A Hybrid Approach for Named Entity Recognition in Clinical Case Reports

Fernanda de O. Ramos[1], **Cauã G. Marvila**[1], **Cristiano da S. Colombo**[1]

[1] Coordenadoria de Sistemas de Informação – Instituto Federal do Espírito Santo (IFES)
Cachoeiro de Itapemirim – ES – Brazil

{fernandaoliveira.rms,caua.privg}@gmail.com, cristianos@ifes.edu.br

Abstract. ***Context****: Natural Language Processing (NLP) has shown significant impact in healthcare by enabling knowledge extraction from unstructured texts. However, in Brazilian Portuguese, there remains a gap in developing models capable of processing clinical case reports, rich sources of crucial information for diagnostics, treatments, and decision-making.* ***Problem****: While some models exist (e.g., BioBERTpt), they are scarce and deliver limited performance for Portuguese clinical texts. The lack of accurate entity recognition reduces the ability to transform raw data into structured information that supports clinical decision-making.* ***Solution****: This work presents ClinptHyb, a hybrid Named Entity Recognition (NER) model combining a spaCy-based pipeline (ClinptCy, trained from scratch on SemClinBr) with a Large Language Model (Llama-3.3-70b) as reviewer. The hybrid design enhances entity identification and helps detect inconsistencies or omissions in the corpus.* ***IS Theory****: Grounded in the Organizational Information Processing Theory (OIPT), the study emphasizes how structuring clinical narratives reduces information uncertainty and increases organizational capacity to process data, thereby improving decision-making in healthcare contexts.* ***Method****: Following Design Science Research, the study included corpus preprocessing, stratified dataset splitting, spaCy model training, and prompt-based LLM revision. Evaluation employed Precision, Recall, and F1-score against the SemClinBr gold standard.* ***Summary of Results****: ClinptHyb consistently improved over the baseline ClinptCy and outperformed BioBERTpt in 8 of the 10 evaluated categories. The Negation class reached an F1-score of 0.97, a critical result for safe clinical interpretation. The model also revealed annotation inconsistencies, highlighting its potential as a support tool for dataset refinement.* ***Contributions and Impact on the IS Area****: This research presents the first hybrid NER approach that combines spaCy with an LLM for Portuguese clinical texts, aligned with IS for data, information, and knowledge management. It contributes to decision support systems, clinical annotation workflows, and the advancement of NLP technologies adapted to the Brazilian healthcare context.*

1. Introdução

O Processamento de Linguagem Natural (*Natural Language Processing* - NLP) tem desempenhado um papel crescente na área da saúde, especialmente em países de língua inglesa, onde as técnicas e modelos de NLP têm avançado consideravelmente. No entanto, quando se trata de outros idiomas, como o Português do Brasil, ainda há uma lacuna significativa no desenvolvimento e aplicação dessas tecnologias, principalmente no

contexto clínico [Névéol et al. 2018]. Isso se torna um desafio particular ao lidar com dados médicos não estruturados, como os relatos de casos clínicos, que contêm informações cruciais para a prática médica, incluindo sintomas, diagnósticos e tratamentos, registrados em linguagem natural pelos profissionais de saúde [Pagano et al. 2024].

Apesar da riqueza de dados presentes nesses documentos, a extração de informações a partir de um grande volume de relatos clínicos através da leitura manual pode ser um processo desafiador. A falta de um padrão de escrita resulta em registros com formatos distintos, que variam de acordo com o profissional responsável [Pagano et al. 2024].

Neste contexto, diante da dificuldade de interpretação e correlação entre os relatos, este trabalho propõe o desenvolvimento de um modelo para o Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* - NER) em relatos de casos clínicos. O NER tem como finalidade identificar e classificar automaticamente entidades textuais, permitindo organizar e estruturar dados relevantes para a prática médica. Essa abordagem possibilita a geração de *insights* úteis aos profissionais de saúde, favorecendo decisões fundamentadas em informações extraídas dos relatos [Pagano et al. 2024]. A construção de um modelo como esse representa uma contribuição significativa para as pesquisas que envolvem NLP aplicada à saúde no Brasil.

