CBIS'24 XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

# Avaliação de grandes modelos de linguagem na extração de informações clínica

# Evaluating of large language models in extracting clinical information

# Evaluación de modelos de lenguaje en la extracción de información clínica

Carlos Eduardo Rodrigues Mello[1], Elisa Terumi Rubel Schneider[2], Lucas Emanuel Silva e Oliveira[3], Juliana Nabbouh do Nascimento[4], Yohan Bonescki Gumie[5], Isabela Fontes de Araújo[6], Claudia Moro[7]

1 Graduando em Ciência da Computação, Pontifica Universidade Católica do Paraná (PUCPR), Curitiba, PR, Brasil
2 Doutora em Informática, Pesquisadora, Instituto do Coração (HC FMUSP), São Paulo - SP, Brasil
3 Doutor em Tecnologia em Saúde, Comsentimento, Curitiba, PR, Brasil
4 Graduanda de Engenharia Biomédica - PUCPR, Curitiba, PR, Brasil
5 Doutor em Tecnologia em Saúde, Pesquisador Instituto do Coração (HC FMUSP), São Paulo - SP, Brasil
6 Mestranda PPGTS/PUCPR, Curitiba, PR, Brasil
7 Doutora Engenharia Elétrica, Professora Titular - PPGTS/PUCPR, Curitiba, PR, Brasil

Autor correspondente: Carlos Eduardo Rodrigues Mello
*E-mail*: carlosrmelloo@gmail.com
*Links: https://github.com/HAILab-PUCPR/LLM_Evaluation_Clinical_Texts*

**Resumo**

**Objetivo:** investigar a eficácia dos modelos de linguagem de grande escala (LLM) no reconhecimento de entidades nomeadas (NER) em notas clínicas em português. **Método**: Foi analisado o desempenho dos modelos de linguagem GPT-3.5, Gemini, Llama-3 e Sabiá-2, na realização de NER em 30 notas clínicas para identificação das entidades "Sinais ou Sintomas", "Doenças ou Síndromes" e "Dados Negados". A tarefa de NER foi avaliada pelos resultados da precisão, recall e *F-score* em cada um destes LLMs. **Resultados:** O modelo Llama-3 apresentou desempenho superior, especialmente em sensibilidade, alcançando um *F-score* de 0,538. O GPT-3.5 demonstrou desempenho equilibrado, enquanto o Gemini mostrou maior precisão, mas menor sensibilidade. **Conclusão:** Os resultados indicam que a escolha do modelo

depende da ponderação adequada desses fatores em relação aos requisitos individuais de cada aplicação clínica.

**Descritores:** Processamento de Linguagem Natural; Sinais e Sintomas; Síndrome

**Abstract**

**Objective:** investigate the effectiveness of large language models (LLMs) in named entity recognition (NER) in clinical notes in Brazilian Portuguese. **Method:** We evaluated the NER task in 30 clinical notes using the metrics and methods of precision, recall, and F-score. In the experiment conducted, we compared the performance of the models GPT-3.5, Gemini, Llama-3, and Sabiá-2 in extracting the entities "Signs or Symptoms," "Diseases or Syndromes," and "Negated Data." **Results:** We found that the Llama-3 model showed superior performance, especially in sensitivity, achieving an F-score of 0.538. GPT-3.5 demonstrated balanced performance, while Gemini showed higher precision but lower sensitivity. **Conclusion:** Our results indicate that the choice of model depends on the appropriate weighting of these factors concerning the individual requirements of each clinical application.

**Keywords:** Natural Language Processing; Machine Learning; Signs and Symptoms; Syndrome

**Resumen**

**Objetivo:** investigar la efectividad de los grandes modelos de lenguaje (LLMs) en el reconocimiento de entidades nombradas (NER) en notas clínicas en portugués brasileño. **Método:** Evaluamos la tarea de NER en 30 notas clínicas utilizando las métricas y métodos de precisión, recall y F-score. En el experimento realizado, comparamos el rendimiento de los modelos GPT-3.5, Gemini, Llama-3 y Sabiá-2 en la extracción de las entidades "Signos o Síntomas", "Enfermedades o Síndromes" y "Datos Negados". **Resultados:** Encontramos que el modelo Llama-3 mostró un rendimiento superior, especialmente en sensibilidad, logrando un F-score de 0.538. GPT-3.5 demostró un rendimiento equilibrado, mientras que Gemini mostró mayor precisión, pero menor sensibilidad. **Conclusión:** Nuestros resultados indican que la elección del modelo depende de la ponderación adecuada de estos factores con respecto a los requisitos individuales de cada aplicación clínica.

