3.2. Pré-processamento dos Dados

Nesta etapa, foram removidos numerações de tópicos, datas, horas e caracteres "_", que representavam separações de seções ou informações que foram ocultas no processo de desidentificação. Além disso, os textos foram convertidos para minúsculas e a acentuação foi removida.

Adicionalmente, uma cópia do conjunto de dados foi criada para que técnicas de pré-processamentos diferentes fossem analisadas separadamente. No conjunto usado para treinar algoritmos clássicos e profundos foram descartadas as *stopwords* e as pontuações, enquanto que no treinamento dos modelos baseados no BERT, essas informações foram mantidas, pois elas são importantes para a inferência de contexto.

3.3. Modelos de Aprendizado de Máquina

Neste estudo, foram usados 8 (oito) modelos distintos, que englobam abordagens clássicas, redes neurais profundas e modelos de língua BERT.

Para os modelos clássicos, as escolhas foram *Linear Support Vector Classification (LinearSVC)* e *Multinomial Naive Bayes*, pois esses são algoritmos indicados pelo *Scikit-learn* para classificação de textos com menos de 100 mil amostras. Já para os modelos profundos, explorou-se as *Recurrent Neural Networks (RNN)* e as *Convolutional Neural Networks (CNN)*, devido à popularidade e eficácia em uma ampla gama de problemas. As Figuras 3 e 4 mostram a visão geral das arquiteturas utilizadas.

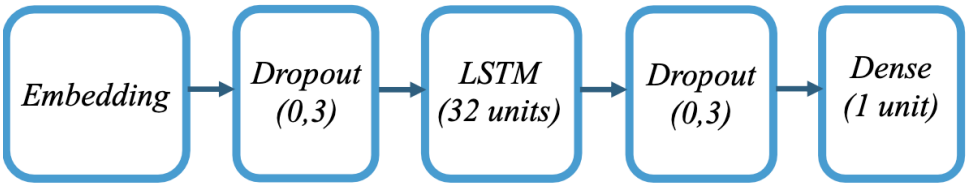

Figura 3. Arquitetura para a RNN.

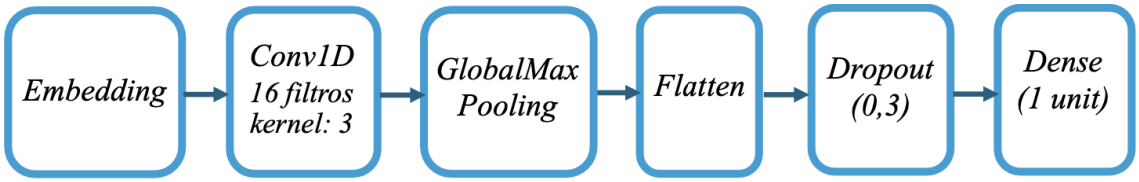

Figura 4. Arquitetura para a CNN.

Destaca-se que que algumas camadas são comuns para as duas arquiteturas, desempenhando papéis semelhantes. São elas:

- **Embedding**: Camada responsável por transformar as representações de palavras em vetores densos de números reais de tamanho fixo. Os vetores de saída possuem dimensões iguais a 64 para a RNN e 16 para a CNN. O tamanho do vocabulário foi determinado pelo número de palavras únicas no conjunto de treinamento.
- **Dropout**: Técnica de regularização frequentemente utilizada em redes neurais para prevenir *overfitting* (sobreajuste). Durante o treinamento, uma fração dos neurônios (nesse caso, 30%) foi aleatoriamente desativada, o que força a rede

aprender representações mais robustas e reduzir a dependência de neurônios específicos. Isso ajuda a melhorar a capacidade de generalização do modelo, permitindo que ele se adapte melhor a novos dados de entrada.

- **Dense**: Camada densamente conectada com uma única unidade de saída, ativada pela função de ativação *sigmoid*. Esta configuração é comum em problemas de classificação binária, onde o objetivo é prever a probabilidade de uma classe pertencer a uma determinada categoria. A função de perda utilizada foi a *Binary Cross-Entropy* e o otimizador foi o *Adam*.

