Extração de Sinais Vitais em Textos Clínicos usando LLMS

Sami Sandy Bezerra Nagahama[1] **, Ticiana Linhares Coelho da Silva**[1]

[1]Instituto Universidade Virtual, Universidade Federal do Ceará – Fortaleza – Ceará – Brasil

saminagahama@alu.ufc.br, ticianalc@virtual.ufc.br

Abstract. *This study comparatively analyzes large language models for named entity recognition (NER) in Portuguese clinical texts, focusing on vital sign identification. Using 320 nursing records from Hospital Geral de Fortaleza, manually annotated for five vital sign categories, we evaluated BioBERTpt, Clinical-LLaMA-BR-7b, and NuExtract 1.5. Metrics such as accuracy, precision, recall, and F1-score were used. NuExtract demonstrated superior performance, with accuracy over 0.89 for all entities and consistent identification distribution with the ground truth. BioBERTpt showed limitations with entity overlap, while Clinical-LLaMA-BR-7b exhibited a tendency for hallucination.*

Resumo. *Este trabalho analisa comparativamente modelos de linguagem para reconhecimento de entidades nomeadas (REN) em textos clínicos em português, focando na identificação de sinais vitais. Utilizando 320 registros de enfermagem do Hospital Geral de Fortaleza, anotados com cinco categorias de sinais vitais, avaliamos BioBERTpt, Clinical-LLaMA-BR-7b e NuExtract 1.5. Métricas como acurácia, precisão, revocação e F1-score foram empregadas. O NuExtract demonstrou desempenho superior, com acurácia acima de 0,89 para todas as entidades e distribuição consistente com o ground truth. O BioBERTpt apresentou limitações de sobreposição, e o Clinical-LLaMA-BR-7b, tendência à alucinação.*

1. Introdução

O avanço nas áreas de aprendizado de máquina e inteligência artificial tem impulsionado significativamente o uso do Processamento de Linguagem Natural (PLN) em aplicações práticas, especialmente na extração de informações em textos clínicos. Esses documentos contêm dados ricos e não estruturados, como registros de enfermagem, compostos por terminologia especializada e conteúdos complexos e não padronizados.

O presente estudo insere-se no projeto colaborativo UFC-UECE *"Computação de Alto Desempenho em Nuvens Aplicada à Escalabilidade de um Sistema de Predição de Riscos de Deterioração Clínica de Pacientes via Aprendizagem de Máquina"*, financiado pela FAPESP. Os dados utilizados neste trabalho são oriundos de registros de enfermagem de pacientes internados no Hospital Geral de Fortaleza.

O problema central abordado nesta pesquisa é como extrair eficientemente informações específicas desses documentos clínicos, particularmente os sinais vitais, considerando que são registrados de forma não padronizada nos textos, dificultando sua identificação e extração automática.

Para abordar este problema, propõe-se o uso da técnica de Reconhecimento de Entidades Nomeadas (REN), que identifica e classifica entidades específicas em textos

não estruturados [Yadav and Bethard 2018, Nadeau 2007]. Na área da saúde, o REN é aplicado para extrair conceitos clínicos de grandes volumes de dados não estruturados, viabilizando análises em larga escala, predição de eventos clínicos e desenvolvimento de sistemas de apoio à decisão clínica [Schneider et al. 2020, Demner-Fushman et al. 2009].

Este estudo avalia comparativamente três modelos de linguagem de grande porte (LLMs) baseados na arquitetura Transformer para a tarefa de REN: BioBERTpt, Clinical-LLaMA-BR-7b e NuExtract 1.5. A avaliação desses modelos é realizada sobre o conjunto de dados de registros de enfermagem, previamente anotados manualmente para as cinco categorias de sinais vitais mencionadas.