**Descriptores:** Procesamiento de Lenguaje Natural; Aprendizaje Automático; Signos y Síntomas; Síndrome;

## Introdução

Geralmente quando um paciente interage no sistema de saúde, como nas consultas, avaliações periódicas, internações hospitalares, realização de exames, entre outros; é criada uma documentação clínica que inclui dados tanto estruturados quanto não estruturados, em forma de texto, ou narrativa clínica. Nesse contexto, existem bilhões de registros médicos eletrônicos disponíveis [1]. Estes registros são essenciais para a pesquisa médica orientada por dados e para o desenvolvimento de práticas de saúde [1]. Por exemplo, os textos clínicos possuem informações importantes, como os métodos de tratamento de doenças, avaliações como nos laudos de exames, e prescrições de medicamentos, sendo cada vez mais valorizados para estudos clínicos e aplicações médicas, pois favorece a extração de informações importantes para tratamentos individuais e para população [2]. Na verdade, uma quantidade significativa das informações contidas nos registros eletrônicos de saúde é não estruturada. A identificação de informações contidas em textos, não é uma tarefa trivial, restringindo assim sua utilização adicional para aprimorar a pesquisa clínica e criar ferramentas de suporte ao cuidado dos pacientes [3]. A preferência por texto livre se baseia na facilidade que ele proporciona na comunicação entre a equipe de saúde, permitindo que os profissionais de saúde ofereçam informações mais detalhadas, já que não estão limitados a campos estruturados. O texto também pode ser originado pela transcrição de áudios, o que facilita muito o registro de informações relacionadas aos pacientes.

A utilização de técnicas e métodos de Processamento de Linguagem Natural (PLN) é uma das principais forma de identificação de informações de textos, e vem sendo cada vez mais aplicada à área de saúde, em especial na identificação de informações em narrativas clínicas.

O Reconhecimento de Entidades Nomeadas (NER, do inglês *Named Entity Recognition*) é crucial para várias tarefas PLN, como Extração de Informações (IE), Respostas a Perguntas (QA), e contribui com a tarefa de extração de relação entre termos que pode ser utilizada na estruturação de textos livres [3]. Seu propósito é identificar e categorizar informações que são necessárias ao estudo, como por

exemplo nomes de pessoas (PER), locais (LOC), organizações (ORG) e expressões numéricas, incluindo datas, moedas e porcentagens [4], dentro da área da saúde, as categorias podem ser personalizadas e incluir opções como medicamentos, doenças, exames, resultados e órgão [5]. Diante disso, é importante notar que as entidades clínicas dependem das aplicações específicas; por exemplo, em NER clínico, as entidades como tratamentos, sinais e sintomas são cruciais para o contexto médico.

Inicialmente, as técnicas de NER eram baseadas em regras, onde os padrões específicos eram manualmente definidos para identificar entidades nos textos. Com o avanço da tecnologia, o campo evoluiu para o uso de aprendizado de máquina tradicional, que depende de características manualmente extraídas para treinar os modelos, isto é, de anotação (*label)* nas entidades que se deseja identificar/extrair do texto. Porém, o processo de anotação de um corpus de textos clínicos é uma tarefa que demanda muito esforço manual e tempo. Recentemente, as arquiteturas de modelos de linguagem revolucionaram o NER ao oferecerem abordagens ainda mais sofisticadas e eficazes, especialmente com a introdução de *Large Language Models* (LLMs).