A RNN ainda possui a camada LSTM, uma unidade recorrente projetada para lidar com sequências de dados, especialmente eficaz em capturar dependências temporais de longo prazo. Em contraste com as RNN tradicionais, as LSTMs possuem mecanismos internos de controle de fluxo de informação, como portões de entrada, esquecimento e saída. Isso permite que elas aprendam e memorizem padrões em dados sequenciais de maneira mais eficaz, evitando o problema de degradação de gradientes e preservando informações relevantes ao longo do tempo. Neste modelo, a camada LSTM possui 32 unidades e também incorpora uma regularização L2 com coeficiente de 0.01 para mitigar o sobreajuste.

A CNN é formada por mais algumas camadas específicas para o seu funcionamento, que comtemplam:

- **Conv1D**: Aplica filtros convolucionais unidimensionais à sequência de entrada (*embeddings das palavras*). Os filtros convolucionais aprendem padrões locais na sequência de entrada, permitindo que o modelo capture características relevantes das palavras e de suas vizinhanças. O número de filtros especificado foi 16 e o tamanho do kernel definido foi 3, o que significa que a cada passo da convolução, o filtro é aplicado a subsequências de três palavras. A função de ativação *ReLU* foi aplicada após a convolução para introduzir não-linearidade.
- **GlobalMaxPooling**: Realiza a operação de *pooling global* na saída da camada convolucional. Isso significa que, para cada filtro, o valor máximo dentro de toda a saída é retido, reduzindo a dimensionalidade dos mapas de características resultantes, mantendo as características mais importantes.
- **Flatten**: Responsável por reduzir a dimensionalidade dos dados de entrada resultando em um vetor unidimensional. Isso é necessário para que se possa conectar os recursos a uma camada subsequente densamente conectada.

Quanto aos modelos de língua BERT, foram analisadas quatro variantes distintas, utilizando a função de perda *BCEWithLogitsLoss* e o otimizador *AdamW*, como descrito abaixo:

1. **RoBERTa**[6]: Reconhecido por seu aprimoramento em tarefas de compreensão de texto em relação ao BERT original.
2. **BERTimbau**[7]: Otimiza a arquitetura BERT para proporcionar melhor desempenho em tarefas específicas da língua portuguesa.
3. **BioBERTpt**[8]: Melhora o desempenho em tarefas relacionadas à saúde e biologia.

[6]https://huggingface.co/FacebookAI/roberta-base
[7]https://huggingface.co/neuralmind/bert-base-portuguese-cased
[8]https://huggingface.co/pucpr/biobertpt-all

4. **BERT Multilingual**[9]: Projetado para processar e compreender diversas línguas.

Enquanto a escolha de modelos clássicos visa a interpretação e observação das características mais relevantes no processo de treinamento, uma vez que oferecem compreensão mais direta de seus processos internos, as demais técnicas têm o objetivo de fornecer uma análise comparativa abrangente, proporcionando percepções sobre a eficácia de diferentes abordagens na classificação de documentos médicos.

4. Resultados e Discussões

Os modelos de AM utilizados neste trabalho possuem diversos parâmetros de configuração, que desempenham um papel crucial nos resultados obtidos, podendo levar a situações de *overfitting* ou *underfitting*. Destaca-se que *overfitting* ocorre quando o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados, capturando padrões irrelevantes ou aleatórios. Já o *underfitting* acontece quando o modelo é muito simples para capturar a complexidade dos dados, resultando em baixo desempenho tanto nos dados de treinamento quanto nos de teste. Os valores atribuídos aos parâmetros foram definidos após uma extensa avaliação, utilizando dados de validação correspondentes a 10% do conjunto total de dados disponíveis. Os testes foram feitos de maneira automática com uso de *GridSearch* para os modelos clássicos e para as redes neurais e de forma manual para os modelos BERT.