A importância dos sinais vitais no contexto clínico justifica o foco deste trabalho, pois segundo [Churpek et al. 2016], eles são os preditores mais precisos de deterioração clínica, sendo fundamentais para detecção precoce de quadros críticos, decisões sobre transferências para UTI e desenvolvimento de sistemas de alerta precoce eficientes. Além disso, a Resolução COFEN nº 736/2024 [COFEN 2024] determina que nas Anotações de Enfermagem devem constar os sinais vitais observados, sendo que ”os sinais vitais mensurados devem ser registrados pontualmente, ou seja, os valores exatos aferidos”[COFEN 2015, p. 16], o que reforça a relevância da extração automática dessas informações para a prática clínica.

2. Trabalhos relacionados

Diversos trabalhos recentes exploram o Processamento de Linguagem Natural (PLN) para extração de informações clínicas. [Song et al. 2021] apresentam uma revisão abrangente de métodos de deep learning para REN biomédico. Em português, [Torres et al. 2024] desenvolveram um modelo baseado na biblioteca spaCy para extrair entidades de casos clínicos. [Akhtyamova 2020] utilizou um modelo BERT treinado especificamente para textos biomédicos em espanhol. [Wang et al. 2023] conduziram revisão sistemática sobre modelos de linguagem pré-treinados para o domínio biomédico, evidenciando que o pré-treinamento específico com textos biomédicos supera significativamente abordagens com textos de domínios mistos.

O presente estudo avança em relação a estes ao focar na avaliação comparativa de modelos Transformer para o português do Brasil e especificamente na extração de um conjunto definido de sinais vitais.

3. Dados e métodos

Nesta seção, são abordados o conjunto de dados e os modelos utilizados para a extração de sinais vitais nos textos clínicos.

3.1. Conjunto de Dados e Anotações

Os textos não estruturados foram anotados manualmente com o software Label Studio para cinco rótulos de sinais vitais: pressão arterial, frequência cardíaca, frequência respiratória, temperatura e saturação de oxigênio.

O Ground Truth compreende 320 registros selecionados aleatoriamente com distribuição relativamente equilibrada das entidades (PA: 162, FC: 166, FR: 159, Temperatura: 174, SpO2: 127), com uma média de 2,46 entidades por texto, sendo que 36,88% dos textos não apresentaram entidades identificáveis.

A análise revelou que sinais vitais tendem a aparecer em conjunto nos registros, o que é consistente com a prática clínica de avaliação simultânea de múltiplos indicadores, conforme demonstra a distribuição exposta na Figura 1.

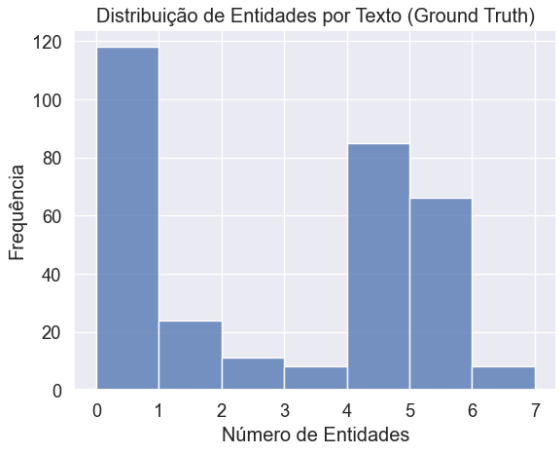

Figura 1. Distribuição de entidades por texto

3.2. Modelos de Linguagem utilizados.

Os Modelos de linguagem baseados na arquitetura Transformer possuem bilhões de parâmetros e são pré-treinados com grandes volumes de dados. O elemento fundamental dos Transformers é o mecanismo de auto-atenção que permite ao modelo focar seletivamente nas partes mais relevantes da entrada para gerar a saída, capturando dependências em longas distâncias de forma eficiente [Vaswani et al. 2017]. A arquitetura possui codificador (processa entrada) e decodificador (gera saída), sendo que alguns modelos, como o BERT, usam apenas o codificador; o GPT, apenas o decodificador; e o T5, ambos.

Para domínios específicos, como o de saúde, os modelos requerem adaptação através de novos conjuntos de dados anotados ou few-shot prompting (aprendizado em contexto). Exemplos de LLMs aplicados ao REN em documentos clínicos incluem BioGPT [Luo et al. 2022], BioBERTpt [Schneider et al. 2020] e GPT2-BioPT [Schneider et al. 2021], sendo os dois últimos desenvolvidos por brasileiros e adaptados para português e domínio médico, fundamentando a aplicação em extração de informações de registros clínicos.