Esses modelos de linguagem de grande escala na área de processamento de linguagem natural têm recebido significativa atenção do público geral, particularmente com o surgimento do ChatGPT [6]. O desenvolvimento de LLMs avançados, como o GPT-3 e o GPT-4, que são pré-treinados em vastos conjuntos de dados, facilitou o aprendizado de *zero-shot* e *few-shot in-context*, técnicas de engenharia de *prompt*, ampliando suas aplicações em cenários do mundo real [6].

Existem diferentes LLMs, que possibilitam a utilização de forma gratuita (GPT-3) ou com mais funcionalidades em versões pagas (GPT-4), e alguns *opensource co*mo o Llma. Estes últimos possibilitam o processamento dos dados localmente, o que é importante considerando a necessidade de manter a privacidade das informações das narrativas clínicas. Por outro lado, necessitam de poder computacional, que dependem de recursos financeiros,

Neste contexto, apresenta-se a oportunidade de explorar os LLMs para a tarefa de NER clínica. Existem alguns estudos que utilizam estes LLMs, mas raros em narrativas clínicas escritas em português. Além disto, a avaliação dos resultados da aplicação dos LLMs, geralmente não é descrita. A avaliação dos resultados de tarefa

de NER em textos clínicos, é fundamental para analisar se o modelo proposto tem realmente potencial para aplicação na área clínica. Porém, avaliação deve ser realizada comparado os resultados do LLM com o mesmo conjunto de narrativas que já tenham a marcação/anotação das entidades clínicas que serão extraídas. Se avaliação for realizada analisando o resultado do experimento, este terá viés. Neste estudo a proposta é a avaliação ser realizada utilizando narrativas de um corpus clínico, a que as entidades clínicas nele anotadas sejam comparadas aos resultados da tarefa de NER obtida com a aplicação do LLM.

Uma das principais contribuições deste estudo é a utilização de notas clínicas em português brasileiro, uma área que ainda é pouco explorada, como demonstram os trabalhos de Schneider *et al.* [7] e Schneider *et al.* [8]. Até onde sabemos, não há estudos que avaliem o impacto dos LLMs na extração de entidades nomeadas clínicas em português brasileiro. Desta forma, o objetivo deste trabalho é descrever a abordagem de avaliação e os resultados da realização da tarefa de NER para identificação achados clínicos (sinais, sintomas e diagnósticos) em narrativas clínicas.

Para realizar a avaliação foi utilizado o corpus SemClinBR [9], que contém anotação de achados clínicos, como o *baseline para a comprarão com os resultados dos LLMs.* A utilização de um corpus anotado evita o viés na avaliação após a obtenção dos resultados, e permite analisar um número maior de narrativas clínicas.

## Método

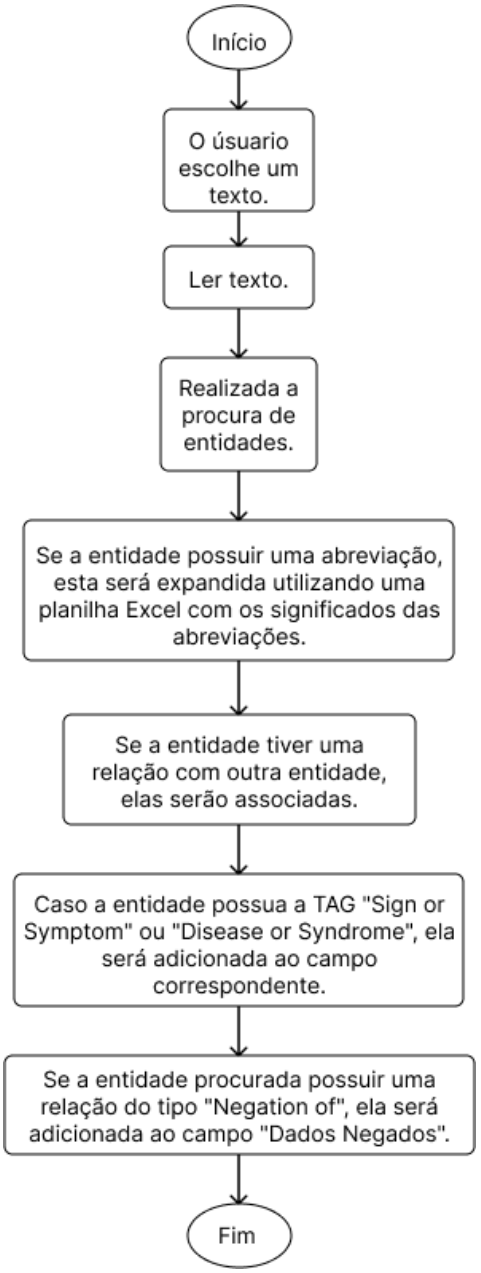
Foi utilizado o SemClinBR [8], um *corpus* semanticamente anotado em português formado por 1.000 notas clínicas contendo 65.117 entidades clínicas anotadas e 11.263 relações entre elas. Dentro das 1.000 narrativas, 30 foram escolhidas para realizar os experimentos e 6 foram escolhidas para serem utilizados de exemplos *few-shot* no *prompt* para o modelo processar. O processo realizado pode ser replicado para a todas as narrativas do SemClinBR.