A Tabela 1 mostra as melhores configurações encontradas. Valores de C para o LinearSVC e de alpha para o Multinomial Naive Bayes foram testados no intervalo [0.001, 0.1, 1, 10, 100]. Já para os demais modelos, foram testadas todas as variações com *batch size* no intervalo [5, 10, 20] e *learning rate* em [1e-5, 1e-4 , 1e-3].

Modelos	**Configurações**
LinearSVC	C = 0.1
Multinomial Naive Bayes	alpha = 0.1
RNN	batch = 5 learning rate = 1e-3
CNN	batch = 20 learning rate = 1e-3
RoBERTa	batch = 5 learning rate = 1e-4
BERTimbau	batch = 10 learning rate = 1e-5
BioBERTpt	batch = 10 learning rate = 1e-5
BERT Multilingual	batch = 10 learning rate = 1e-4

Tabela 1. Melhores configurações encontradas para cada modelo.

Das quatro variantes do BERT, apenas as duas com melhores resultados foram escolhidas para os experimentos finais. São elas: *BERTimbau* e *BioBERTpt*.

[9]https://huggingface.co/google-bert/bert-base-multilingual-uncased

4.1. Treinamento dos Modelos

Dispondo das configurações e tendo a base de dados dividida, de maneira estratificada, em 80% para o treinamento, 10% para validação e 10% para testes, foi realizado o treinamento, onde TF-IDF foi a entrada para os modelos clássicos e *Embeddings* não pré-treinadas foram a entrada para as redes neurais e para os modelos BERT. Além disso, houve o uso de *Early Stopping* para a parada do treinamento antes do modelo se ajustar demais aos dados de treinamento. Os resultados obtidos são apresentados na Tabela 2.

Modelos	**Precisão**		**Revocação**		**Medida-F1**		**Acurácia**
	0	1	0	1	0	1	
LinearSVC	0.924	0.914	0.912	0.925	0.918	0.919	0.919
Multinomial Naive Bayes	0.907	0.829	0.810	0.918	0.856	0.871	0.864
RNN	0.962	0.937	0.935	0.963	**0.948**	**0.950**	**0.949**
CNN	0.925	0.911	0.910	0.927	0.917	0.919	0.918
BERTimbau	0.941	0.932	0.930	0.943	0.939	0.942	0.940
BioBERTpt	0.923	0.917	0.936	0.949	0.933	0.935	0.933

Tabela 2. Resultados dos testes para cada modelo.

Os resultados dos algoritmos clássicos foram razoáveis, mas foram inferiores aos resultados dos modelos profundos. Porém, os modelos clássicos possuem a vantagem de permitir a avaliação direta das *features* mais importantes para a classificação da condição de um paciente de acordo com os laudos médicos. A Figura 5 mostra as 20 palavras mais importantes de acordo com o modelo *LinearSVC*.