A teoria que fundamenta este trabalho é a *Organizational Information Processing Theory* (OIPT), que se concentra na forma como as organizações lidam com a incerteza e processam informações complexas. No contexto da saúde, a extração de dados de relatos clínicos pode ser entendida como um processo de aquisição, filtragem e interpretação de informações essenciais à prática médica. De acordo com a OIPT, organizações buscam reduzir incertezas e aprimorar a tomada de decisões por meio de mecanismos que aumentam sua capacidade de processar dados [Kmetz 2018]. Nessa perspectiva, o modelo NER proposto contribui para a gestão de dados clínicos ao transformar informações não estruturadas em registros estruturados, fortalecendo a qualidade das informações utilizadas em processos de análise e suporte à decisão.

Este trabalho está estruturado da seguinte forma: a Seção 2 apresenta a fundamentação teórica do estudo, abordando Processamento de Linguagem Natural, Extração de Informação, Reconhecimento de Entidades Nomeadas, Relatos de Casos Clínicos, o *dataset* SemClinBr, a biblioteca spaCy, LLMs e *Prompting*; a Seção 3 discute os trabalhos correlatos; a Seção 4 descreve a metodologia adotada; a Seção 5 analisa os experimentos e resultados obtidos; a Seção 6 reúne as considerações finais; por fim, a Seção 7 apresenta as propostas de trabalhos futuros.

2. Fundamentação Teórica

2.1. Processamento de Linguagem Natural

O Processamento de Linguagem Natural, do inglês *Natural Language Processing* (NLP), é uma área de estudo voltada à investigação e resolução de problemas que envolvem a linguagem humana através de ferramentas computacionais. No trabalho de [Caseli et al. 2024], os autores destacam que por meio de abordagens estatísticas e neurais, os sistemas de NLP podem aprender com exemplos, utilizando técnicas de Aprendizado de Máquina (*Machine Learning* - ML) para extrair informações relevantes de dados não estruturados.

> ”De modo geral, em PLN buscam-se soluções para problemas computacionais, ou seja, tarefas, sistemas, aplicações ou programas, que requerem o tratamento computacional de uma língua (português, inglês etc.), seja escrita (texto) ou falada (fala). Línguas como as de sinais também têm sido alvo de estudos da área.” [Caseli et al. 2024]

2.2. Extração de Informação

A Extração de Informação, ou *Information Extraction* (IE), é uma subárea do Processamento de Linguagem Natural cujo objetivo é obter o conteúdo semântico de um texto, convertendo dados não estruturados em informação estruturada [Claro et al. 2024]. A IE pode ser subdividida com base no tipo de informação a ser extraída do texto, abrangendo diferentes tarefas, dentre as quais o Reconhecimento de Entidades Nomeadas destaca-se como foco principal do presente estudo.

2.3. Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* - NER) é a tarefa de identificar e classificar expressões linguísticas chamadas Entidades Nomeadas (*Named Entities* - NE). Estas por sua vez fazem referência a entidades específicas em um determinado domínio, como nomes próprios, expressões temporais e espécies biológicas [Nadeau 2007]. Em geral, o processo do NER é dividido em duas etapas: primeiro, a identificação (ou delimitação), onde são selecionadas as palavras que compõem as NEs. Em seguida, a classificação atribui uma categoria semântica à Entidade Nomeada [Claro et al. 2024].

No contexto do NER, considere a frase:

> ”Estar acima do peso pode fazer com que apareça o diabetes tipo 2.” [Pavanelli et al. 2023]

Neste exemplo, a etapa de identificação consiste em reconhecer ”**Estar acima do peso**” e ”**diabetes tipo 2**” como entidades relevantes. Em seguida, o processo de classificação define ”**Estar acima do peso**” como uma Complicação (*Complication*) e ”**diabetes tipo 2**” como Tipo de Diabete (*DiabetesType*).