As etapas para a extração das entidades contidas na narrativa clínica são realizadas conforme mostrado no fluxograma da Figura 1. É importante notar que alguns achados clínicos podem estar associados a uma negação, indicando que síndromes, sinais ou sintomas não estão presentes na narrativa ou foram negados pelo paciente. Para explicar de forma simplificada o processo, o código lê a narrativa e

identifica as entidades marcadas. Se uma entidade tem uma abreviação, ela é expandida conforme o Excel de abreviações do SemClinBR. Entidades relacionadas são combinadas e associadas. Depois, verifica-se se a entidade está marcada como "Sign or Symptom" ou "Disease or Syndrome". Se sim, a entidade é salva no campo correspondente, caso contrário, é ignorada. Se a entidade tem a relação "negation of", é salva em "Dados Negados".

Portanto, o processo de extração deve ser capaz de identificar essas entidades para evitar conclusões errôneas. Isso é crucial para garantir a precisão dos dados extraídos e fornece uma base sólida para a análise dos dados. Além disso, a correta identificação das entidades melhora a qualidade das análises subsequentes, assegurando que as informações refletidas na narrativa sejam interpretadas de forma fiel e confiável.

**Figura 1:** Etapas do pré-processamento de texto realizado




Após o pré-processamento inicial, uma etapa manual foi realizada para a avaliação das entidades. Apesar de as notas do SemClinBr conterem *tag's* para a identificação das entidades, uma etapa de pós-processamento se mostrou necessária, para ajustar eventual segmentação incorreta de dados, a repetição de informações ou a ausência de marcações para certos dados. Por exemplo, na primeira narrativa selecionada, que descreve o sintoma "dor retroesternal, em aperto, intensa sem irradiação", no corpus essa frase está segmentada em três entidades distintas de "Sinal ou Sintoma": "dor retroesternal", "em aperto" e "intensa sem irradiação". Embora a marcação individual desses dados esteja correta, estamos buscando compreender o

conjunto completo. Se apenas a entidade "intensa sem irradiação" fosse destacada sem o contexto completo, seria difícil interpretar adequadamente o sintoma. Outro exemplo ocorre com o sintoma "abdômen flácido", em que apenas "flácido" é identificado, deixando de fora o restante do contexto do sintoma. O resultado desse processo é o *Golden standard*, sendo a retirada de todas as entidades de interesse de acordo com que as *tag's (etiquetas de anotação)* dizem ser, sendo esses os dados que o modelo deve marcar corretamente.

Com os textos processados, a próxima etapa foi fazer as requisições para os modelos de linguagem. Os modelos utilizados foram o GPT-3.5 [9], desenvolvido pela OpenAI, o Gemini [10], desenvolvido pela Google, o Llama-3 70B [11], desenvolvido pela Meta, e o Sabiá-2 Medium [12], desenvolvido pela equipe Maritaca AI.