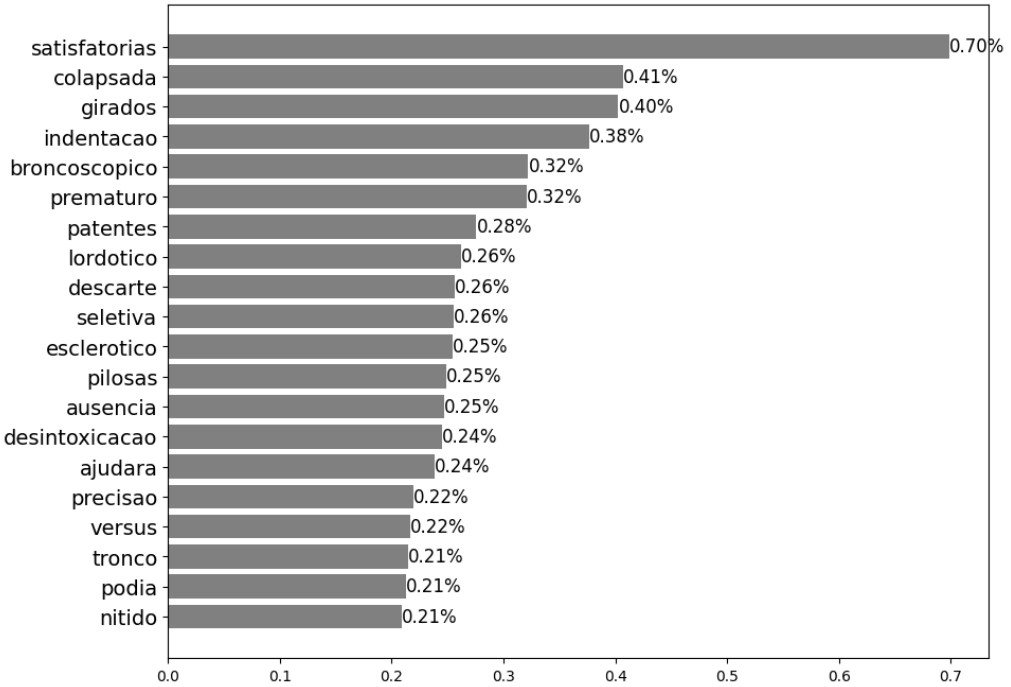

Figura 5. LinearSVC: as 20 *features* mais importantes para a classificação.

Para avaliar como as *features* implicam na classificação, foi calculada a *correlação de Pearson* entre elas e a condição do paciente (ver Figura 6). Como a condição pode as-

sumir os valores 0 e 1, para as classes: normal e alterada, respectivamente, uma correlação positiva indica a relação da *feature* com a alteração da saúde e uma correlação negativa indica a normalidade da saúde. É importante salientar que correlação não implica causalidade, pois ela se limita a associações estatísticas.

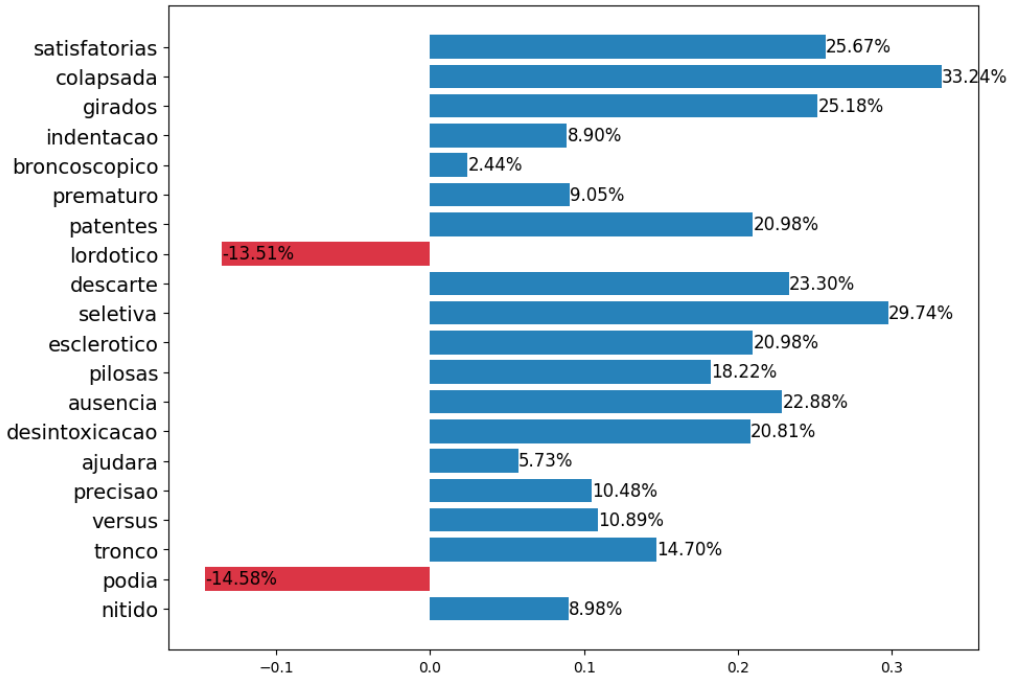

Figura 6. Correlação de Pearson: features vs condição do paciente.