2.4. Relatos de Casos Clínicos como fonte de dados

Relatos de Casos Clínicos são descrições detalhadas de casos clínicos observados por profissionais de saúde, contendo informações essenciais sobre o paciente, como sintomas, histórico médico e estilo de vida [Pagano et al. 2024]. Essas narrativas podem incluir questionários preenchidos pelo paciente, registros de exames, e anotações feitas por profissionais de saúde, como notas de evolução de enfermagem, sumários de alta, boletins médicos e observações em texto livre nos prontuários eletrônicos.

Partindo de um modelo treinado para o Reconhecimento de Entidades Nomeadas em textos clínicos, é possível identificar e classificar automaticamente entidades relevantes como medicamentos, doenças, sintomas, entre outros. A análise desses dados pode ser empregada para detectar padrões e organizar o conhecimento presente nos relatos, favorecendo uma compreensão mais aprofundada da condição do paciente [Pagano et al. 2024]. Como resultado, fornece *insights* relevantes aos profissionais de saúde, possibilitando tomadas de decisões clínicas mais assertivas.

2.5. SemClinBr

O SemClinBr é um dataset anotado desenvolvido para apoiar tarefas de NLP no contexto clínico em português [Oliveira et al. 2022]. Construído a partir de textos clínicos obtidos de diversas instituições e especialidades médicas, o corpus é composto por 1.000 notas clínicas que foram rotuladas com 65.117 entidades e 11.263 relações. O SemClinBr destaca-se por oferecer anotações detalhadas que seguem padrões semânticos específicos para a área da saúde, abrangendo categorias como problemas de saúde, exames, tratamentos e sinais clínicos. As anotações foram validadas por especialistas médicos e linguistas, garantindo a sua qualidade e a sua consistência, sendo um recurso valioso para treinar e avaliar modelos voltados à compreensão de linguagem em cenários clínicos.

2.6. spaCy

A documentação oficial define o spaCy como uma biblioteca gratuita de código aberto voltada para o Processamento de Linguagem Natural em Python. Essa biblioteca fornece ferramentas robustas que realizam análise e compreensão de grandes volumes de dados, permitindo sua utilização para diferentes tarefas de NLP [spaCy 2025].

Uma das vantagens do spaCy é a modularização, que permite a criação de uma sequência customizada de componentes na *pipeline*. Além disso, a biblioteca possibilita o treinamento individual de cada componente, facilitando a adaptação com base nas tarefas de NLP desejadas. O presente estudo apresenta o treinamento do componente NER a partir do zero, utilizado para a identificação de entidades relevantes em relatos de casos clínicos.

Dentre os modelos pré-treinados disponíveis na língua portuguesa, o *pt_core_news_lg* [1] foi selecionado e sua utilização será descrita posteriormente na metodologia.

2.7. LLMs

[Wulff et al. 2025] define os *Large Language Models* (LLMs) como modelos baseados na arquitetura *Transformer* [2] que, através do treinamento auto-supervisionado em grandes volumes de textos, desenvolvem a capacidade de compreender diferentes dimensões da linguagem humana (sintaxe, semântica e pragmática), além de incorporar conhecimento de mundo. Essas habilidades permitem sua aplicação em diferentes tarefas de NLP, como tradução, sumarização e NER. Entretanto, o desempenho dos LLMs depende diretamente da representatividade dos dados de treinamento, o que limita sua aplicação em domínios especializados [Wulff et al. 2025]. Os termos técnicos e contextos específicos dos domínios biomédico e o clínico, por exemplo, são pouco frequentes nos corpora generalistas, o que tende a impactar negativamente o desempenho desses modelos.

Uma estratégia para adaptar os LLMs a contextos específicos é o uso do *fine-tuning* [3]. No entanto, conforme descrito por [Wulff et al. 2025], essa abordagem exige

[1] O sufixo lg (*large*) indica uma maior quantidade de dados e vetores de palavras, o que o torna mais robusto e preciso para tarefas que exigem maior detalhamento semântico [spaCy 2025].