3.3. Critérios de avaliação.

A avaliação dos modelos utiliza matriz de confusão 6x6 (cinco sinais vitais + categoria ”outros”) com análise de Verdadeiro Positivo (TP), Verdadeiro Negativo (TN), Falso Positivo (FP) e Falso Negativo (FN). As métricas aplicadas são: revocação/sensibilidade ($S = \frac{VP}{VP+FN}$), que mede a proporção de casos positivos corretamente identificados; acurácia ($A = \frac{VP+VN}{VP+VN+FP+FN}$), que indica casos corretamente identificados no total da amostra; precisão (Valor Preditivo Positivo); e F1-Score ($F1 = 2/(\frac{1}{S} + \frac{1}{VPP})$), média harmônica entre sensibilidade e precisão [Izbicki and Santos 2020].

4. Resultados

Esta seção apresenta os resultados alcançados ao utilizar os modelos BioBERTpt, Clinical-LLaMA-BR-7b, e NuExtract 1.5 no conjunto de dados deste trabalho.

4.1. BioBERTpt

Este modelo é baseado em BERT multilingual com ajuste fino em textos clínicos brasileiros (2,1M notas) e biomédicos (16,4M palavras), contemplando 13 rótulos de entidades médicas através do esquema BIO, servindo como baseline por sua consolidação acadêmica e processamento rápido sem GPU.

No contexto do presente estudo, que tem como objetivo específico a identificação de sinais vitais, dentre as treze categorias de entidades definidas pelo BioBERTpt [Schneider et al. 2020], a análise concentrou-se especificamente na entidade "Sinal ou Sintoma", visto que este modelo não distingue entre os cinco sinais objetos da presente pesquisa. Entretanto, o modelo teve um desempenho inferior ao esperado, manifestando inconsistências na classificação, com sobreposição de múltiplas entidades em um mesmo token, impossibilitando uma avaliação sistemática dos resultados obtidos.

4.2. Clinical-LLaMA-BR-7B

Este modelo (6,74B parâmetros) deriva do LLaMA-2 [Touvron et al. 2023] com ajuste fino em 2,4GB de dados clínicos brasileiros (SemClinBr, BRATECA e textos neurológicos), requerendo prompt engineering com one-shot para REN devido à sua natureza generativa [Clinical-BR-LlaMA-2-7B 2024]. Foi utilizado com processamento em GPU T4 em 2,5h de execução.

Apesar das configurações do prompt visando maior determinismo e geração de sequência única, observou-se alta ocorrência de alucinações, reproduzindo frequentemente o exemplo do prompt.

Os resultados mostaram que o Clinical-LLaMA-BR-7b apresentou uma concentração desproporcional de identificações para "Temperatura"(320 ocorrências) em contraste com outros sinais vitais, indicando inconsistências em relação ao Ground Truth, que tinha 36,88% dos textos sem anotações.

A análise de desempenho, incluindo a matriz de confusão e métricas, mostrou resultados insatisfatórios, especialmente para Pressão Arterial e Saturação de Oxigênio, com valores nulos de precisão, revocação e F1-score. A alta acurácia para a maioria dos sinais vitais foi devido à correta não identificação de entidades ausentes, ressaltando a necessidade de múltiplas métricas. Já o rótulo "Temperatura"exibiu muitos falsos positivos e baixa acurácia, atribuído à alucinação baseada no exemplo do prompt, mesmo após testes com múltiplas variações.

4.3. NuExtract-1.5

Este modelo (3,82B parâmetros) é o resultado do ajuste fino do Phi-3.5-mini-instruct, focado na extração de informações estruturadas multilíngue baseada em templates JSON [Cripwell et al. 2024]. Além disso, ele não é especializado em domínios específicos.