Por outro lado, os demais modelos apresentaram desempenho superior, com destaque especial para a RNN, que registrou uma acurácia de 95%, superando o modelo BERTimbau (acurácia = 94%). Para os modelos profundos, foram geradas as curvas de aprendizado, considerando a acurácia x época. A intenção é avaliar o quanto eles generalizam no decorrer do treinamento. A Figura 7 mostra as curvas geradas para os modelos RNN e CNN e a Figura 8 para os modelos BERTimbau e BioBERTpt.

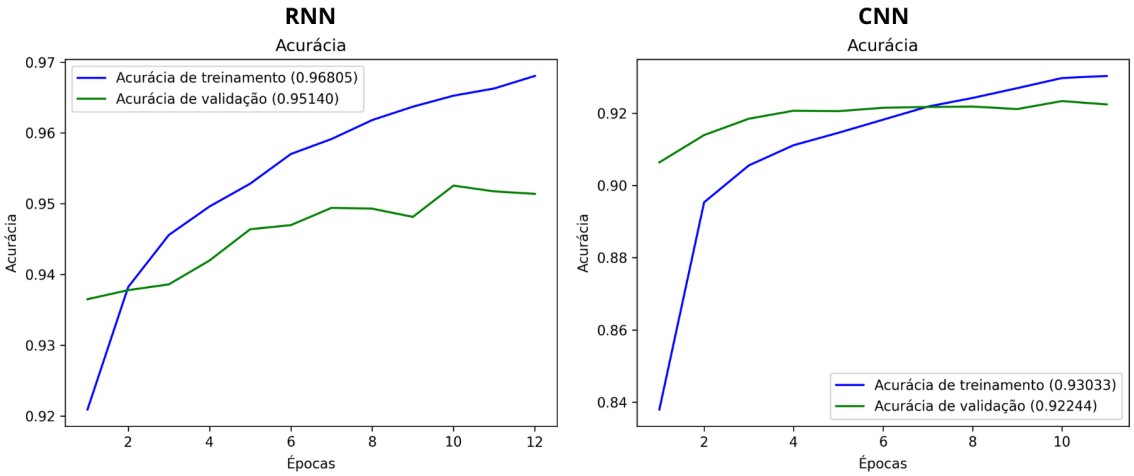

Figura 7. Curvas de aprendizado para RNN e CNN.

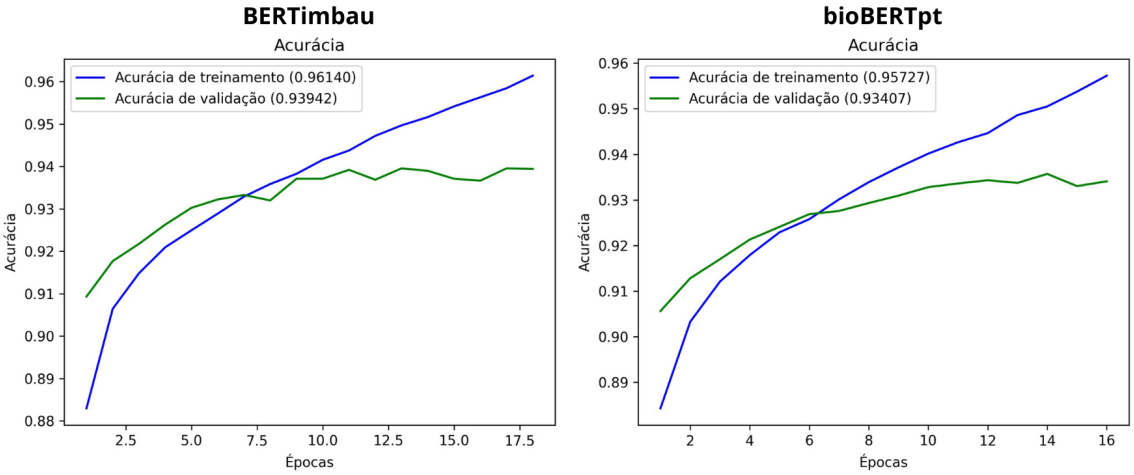

Figura 8. Curva de aprendizado para BERTimbau e bioBERTpt.