[2] A arquitetura *Transformer*, proposta por [Vaswani et al. 2017], introduziu mecanismos de atenção que permitem ao modelo relacionar diferentes partes de uma sequência de texto, tornando-se a base de modelos amplamente utilizados, como BERT e GPT.

[3] Processo de ajuste de um modelo pré-treinado em um conjunto de dados específico, com o objetivo de adaptá-lo a uma nova tarefa ou domínio [Jurafsky and Martin 2024].

elevado custo computacional e grandes volumes de dados anotados, o que a torna inviável no contexto deste estudo. Como alternativa, esta pesquisa propõe uma uma abordagem híbrida, combinando a especialização de um modelo spaCy treinado no corpus SemClinBr com a versatilidade de um LLM revisor. O objetivo é equilibrar precisão em domínios específicos e capacidade de generalização na tarefa de NER.

2.8. *Prompting*

Segundo [Jurafsky and Martin 2024], o *prompting* consiste em fornecer instruções ou exemplos diretamente no texto de entrada de um LLM, orientando a forma como o modelo produz sua saída. Ou seja, um *prompt* é uma sequência textual que guia o LLM na execução de uma tarefa específica, como responder perguntas, realizar traduções ou classificar sentimentos. O processo de elaborar instruções claras e eficazes para otimizar o desempenho do modelo é conhecido como *prompt engineering* (engenharia de prompt).

[Jurafsky and Martin 2024] destaca que existem diferentes estratégias de prompting, entre elas:

- *Zero-shot prompting*: o modelo recebe apenas a instrução, sem exemplos adicionais.
- *Few-shot prompting*: o modelo recebe alguns exemplos de entrada e saída no prompt para orientar a tarefa.

A elaboração do *prompt* tem grande influência na qualidade da resposta, sendo possível estruturar instruções mais restritivas (como oferecer opções fixas de resposta) para aumentar a consistência e reduzir ambiguidades [Jurafsky and Martin 2024].

3. Trabalhos Relacionados

O trabalho de [Lee et al. 2020] introduziu o BioBERT, um modelo de linguagem baseado no BERT, pré-treinado com grandes corpora biomédicos, como os resumos do PubMed e os artigos completos do PubMed (*PubMed Central full-text articles* - PMC). O modelo foi avaliado em tarefas como NER, Extração de Relações e Resposta a Perguntas, demonstrando um desempenho superior a modelos genéricos. Em contraste com o BioBERT, voltado ao inglês e a textos biomédicos gerais, o presente estudo busca adaptar a tarefa de NER ao contexto específico dos relatos clínicos na língua portuguesa.

O trabalho de [Schneider et al. 2020], por sua vez, apresenta o BioBERTpt, um modelo de linguagem adaptado para o português, focado em tarefas NER em textos clínicos e biomédicos. Foram desenvolvidas três versões do modelo: o BioBERTpt(clin), treinado com notas clínicas; o BioBERTpt(bio), utilizando textos biomédicos; e o BioBERTpt(all), que combina ambos os conjuntos de dados. Os resultados da avaliação indicaram que o BioBERTpt(clin) superou os modelos de base [4] em 11 das 13 categorias de entidades no corpus SemClinBr e alcançou um *F1-score* de 92,6% no CLINpt, estabelecendo um novo padrão de excelência para esse corpus. Em contrapartida, esta pesquisa adota uma abordagem híbrida de menor custo, que combina um modelo spaCy treinado do

[4]Modelos de base, treinados em grandes conjuntos de dados gerais, servem como referência para avaliar o desempenho de novos modelos em tarefas específicas. Eles permitem uma comparação eficaz com modelos especializados, como o BioBERTpt, adaptado para textos clínicos e biomédicos, facilitando a avaliação na tarefa de NER.

zero com um LLM revisor, dispensando o re-treinamento completo de grandes modelos *Transformer*.

