Utilização de modelos BERT em língua portuguesa para predição de códigos CID em contexto neonatal

Ricardo da S. Santos [1], Murilo G. Gazzola[2], Renato T. Souza[3], Rodolfo C. Pacagnella [3], Cristiano Torezzan [4]

[1]Instituto de Matematica, Estatística e Computação Cientifica (IMECC)
Universidade Estadual de Campinas (UNICAMP)

[2]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie (MACKENZIE)

[3]Faculdade de Ciências Médicas (FCM)
Universidade Estadual de Campinas (UNICAMP)

[4]Faculdade de Ciências Aplicadas (FCA)
Universidade Estadual de Campinas (UNICAMP)

ricardo.santos@ime.unicamp.br, gazzola@alumni.usp.br,
rtsouza@g.unicamp.br,rodolfop@unicamp.br, torezzan@unicamp.br

Abstract. *The task of predicting codes from the International Classification of Diseases (ICD) represents a contemporary research challenge in the field of artificial intelligence applied to healthcare. This approach is seen as a promising solution for optimizing recurrent clinical record tasks, increasing diagnostic accuracy, and enhancing medical decision-making. Accurate prediction of ICD codes can streamline and automate administrative processes in healthcare settings and contribute to more personalized and effective medicine. Despite the relevance of this topic, there is still limited research on predicting ICD codes in Brazilian Portuguese. In this context, this work presents preliminary results from an ongoing research project aiming to train algorithms for predicting ICD codes in the context of neonatal primary care, focusing on predicting ICD codes in admissions and discharge reports of newborn pediatric hospitalizations. The algorithms employ models based on BERT - Bidirectional Encoder Representations from Transformers, and preliminary results indicate a promising path, although further adjustments are needed for practical clinical application.*

Resumo. *A tarefa de prever códigos da Classificação Internacional de Doenças (CID) representa um desafio contemporâneo de pesquisa na área de inteligência artificial aplicada à saúde. Essa abordagem é vista como uma solução promissora para otimizar tarefas recorrentes de registros clínicos, aumentar a precisão de diagnósticos e aprimorar a tomada de decisões médicas. Uma previsão acurada de códigos CID pode permitir agilizar e automatizar processos administrativos em ambientes de saúde e contribuir para uma medicina mais personalizada e eficaz. Apesar da relevância desse tema, ainda existem poucas pesquisas para a predição do código CID em português brasileiro. Neste contexto, este trabalho apresenta resultados preliminares de uma pesquisa que está em desenvolvimento, com objetivo de treinar algoritmos para a predição de códigos CID*

no contexto da atenção primária neonatal, com foco na previsão dos códigos CID em admissões e relatórios de alta de internações pediátricas de recém-nascidos. Os algoritmos utilizam modelos baseados em BERT - Representações Codificadoras Bidirecionais de Transformadores e os resultados preliminares indicam que o caminho é promissor, mas ainda há necessidade de ajustes para que se tenha uma aplicação que possa ser utilizada na prática clínica.

1. Introdução

As notas clínicas de pacientes, conhecidas como Registros Eletrônicos de Saúde (EHR, do inglês *Electronic Health Record*), compreendem uma variedade de informações sobre o estado de saúde atual, englobando admissão médica, resultados de exames, histórico e condições pré-existentes.

Quando bem preenchidos, são documentos confiáveis na prática médica, todavia a em geral toda informação contida nesses registros está registrada na forma de textos não estruturados, ou campos abertos com informações numéricas e textuais, armazenados formatos de arquivos variados (como pdf ou doc, txt). Esses formatos dificultam o processamento das informações contidas nos EHR, que raramente são utilizados para fins de gerenciamento clínico, planejamento financeiro ou geração de métricas abrangentes sobre seu conteúdo.

Uma maneira mais objetiva para gerar informações consolidadas para gestores e operadoras de saúde consiste no uso dos códigos da Classificação Internacional de Doenças (CID). A atribuição desses códigos é crucial em diversos aspectos da prática hospitalar e permitem, desde assegurar um faturamento correto até criar um registro válido do histórico de atendimento de pacientes. No entanto, a atribuição manual desses códigos, como destacado em estudos como [Whiteley et al. 2022], muitas vezes é um processo lento e impreciso.