O *prompt,* utilizado para comunicar-se com um modelo de linguagem, contém instruções para solicitar a realização de uma tarefa específica e é a principal forma de comunicação com o modelo. No nosso caso, a tarefa é a extração de "Sinais ou Sintomas", "Doenças ou Síndromes" e "Dados Negados" presentes no texto. Após múltiplas tentativas de construção do *prompt*, incluindo abordagens de *zero-shot* e alguns modelos de *few-shot* [13], foi desenvolvido um *prompt* que atendeu o objetivo do estudo. Ele inclui 6 narrativas do SemClinBR, que servem como exemplos para o *few-shot learning*, onde o modelo aprende como a tarefa deve ser realizada e quais instruções devem ser seguidas através de exemplos descritos no *prompt* [13]. Esses 6 textos também passaram pelo processamento de texto mencionado anteriormente, incluindo a remoção das *stop-words* presentes em cada um utilizando a biblioteca NLTK [14], sendo as *stop-words* palavras que não agregam muita informação para o texto e não ajudam no processo de seleção de informações relevantes. Os 30 textos foram enviados junto com o *prompt* para que o modelo possa realizar a tarefa de extração e devolver como saída os "Sinais ou Sintomas", "Doenças ou Síndromes" e "Dados Negados" que foram identificados.

Após o envio de cada requisição ao modelo e a entrega da saída, realizamos uma análise do desempenho de cada modelo. A avaliação conduzida foi distribuída em sete categorias: "Acertos", "Acertos Parciais", "Achados Faltantes", "Achados que não estavam na narrativa", "Acertos de dados negados", "Achados Negados Faltantes" e

"Achados Negados que não estavam na narrativa". De maneira simplificada, avaliamos os acertos e erros do modelo na extração dos dados.

Quando o modelo classifica corretamente e extrai a informação esperada, isso é registrado como um Acerto. Se o modelo extrai a informação correta, mas erra na classificação, isso é considerado um Acerto Parcial, também considerado um meio acerto. Os Achados Faltantes são as informações corretas que o modelo não mencionou, enquanto os Achados que não estavam na narrativa referem-se às informações mencionadas pelo modelo que não eram esperadas, indicando erros ou alucinações. A análise também se aplica aos dados negados: se o modelo classifica corretamente dados negados, isso é registrado como um Acerto de dados negados; informações negadas que o modelo não mencionou são consideradas Achados Negados Faltantes; e informações negadas mencionadas erroneamente pelo modelo são os Achados Negados que não estavam na narrativa.

Após a análise de cada texto, utilizamos também as métricas de desempenho precisão, sensibilidade e *F-score*. A precisão corresponde à proporção de previsões positivas que estão realmente corretas, enquanto a sensibilidade é a proporção de casos positivos identificados pelo modelo. O *F-score* é a média harmônica entre precisão e sensibilidade, fornecendo um único valor que equilibra as métricas, essa maneira de avaliação é baseada na maneira de avaliação de modelos chamado de Matriz de Confusão.

A precisão é dada pela equação (1):

$$P = \frac{VP}{VP + VP} \qquad (1)$$

onde VP equivale aos verdadeiros positivos, ou seja, as previsões corretas do modelo sobre o que estava no texto (a soma dos Acertos Completos, Acertos de dados Negados e Acertos Parciais, sendo os Acertos Parciais multiplicados por 0.5 para serem considerados meio acertos) e FP equivale aos falsos positivos, ou seja, os dados que o modelo errou ou alucinou (a soma dos Achados que não estavam na narrativa e dos Achados Negados que não estavam na narrativa).

A sensibilidade é dada pela equação (2):

$$S = \frac{VP}{VP + FN} \quad (2)$$

onde VP representa os verdadeiros positivos e FN representa os falsos negativos, que são os dados presentes no texto, mas não mencionados pelo modelo (a soma dos Achados Faltantes e dos Achados Negados Faltantes).

O *F-score* é dado pela equação (3):

$$F = 2 \cdot \frac{P \cdot S}{P + S} \quad (3)$$

onde P é a precisão e S a sensibilidade.

Por fim, foi calculada a média e determinamos as métricas agregadas de cada atributo para cada um dos modelos utilizados.

## Resultados e Discussão

Os resultados da média de precisão, sensibilidade e *F-score* de cada modelo são apresentados na Tabela 1.