[Torres 2024] aborda a utilização do spaCy para o reconhecimento de entidades nomeadas em textos clínicos. Um dos modelos desenvolvidos por [Torres 2024] utilizou uma amostra do dataset SemClinBr, com destaque para a categoria *Finding* (retratada como Conclusão) que obteve *F1-score* de 71%. Os resultados obtidos indicam um potencial promissor do spaCy para a classificação de entidades clínicas em português. Enquanto o trabalho de [Torres 2024] avaliou apenas o desempenho isolado do spaCy, o presente estudo amplia essa abordagem ao integrá-lo com um LLM, melhorando o desempenho alcançado.

[Mello et al. 2024], por fim, avaliou a eficácia de diferentes LLMs na tarefa de NER clínico em portugues, utilizando 30 relatos do corpus SemClinBr. Foram avaliados os modelos GPT-3.5, Gemini, Llama-3 e Sabiá-2, considerando três categorias de entidades: Sinais ou Sintomas, Doenças ou Síndromes e Dados Negados. Os resultados mostraram que o Llama-3 apresentou o melhor desempenho, com *F1-score* de 0,538 e destaque para o *Recall*. O GPT-3.5, por sua vez, obteve desempenho equilibrado, enquanto o Gemini obteve maior *Precision*, mas com *Recall* reduzido. O estudo demonstra o potencial dos LLMs para a extração de informações clínicas em português, ressaltando que a escolha do modelo deve considerar os requisitos específicos de cada aplicação. Diferentemente dos experimentos comparativos entre LLMs realizados por [Mello et al. 2024], o presente trabalho amplia a análise ao apresentar métricas detalhadas por categoria de entidade e comparações diretas com o estado da arte do corpus SemClinBr.

Os trabalhos correlatos apresentados nesta seção reforçam a relevância desta pesquisa ao buscar soluções alternativas para desenvolver um modelo adaptado às especificidades dos relatos clínicos em português. Nesse contexto, pretende-se explorar o dataset SemClinBr e avaliar potenciais melhorias na classificação de suas entidades por meio uma abordagem híbrida, que combina um modelo spaCy com um LLM revisor na tarefa de NER.

4. Metodologia

Este capítulo descreve a metodologia utilizada no presente estudo, que foi estruturada em etapas bem definidas, visando a transparência e a reprodutibilidade. O processo iniciou-se com a análise do *dataset*, no qual foi aplicado um algoritmo de estratificação para a divisão em conjuntos de treino, teste e validação. Ainda nessa fase, os arquivos originais em formato XML foram convertidos para JSON, de modo a torná-los compatíveis com a biblioteca spaCy. Em seguida, realizou-se o treinamento do modelo spaCy, com a criação do componente NER a partir do zero. Por fim, foi desenvolvido o algoritmo híbrido, que integra o novo modelo spaCy a um LLM revisor.

4.1. Análise do Dataset

Nessa etapa, foi realizada a análise do *dataset* e a definição das categorias de interesse, a fim de facilitar sua utilização no treinamento do modelo. Inicialmente, desenvolveu-se um *script* que percorreu todos os arquivos e retornou uma lista com todas as entidades nomeadas encontradas. No total foram identificadas 923 entidades distintas, número elevado em razão da presença de múltiplas *tags* (etiquetas) atribuídas a uma mesma sequência de

caracteres. Por exemplo, o trecho "**COMORBIDADES**" aparecia anotado como "*Disease or Syndrome/Disease or Syndrome*", o que seria interpretado pelo spaCy como uma categoria diferente de "*Disease or Syndrome*".

Optou-se por não trabalhar com múltiplas anotações para a mesma entidade, uma vez que o spaCy não possui suporte nativo para tal funcionalidade. Para isso, a primeira fase consistiu em mapear todas as anotações duplicadas de uma mesma entidade - como o exemplo acima - e substituí-las por uma única classificação. Assim, "**COMORBIDADES**" passou a ser anotado apenas como "*Disease or Syndrome*". Na segunda etapa realizou-se o mapeamento das abreviações, mantendo apenas a entidade clínica associada. Por exemplo, o termo "**AVC**", anteriormente anotado como "*Disease or Syndrome/Abbreviation*" passou a ser classificado exclusivamente como "*Disease or Syndrome*".