Uma alternativa que tem se mostrado promissora é atribuição automática (ou semi-automática) de códigos da CID. O problema de automatizar a extração de códigos médicos de textos clínicos não estruturados tem sido um objetivo de longa data da pesquisa médica em processamento de linguagem natural [Soroush et al. 2024], por exemplo, na década de 90, [Larkey and Croft 1995] já propunham o uso de técnicas de aprendizado não supervisionado, para tentar determinar o código CID-9 automaticamente ao resumo de alta do paciente. Estudos posteriores exploraram abordagens baseadas em *machine learning* e regras para predizer códigos CID [Pakhomov et al. 2006, Farkas and Szarvas 2008]. No entanto, tais abordagens encontravam barreiras no processamento adequado dos campos textuais. Com a recente evolução dos modelos de Processamento de Linguagem Natural, sobretudo a partir do desenvolvimento dos *Transformers*, esse assunto voltou a chamar atenção da comunidade científica, com resultados promissores como em [Nguyen et al. 2023].

Além do desafio técnico, há particularidades linguísticas associada ao idioma que necessitam atenção. No contexto da língua portuguesa, em 1998, [de Lima et al. 1998] propuseram um modelo hierárquico para prever a categoria do código CID. No entanto, esse modelo não determinava um código CID específico. Em uma abordagem mais recente, [Duarte et al. 2018] utilizou modelos de redes neurais profundas para realizar a codificação de CID pós-morte a partir de certidões de óbitos e relatórios de autópsia. Por

outro lado, [Oleynik et al. 2017] propuseram prever grupos de códigos CID oncológicos a partir de relatórios de patologia. Embora esses artigos apresentem aspectos interessantes, é importante notar que esses estudos foram realizados em português europeu, que possui diversas particularidades linguísticas em relação ao português brasileiro.

Dentro do cenário brasileiro a literatura é bastante escassa, com raros trabalhos como [Reys et al. 2020], onde técnicas de processamento natural, especialmente uma *Convolutional Neural Network* (CNN) com atenção por rótulo, foram empregadas para prever códigos da CID baseados em alguns documentos de Evolução Clinica, Exames Físicos e Sumários de Alta do Hospital Sírio Libanês. No total foram mais de 100 mil documentos analisados e o melhor resultado obtido foi de uma métrica F_1 igual a 0,485, considerando os top-50 códigos CIDs mais frequentes.

Neste trabalho, investigamos o uso de métodos baseados em BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2019] , para predição de códigos CIDs em adendos de internação pediátrica e em relatórios de alta pediátrica de recém nascidos, totalmente redigidos em português brasileiro.

2. Materiais e Método

2.1. Banco de Dados

Neste estudo, realizou-se a análise de 5707 adendos de internação pediátrica e 5703 relatórios de alta, todos anonimizados e provenientes de um Hospital do Estado de São Paulo. Utilizando técnicas de processamento textual e expressões regulares (Regex), foi possível extrair 2164 adendos e 1379 relatórios de alta que se enquadravam nos critérios da pesquisa, ou seja, continham informações relevantes sobre as doenças, especialmente os códigos CIDs associados. É relevante destacar que, dentro deste contexto, as notas clínicas podem conter mais de um rótulo, abrangendo diversos aspectos do quadro clínico e do tratamento dos pacientes. A Figura 1 apresenta um resumo do fluxo de pré-processamento dos registros utilizados.

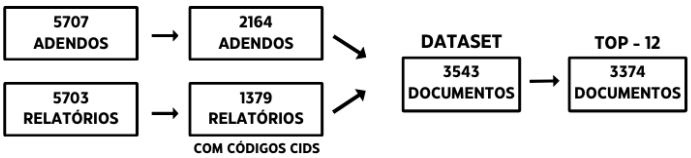

Figura 1. Banco de Dados

Após a etapa inicial de padronização dos códigos CIDs, na qual foram removidos pontuações, espaços em branco e traços para garantir uniformidade na representação dos códigos, diversos experimentos foram conduzidos utilizando modelos pré-treinados em língua portuguesa.