**Tabela 1 –** A média da precisão, sensibilidade e *F-score* de cada modelo.

| Modelo | Precisão | Sensibilidade | *F-Score* |
|---|---|---|---|
| **GPT 3.5 [9]** | 0.470 | 0.530 | 0.477 |
| **Gemini-Pro [10]** | 0.496 | 0.468 | 0.466 |
| **Llama-3 70B [11]** | 0.503 | 0.624 | 0.538 |
| **Sabiá 2 Medium [12]** | 0.398 | 0.544 | 0.426 |

Os resultados das métricas agregadas de cada modelo estão apresentados na Tabela 2.

**Tabela 2 –** As métricas agregadas de cada atributo para cada modelo.

| Modelo | Precisão | Sensibilidade | *F-Score* |
|---|---|---|---|
| **GPT 3.5 [9]** | 0.442 | 0.517 | 0.476 |
| **Gemini-Pro [10]** | 0.523 | 0.441 | 0.478 |
| **Llama-3 70B [11]** | 0.492 | 0.610 | 0.545 |

| | | | |
|---|---|---|---|
| **Sabiá 2 Medium [12]** | 0.365 | 0.491 | 0.419 |

Essas medidas são úteis para avaliar o desempenho de todos os modelos na extração de dados e compreender a abordagem de cada um nesse processo.

O modelo Llama 3 teve melhor desempenho todas as métricas, especialmente na sensibilidade, com um valor de 0.624. Isso indica que o Llama-3 conseguiu identificar os verdadeiros positivos com mais assertividade, ou seja, foi eficaz em detectar e extrair a maior quantidade de dados relevantes no texto fornecido. A precisão de 0.504 sugere que ele também foi mais eficiente em minimizar falsos positivos em comparação com outros modelos, sendo o modelo mais confiável para a tarefa avaliada em relação aos demais. O *F-score* de 0.538 foi o mais alto entre os modelos, indicando que ele mantém um melhor equilíbrio entre precisão e sensibilidade, garantindo um desempenho consistente.

O modelo GPT 3.5 apresentou um desempenho equilibrado, com valores medianos em todas as métricas. Sua precisão de 0.470 e sensibilidade de 0.530 indicam que ele é razoavelmente confiável em suas predições, mas não tão eficaz quanto o Llama 3. O *F-score* de 0.477 reflete esse equilíbrio, mostrando que o GPT 3.5 é uma escolha sólida, embora não excepcional, para tarefas de classificação onde tanto a precisão quanto a sensibilidade são importantes, mas não críticas.

O modelo Gemini demonstrou uma precisão relativamente alta de 0.496, sugerindo que é eficaz em evitar falsos positivos. No entanto, sua sensibilidade mais baixa, de 0.468, indica que ele pode deixar de identificar alguns verdadeiros positivos, resultando em uma maior taxa de falsos negativos. Isso faz do Gemini um modelo mais conservador, que pode ser útil em contextos em que é mais importante minimizar erros de classificação positiva já que classificações positivas incorretas podem levar a resultados prejudiciais, como diagnósticos errôneos ou sintomas incorretos. O *F-Score* de 0.467 confirma esse viés para a precisão em detrimento da sensibilidade.

O modelo Sabiá 2 Medium teve o pior desempenho médio entre os modelos avaliados. Com a precisão mais baixa de 0.399, ele apresenta uma maior propensão a falsos positivos, o que pode reduzir sua confiabilidade em aplicações onde a precisão é crucial. No entanto, sua sensibilidade de 0.545 é relativamente alta, indicando que ele é melhor em detectar verdadeiros positivos em comparação com Gemini, embora

ainda inferior ao Llama 3. O *F-Score* de 0.427 sugere que, apesar de suas limitações, o modelo Sabia-2 Medium pode ser útil em situações em que a sensibilidade é mais valorizada que a precisão.

## Conclusão

A escolha final do modelo deve levar em conta não apenas as métricas de desempenho, mas também os requisitos específicos da aplicação, custos e velocidade de processamento. O Llama-3 é indicado para aplicações que requerem muita precisão e sensibilidade, como em sistemas de análise de dados complexos e críticos. O GPT-3.5, com seu desempenho equilibrado e rápida resposta, pode ser ideal para aplicações que demandam uma boa combinação de precisão e sensibilidade junto com uma rápida execução, como em *chatbots* e sistemas de suporte ao cliente.