Por fim, foram selecionadas 12 categorias de interesse, descritas a seguir:

1. *Chemicals and Drugs (Chem)*: substância farmacológica ou preparação usada no diagnóstico, tratamento ou prevenção de doenças.
2. *Diagnostic Procedure (DProced)*: procedimento, método ou técnica utilizada para determinar a natureza ou identidade de uma doença ou distúrbio. Excluem-se aqui procedimentos realizados exclusivamente em espécimes de laboratório.
3. *Disease or Syndrome (DisSynd)*: condição que altera ou interfere no funcionamento normal de um organismo, geralmente caracterizada por alterações patológicas em sistemas, órgãos ou partes do corpo.
4. *Disorders (Disor)*: alteração estrutural ou lesão de uma parte do corpo resultante de causa externa ou trauma. Inclui fraturas, luxações ou lesões traumáticas.
5. *Finding (Find)*: constatação clínica obtida por meio de observação direta ou pela mensuração de um atributo ou condição de um organismo, incluindo o histórico clínico do paciente.
6. *Health Care Activity (Health)*: atividade relacionada à prática médica ou ao cuidado de pacientes.
7. *Laboratory or Test Result (LResult)*: resultado ou interpretação derivada de exame laboratorial ou teste diagnóstico.
8. *Laboratory Procedure (LProced)*: procedimento ou técnica usada para determinar composição, quantidade ou concentração de uma amostra, realizado em laboratório clínico.
9. *Medical Device (MDev)*: objeto fabricado utilizado principalmente no diagnóstico, tratamento ou prevenção de distúrbios fisiológicos ou anatômicos.
10. *Negation (Neg)*: expressão linguística que indica ausência ou negação de conceito clínico em texto narrativo.
11. *Sign or Symptom (Sign)*: manifestação clínica de doença ou condição: sinal observável pelo profissional ou sintoma subjetivo relatado pelo paciente, indicando desvio do normal.
12. *Therapeutic or Preventive Procedure (TProced)*: procedimento, método ou técnica destinado a prevenir doença/distúrbio, melhorar função física ou tratar doença/lesão.

Após a definição das categorias de interesse, realizou-se a divisão estratificada do dataset, com o objetivo de garantir que as entidades nomeadas estivessem proporcionalmente representadas nos conjuntos treino, teste e validação.

Para a divisão, utilizou-se a biblioteca iterative-stratification, implementada em Python como extensão do scikit-learn, a qual disponibiliza métodos específicos para problemas de classificação multirrótulo [Sechidis et al. 2024]. O processo resultou em três subconjuntos na proporção de 8:1:1, correspondentes a treino (80%), teste (10%) e validação (10%). Considerando o total de 1000 relatos clínicos, a divisão produziu 800 arquivos para treino, 100 para validação e 100 para teste. Essa estratégia busca evitar desequilíbrios entre as classes, preservando a distribuição original dos dados e contribuindo para a confiabilidade das avaliações posteriores.

Por fim, os arquivos originais em XML foram convertidos para o formato JSON, no qual cada relato foi representado como um dicionário contendo as chaves *text* e *entities*: a primeira armazena o relato em linguagem natural, enquanto a segunda corresponde a uma lista de anotações estruturada no formato [*start, end, entity, text*]. Nesse caso, *start* e *end* indicam as posições da entidade no texto, *entity* refere-se ao rótulo atribuído e *text* armazena a sequência de caracteres correspondente. A Figura 1 apresenta um exemplo de relato de caso clínico anotado após a conversão.

```
"text": "POI DE LAVAGEM + CURETA DE TECIDO NECRÓTICO. 18:00: PACIENTE RETORNOU DO CC LÚCIDO,
"entities": [
    [ 7, 14, "Procedure", "LAVAGEM" ],

    [ 17, 23, "Procedure", "CURETA"],

    [ 52, 60, "Patient or Disabled Group", "PACIENTE" ],

    [ 76, 82, "Finding", "LÚCIDO"],
```

Figura 1. Exemplo de anotação de um Relato de Caso Clínico no formato JSON. Fonte: a autora.