Os modelos foram configurados para realizar uma tarefa de classificação multi-rótulo supervisionada, visando prever os códigos CIDs mais relevantes para cada registro de saúde. Para isso, uma lista dos 12 códigos CIDs mais frequentemente encontrados nos documentos analisados foi selecionada como o conjunto-alvo. Os dados foram divididos em três conjuntos: 70% para treinamento, 15% para validação e 15% para teste. A distribuição dos códigos CID nos conjuntos pode ser observada na Figura 2.

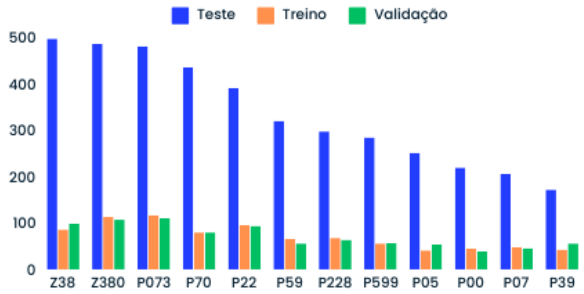

Figura 2. Frequência dos Conjuntos de Treino, Teste e Validação dos 12 principais códigos CIDs

2.2. Métodos

A arquitetura Transformers [Vaswani et al. 2017] trouxe um notável avanço no Processamento de Linguagem Natural, introduz uma abordagem inovadora com mecanismos de atenção multi-cabeça, permitindo interpretações contextuais variadas para o mesmo token. Baseado nisso, o modelo BERT [Devlin et al. 2019] utiliza a atenção bidirecional durante o pré-treinamento da linguagem, o que facilita seu ajuste (*fine-tuning*) para tarefas específicas. Esse processo de *fine-tuning* geralmente necessita apenas de um conjunto pequeno de dados rotulados, pois o BERT já possui uma compreensão abrangente da linguagem após o pré-treinamento.

Nesse contexto, um dos modelos que foram testados para a tarefa de predição multi-rótulo do CID (Classificação Internacional de Doenças) foi o BERTimbau, incluindo sua versão BERTimbau-Large [Souza et al. 2020]. Ambos os modelos do BERTimbau foram treinados em um extenso corpus de texto em português, com o objetivo de aprender representações de palavras e sentenças que capturam as nuances do idioma.

Além disso, também foi testado o modelo biomédico BioBERT-PT [Schneider et al. 2020], que foi treinado utilizando títulos e resumos de artigos científicos em português publicados no PubMed e no Scielo, e por fim testamos também o modelo RoBERTa-XLM [Conneau et al. 2020] que é uma extensão multilíngue do modelo RoBERTa, projetado para compreender e gerar texto em vários idiomas.

Antes dos experimentos, era necessário determinar a quantidade de códigos CID que planejávamos prever, levando em conta a disponibilidade de documentos para este estudo preliminar. Optamos por direcionar nossos esforços para os 12 códigos CID mais comumente encontrados. A figura 3 exemplifica o funcionamento do modelo, delineando os conceitos de predição correta (verdadeiro positivo), predição incorreta (falso positivo) e predição não capturada (falso negativo), no contexto de predição de CID.

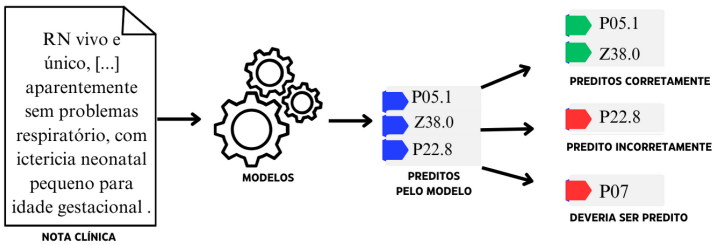

Figura 3. Exemplo do Problema Multi-rótulo aplicado à predição de CID

Para nosso experimento, executamos os modelos BERTimbau-Base, BERTimbau-Large, BioBERTpt e RoBERTa-XLM considerando 5,10 e 15 épocas, todos os testes foram executados em *Python* 3.10.12 utilizando uma GPU-V100.