Foi este modelo que demonstrou o melhor desempenho na análise comparativa de Reconhecimento de Entidades Nomeadas (REN). A distribuição das entidades identificadas pelo NuExtract é consistente com a verificada no Ground Truth, confirmando a eficácia do LLM na identificação de sinais vitais, com uma média de 2,42 entidades por texto.

A matriz de confusão (Figura 2) e as métricas (Figura 3) confirmaram o bom desempenho do modelo, destacando a Frequência Cardíaca (146 acertos), Frequência Respiratória (126) e Pressão Arterial (120). Saturação de Oxigênio (118) e Temperatura (93) também apresentaram bons resultados.

As métricas de acurácia foram superiores a 0,89 para todas as entidades clínicas principais, com a Saturação de Oxigênio atingindo 0,96, e precisão, revocação e F1-score elevados para esta última entidade, Frequência Cardíaca e Frequência Respiratória. Pressão arterial apresentou 58 ocorrências de Falsos Positivos enquanto Temperatura teve 81 Falsos Negativos.

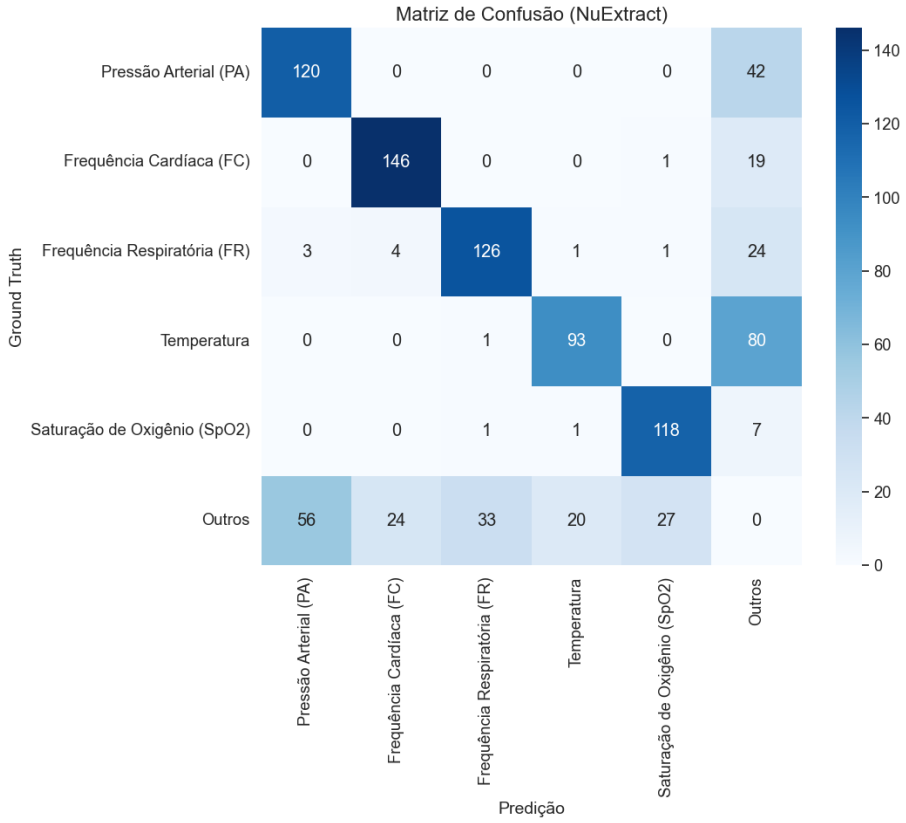

Figura 2. Matriz de confusão: *Ground Truth* versus predições do NuExtract

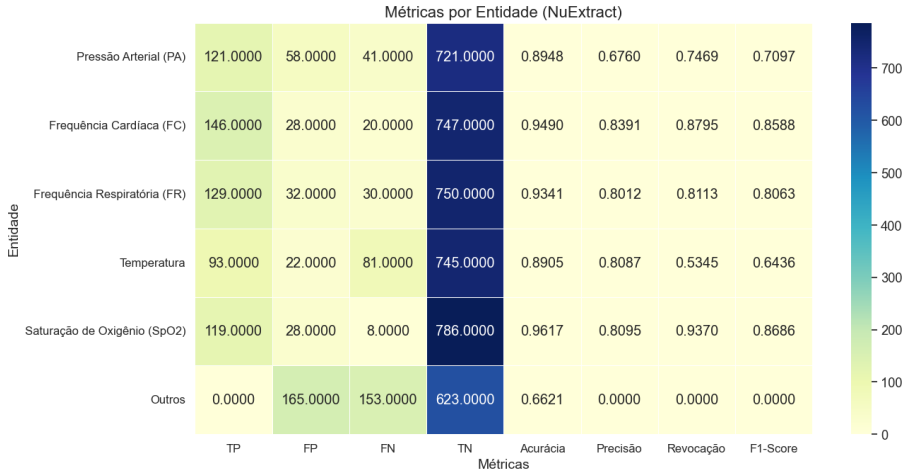

Figura 3. Métricas por entidade das predições do NuExtract

5. Conclusão

Esta pesquisa teve como objetivo desenvolver e avaliar métodos de Reconhecimento de Entidades Nomeadas (REN) em textos clínicos em português, focando na identificação de sinais vitais por serem os preditores mais precisos de deterioração clínica. Comparou-se três abordagens: BioBERTpt, Clinical-LLaMA-BR-7b e NuExtract.

O NuExtract demonstrou desempenho superior na identificação de sinais vitais, com precisão e revocação consistentes, sugerindo grande potencial de melhoria com fine-tuning e expansão de dados.

O BioBERTpt apresentou limitações significativas na tarefa de sinais vitais, enquanto o Clinical-LLaMA-BR-7b evidenciou tendência à alucinação e reprodução excessiva de exemplos do prompt.

As principais contribuições incluem a avaliação comparativa das abordagens de REN, a anotação de 320 documentos de evolução de pacientes para expansão futura e o desenvolvimento de uma solução inicial para extração automática de informações clínicas.