4.2. Treinamento do modelo spaCy

O treinamento de um modelo com spaCy é um processo cíclico que utiliza exemplos anotados como base. Esses exemplos são compostos por um texto e pelos rótulos (*labels*) a serem previstos, como categorias gramaticais ou entidades nomeadas. Durante o treinamento, o modelo recebe o texto, gera uma predição e compara essa saída com o rótulo original. A diferença entre as previsões do modelo e as *labels* de referência permite o cálculo do gradiente, que indica como os parâmetros internos do modelo devem ser ajustados para reduzir o erro. A Figura 2 ilustra esse ciclo, cujo resultado final é um modelo atualizado que pode ser salvo e utilizado em diferentes tarefas de NLP.[spaCy 2025].

O componente selecionado para treinamento nesta pesquisa foi o NER, e a metolodogia segue a documentação oficial da biblioteca spaCy [spaCy 2025]. Inicialmente, foi criado um arquivo de configuração **config.cfg** [5], no qual foram definidos os componentes da *pipeline* e os hiperparâmetros de treino.

A Figura 3 apresenta os componentes do novo modelo, sendo eles:

[5]Esse arquivo centraliza todas as definições da *pipeline* e dos hiperparâmetros, o que favorece a reprodutibilidade e o versionamento do modelo. Assim, diferentes usuários podem treinar *pipelines* equivalentes com as mesmas configurações [spaCy 2025].

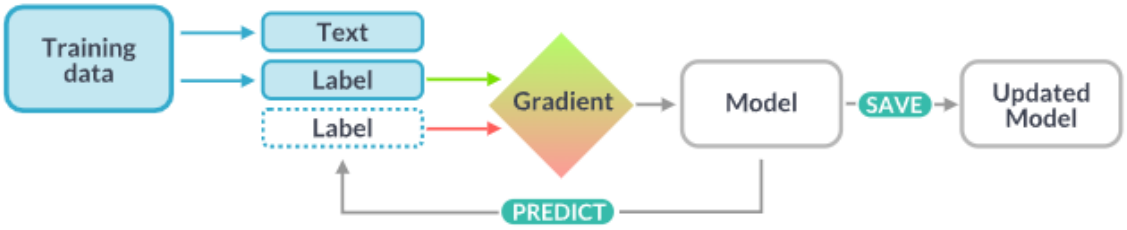

Figura 2. Diagrama demonstrando o processo de treinamento do spaCy. Fonte: [spaCy 2025].

1. *Tokenizer*: divide o texto em *tokens* e cria o objeto Doc. Cada língua possui um *tokenizer* padrão definido pelo spaCy e neste estudo o português foi utilizado.
2. Tok2Vec: converte cada *token* em uma representação vetorial, que será utilizada pelos demais componentes. Foi inicializado a partir do modelo pré-treinado *pt_core_news_lg*, garantindo representações mais adequadas à língua portuguesa.
3. NER: realiza o reconhecimento de entidades nomeadas, identificando *spans* de texto e classificando-os em categorias específicas. Esse componente foi treinado do zero a partir do corpus SemClinBr, visando a aprendizagem de entidades específicas do domínio clínico.

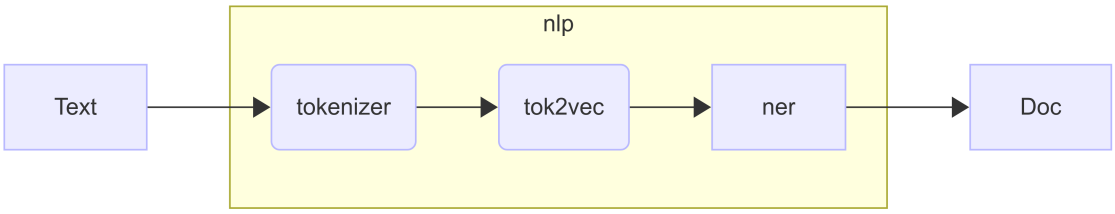

Figura 3. Diagrama demonstrando os componentes da *pipeline* do novo modelo spaCy. Fonte: a autora.