3. Resultados

Para os testes realizados, considerando os top-12 códigos CID mais frequentes, o modelo BERTimbau-Large obteve o melhor resultado com 15 épocas, alcançando um valor de F1-Score de 0,4955. No entanto, os modelos BioBERTpt e BERTimbau também alcançaram pontuações bastante satisfatórias, com F1-Scores de 0,4686 e 0,4674 respectivamente. A Tabela 1 apresenta as pontuações F1 alcançadas por cada um dos modelos testados:

Modelo	5 épocas	10 épocas	15 épocas
BERTimbau - Base	0,3136	0,4157	0,4674
BERTimbau - Large	0,3582	0,4703	0,4955
BioBERTpt	0,2649	0,4561	0,4686
RoBERTa XLM	01468	03144	0,3745

Tabela 1. F1-Score para os doze códigos CIDs mais frequentes

Quando examinamos o desempenho do modelo líder, o BERTimbau-Large com 15 épocas, além do F1-score, é fundamental considerar outras métricas como precisão, recall e acurácia. Este modelo obteve uma precisão de 0,4620 e um recall de 0,5307, além de uma acurácia geral de 0,8598. Dadas as complexidades inerentes à tarefa de predição automática de CID, especialmente devido ao número de rótulos e uma quantidade limitada de documentos clínicos disponíveis para treinamento, esses resultados apontam para uma promissora eficácia da utilização de modelos BERTs nesse contexto específico.

4. Considerações Finais e Perspectivas Futuras

Com base nos resultados deste estudo no domínio específico neonatal, a tarefa de predição de CID mostra-se promissora, mesmo com um dataset não tão extenso. Esses resultados sugerem que os modelos baseados em BERT têm grande potencial, especialmente quando aplicados a conjuntos de dados maiores. Essa descoberta abre caminho para estudos futuros que explorem o uso desses modelos em contextos mais amplos e até mesmo para o desenvolvimento de ferramentas de apoio à decisão médica.

Olhando para frente, é pertinente considerar pesquisas adicionais que avaliem o desempenho desses modelos em datasets mais abrangentes e diversificados, e envolvendo múltiplos domínios médicos. Além disso, a criação de ferramentas práticas e de fácil utilização, que integrem os modelos baseados em BERT, poderia fornecer suporte valioso aos profissionais de saúde em suas tomadas de decisão, ajudando a agilizar e melhorar o processo de diagnóstico e tratamento em ambientes clínicos movimentados.

Referências

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

de Lima, L. R., Laender, A. H., and Ribeiro-Neto, B. A. (1998). An experimental study in automatically categorizing medical documents. In *Journal of the American Society for Information Science and Technology*, CIKM '98, Nova York, NY, EUA.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Duarte, F., Martins, B., Pinto, C. S., and Silva, M. J. (2018). Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text. *Journal of biomedical informatics*, 80.

Farkas, R. and Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9(3).

Larkey, L. S. and Croft, W. B. (1995). Automatic assignment of icd9 codes to discharge summaries. In *Proceedings of the Annual Meeting of the American Medical Informatics Association*.

Nguyen, T. T., Schlegel, V., Kashyap, A. R., and Winkler, S. (2023). A two-stage decoder for efficient icd coding. In *Annual Meeting of the Association for Computational Linguistics*.

Oleynik, M., Patrão, D. F. C., and Finger, M. (2017). Automated classification of semi-structured pathology reports into icd-o using svm in portuguese. *Studies in health technology and informatics*, 235.

Pakhomov, S. V. S., Buntrock, J. D., and Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*.

Reys, A. D., Silva, D., Severo, D., Pedro, S., de Sousa e Sá, M. M., and Salgado, G. A. C. (2020). Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. In *Intelligent Systems*, Cham. Springer International Publishing.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., and Barra, C. M. C. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*.

Soroush, A., Glicksberg, B. S., Zimlichman, E., Barash, Y., Freeman, R., Charney, A. W., Nadkarni, G. N., and Klang, E. (2024). Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, Cham. Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Whiteley, W. et al. (2022). Automated clinical coding: what, why, and where we are? *npj Digital Medicine*, 5(1):159.