O modelo Gemini é recomendado para situações em que minimizar falsos positivos é crucial, oferecendo uma abordagem conservadora que pode ser valiosa em áreas como segurança e diagnóstico médico. Já o Sabiá 2 Medium, apesar de seu desempenho mais modesto, pode ser uma opção viável em contextos em que a detecção de verdadeiros positivos é mais importante do que evitar falsos positivos, como em sistemas de triagem inicial.

Em suma, a escolha do modelo mais adequado deve considerar não apenas as métricas de precisão, sensibilidade e F-score, mas também o contexto específico da aplicação, os custos envolvidos e a necessidade de velocidade no processamento. Cada modelo possui suas próprias vantagens e desvantagens, e a decisão final deve ser baseada em uma avaliação abrangente dessas variáveis.

## Agradecimentos

À Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná (FA) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelas bolsas de iniciação em desenvolvimento tecnológico e inovação (PIBIT), e à Pontifícia Universidade Católica do Paraná (PUCPR) pela isenção da taxa de mestrado, e à Fundação Zerbini e FOXCONN Brasil, pelas bolsas

de pesquisador como parte do projeto de pesquisa "Processamento de Linguagem Natural em Medicina Cardiovascular".

## Referências

1. Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining Electronic Health Records (EHRs). ACM Computing Surveys, 50(6), 1–40. doi:10.1145/3127881

2. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395–405. doi:10.1038/nrg3208

3. Assale, M., Dui, L. G., Cina, A., Seveso, A., & Cabitza, F. (2019). The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. Frontiers in Medicine, 6. doi:10.3389/fmed.2019.00066

3. Sun, Peng et al. "An overview of named entity recognition." 2018 International Conference on Asian Language Processing (IALP). IEEE, 2018. p. 273-278.

4. DA SILVA, Diego Pinheiro et al. "Exploring named entity recognition and relation extraction for ontology and medical records integration". Journal of Informatics in Medicine Unlocked vol. 43 (2023): 2352-9148. doi:10.1016/j.imu.2023.101381

5. Liu, Zhengliang, et al. "Deid-gpt: Zero-shot medical text de-identification by gpt-4." arXiv preprint arXiv:2303.11032 (2023).

6. Schneider, Elisa Terumi Rubel et al. "BioBERTpt: a portuguese neural language model for clinical Named Entity Recognition." Proceedings of the 3rd Clinical Natural Language Processing Workshop. 19 November 2020, 2020.

7.Schneider, E. T. R, et al., "CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese," 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 2023, pp. 378-381, doi: 10.1109/CBMS58004.2023.00247.

8. Oliveira, L.E.S.e., Peters, A.C., da Silva, A.M.P. et al.. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semantics. 2022;13(1):13. Published 2022 May 8. doi:10.1186/s13326-022-00269-1

9. https://openai.com/index/chatgpt/ [Internet]. San Francisco: OpenAI; c2024 [cited 2024 May 31]. Available from: https://openai.com/index/chatgpt/.

10. Apresentando o Gemini: nosso maior e mais hábil modelo de IA. [Internet]. California: Google; c2024 [cited 2024 May 31]. Available from: https://blog.google/intl/pt-br/novidades/tecnologia/apresentando-o-gemini-nosso-maior-e-mais-habil-modelo-de-ia/#mensagem-sundar.

11. https://llama.meta.com/llama3/ [Internet]. California: Meta; c2024 [cited 2024 May 31]. Available from: https://llama.meta.com/llama3/

12. https://www.maritaca.ai/sabia-2 Internet]. São Paulo: Maritaca AI; c2024 [cited 2024 May 31]. Available from: https://www.maritaca.ai/sabia-2

13. GE, Yao et al. "Few-shot learning for medical text: A review of advances, trends, and opportunities". Journal of Biomedical Informatics vol. 144 (2023): 1532-0464. doi: 10.1016/j.jbi.2023.104458

14. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."