Nos modelos profundos foi aplicada a técnica de *Early Stopping*, responsável por parar o treinamento antes que eles se especializassem demais, evitando *overfitting*. Nessa técnica existe um mecanismo de *checkpoint* para salvar os melhores parâmetros dos modelos com base no menor valor de perca (*loss*) para o conjunto de validação.

A partir desses gráficos, pode-se observar alguns padrões comuns a todos eles. Quando a curva de treinamento se encontra abaixo da curva de validação no início do treinamento, isso sugere que o modelo está subajustado. Esse fenômeno indica que o modelo não está conseguindo capturar adequadamente a complexidade dos dados de treinamento. À medida que o treinamento prossegue, a curva de treinamento ultrapassa a de validação, mostrando que os padrões estão sendo aprendidos com sucesso, melhorando a capacidade de generalização.

No entanto, é importante observar a ocorrência de *overfitting* à medida que as duas curvas começam a se distanciar consideravelmente, sugerindo o ajuste excessivo aos dados de treinamento, capturando ruídos e padrões irrelevantes. Além disso, a curva de validação se estabiliza primeiro, indicando que o modelo atingiu seu limite de generalização antes de aprender completamente os padrões nos dados de treinamento. É nesse momento que ocorre a parada antecipada (*Early Stopping*).

5. Conclusões

O estudo realizado evidencia a importância da aplicação de técnicas de PLN e IA na tarefa de classificação de laudos clínicos, oferecendo um meio eficiente para detectar e categorizar condições médicas como normais ou alteradas. Os resultados mostraram que as abordagens de redes neurais (modelo RNN) e os modelos de língua (BERTimbau) apresentaram um desempenho superior em comparação aos demais modelos.

Além disso, a análise das *features* mais importantes identificadas pelo modelo *LinearSVC* mostrou dicas importantes sobre aspectos linguísticos relevantes na classificação entre laudos normais e alterados. Essas *features* podem ser exploradas ainda mais para a aprimoramento de desempenho e fornecimento de uma compreensão mais profunda das informações contidas nos laudos clínicos.

Como limitações observadas, destaca-se a dificuldade em se obter dados médicos

textuais, especialmente em português, pois são dados sigilosos e a burocracia para conseguí-los se torna um obstáculo. Para contornar este problema, o uso da técnica *Back-translation* se mostrou eficiente para a obtenção de dados em português.

Em suma, este estudo investigou o potencial das técnicas de AM para análise de laudos médicos escritos em textos livres. A automação da regulação médica pode oferecer uma solução eficaz para a triagem e classificação dos laudos, contribuindo para uma tomada de decisão mais rápida e precisa no contexto da saúde suplementar, beneficiando as operadoras de planos de saúde em termos de redução de custos e otimização de recursos e os pacientes, na melhoria da qualidade do atendimento.

Agradecimentos

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

Referências

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., and Stanley, H. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23).

Graham, B., Bond, R., Quinn, M., and Mulvenna, M. (2018). Using data mining to predict hospital admissions from the emergency department. *IEEE Access*, 6:10458–10469.

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. (2019a). MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). PhysioNet. `https://doi.org/10.13026/8360-t248`.

Johnson, A., Pollard, T., Mark, R., Berkowitz, S., and Horng, S. (2019b). MIMIC-CXR database (version 2.0.0). PhysioNet. `https://doi.org/10.13026/C2JT1Q`.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019c). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. (2019d). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Magalhães Junior, G. V., Vieira, J. P. A., Santos, R. L. S., Barbosa, J. L. N., Santos Neto, P., and Moura, R. S. (2019). A study of the influence of textual features in learning medical prior authorization. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*.

Roquette, B. P., Nagano, H., Marujo, E. C., and Maiorano, A. C. (2020). Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Networks*, 126:170–177.

Wang, B., Li, W., Bradlow, A., Bazuaye, E., and Chan, A. T. (2023). Improving triaging from primary care into secondary care using heterogeneous data-driven hybrid machine learning. *Decision Support Systems*, 166:113899.