Este estudo ressalta a importância de soluções especializadas para processamento de textos clínicos em português, sendo um passo fundamental para sistemas de monitoramento e predição de riscos em pacientes hospitalares no Brasil.

Referências

Akhtyamova, L. (2020). Named entity recognition in spanish biomedical literature: Short review and bert model. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 1–7. IEEE.

Churpek, M. M., Adhikari, R., and Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, 102:1–5.

Clinical-BR-LlaMA-2-7B (2024). Acesso em 12 jan. 2025.

COFEN (2015). *Guia de Recomendações para Registro de Enfermagem no Prontuário do Paciente e outros Documentos de Enfermagem*. Conselho Federal de Enfermagem, Brasília. Versão Web.

COFEN (2024). Resolução cofen nº 736 de 17 de janeiro de 2024. Dispõe sobre a implementação do Processo de Enfermagem em todo contexto socioambiental onde ocorre o cuidado de enfermagem.

Cripwell, L., Constantin, A., and Bernard, E. (2024). Nuextract 1.5 - multilingual, infinite context, still small, and better than gpt-4o!

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *J Biomed Inform*, 42(5):760–772.

Izbicki, R. and Santos, T. M. d. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, São Carlos, SP.

Luo, R. et al. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Tese de doutorado, University of Ottawa.

Schneider, E. T. R. et al. (2020). Biobertpt: a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*.

Schneider, E. T. R. et al. (2021). A gpt-2 language model for biomedical texts in portuguese. In *2021 IEEE 34th international symposium on computer-based medical systems (CBMS)*, pages 474–479. IEEE.

Song, B., Li, F., Liu, Y., and Zeng, X. (2021). Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):1–18.

Torres, A. M. N., Bersot, R. P. M., and Colombo, C. d. S. (2024). A extração de entidades nomeadas em relatos de casos clínicos. In *Anais do XX Congresso Brasileiro de Informática em Saúde*, Belo Horizonte, MG. Sociedade Brasileira de Informática em Saúde.

Touvron, H., Martin, L., Stone, K., Albert, P., et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Vaswani, A. et al. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., and Fu, J. (2023). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.