Ainda no **config.cfg**, são definidos os caminhos para os arquivos binários de treino e teste no formato .spacy que, conforme descrito anteriormente, contém 800 e 100 relatos de casos clínicos, respectivamente. Para fins didáticos, a partir desse ponto o novo modelo spaCy será tratado como ClinptCy (*Clinical Portuguese spaCy Model*).

O treinamento é executado a partir de uma instrução na linha de comando e esse processo utiliza os arquivos treino.spacy e teste.spacy. O primeiro serve como base para o aprendizado, fornecendo os relatos de casos clínicos já anotados. Já o segundo é utilizado para ajustar os hiperparâmetros do modelo e monitorar seu desempenho durante o treinamento, evitando o *overfitting* [6].

4.3. Treinamento do Modelo Híbrido (spaCy + Llama)

Após experimentos iniciais realizados com o ClinptCy, descritos posteriormente no Capítulo 5, observou-se que, apesar de os resultados serem satisfatórios quando comparados a outros modelos disponíveis na literatura, ainda existia uma margem significativa

[6] *Overfitting* é um problema em aprendizado de máquina que ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, perdendo a capacidade de generalizar para novos dados.

para melhorias. A partir dessa constatação surge a proposta central do presente estudo: o desenvolvimento de um algoritmo híbrido que integra o spaCy e um LLM revisor, buscando aprimorar o desempenho inicial obtido.

A próxima etapa consistiu na construção do *prompt*, elemento fundamental para o uso adequado de LLMs. Os testes iniciais utilizando a estratégia *few-shot prompting* não trouxeram ganhos de consistência, além de aumentar o tamanho e a complexidade do *prompt*. Portanto, optou-se por adotar a abordagem de *zero-shot prompting*, cuja estrutura principal, apresentada na Figura 4, consiste em:

- Descrição geral da tarefa, indicando ao LLM para atuar como um revisor de anotações médicas.
- Objeto JSON contendo o texto clínico e as anotações realizadas pelo ClinptCy.
- Lista de entidades e suas respectivas definições, extraídas diretamente da *UMLS Semantic Network* [7].
- Descrição mais detalhada da tarefa, especificando as instruções através de uma lista numerada.
- Especificação do formato de saída, exigindo que as respostas sejam fornecidas estritamente em JSON.

Além disso, foram adicionadas regras complementares a fim de reduzir ambiguidades e comportamentos inesperados observados durante os testes.

```
messages = [
    SystemMessage(content="""Você é um revisor de anotações médicas.
    Sua função é revisar e corrigir entidades extraídas de prontuários médicos, garantindo consistência e precisão."""),
    HumanMessage(content=f"""
    ### Entrada
    Prontuário médico (texto e anotações atuais):
    {json.dumps(clinical_report, ensure_ascii=False, indent=2)}

    ### Definições de entidades (labels em inglês, descrições em português):
    {json.dumps(ENTITIES_LIST, ensure_ascii=False, indent=2)}
    {json.dumps(ENTITIES_DEFINITION, ensure_ascii=False, indent=2)}

    ---

    ### Tarefa
    1. Revisar entidades anotadas no prontuário.
    2. Corrigir labels incorretos conforme as definições.
    3. Adicionar entidades faltantes.
    4. Remover entidades não suportadas.
    5. Mesclar duplicatas (ignorar maiúsculas/minúsculas, acentuação, espaços).
    6. Evitar anotações artificiais ou redundantes.
    7. **Verificar sistematicamente negação**: identificar e anotar gatilhos de negação (label "Negation") quando presentes.

    ---

    ### Regras do output
    - Retornar **apenas JSON**, sem explicações adicionais.
    - Formato JSON esperado (schema):
    {json_output_str}
```

Figura 4. Estrutura principal do prompt utilizado nos experimentos. Fonte: a autora.

A etapa seguinte consistiu na escolha do LLM e os primeiros testes foram reali-

[7] A *UMLS Semantic Network* fornece uma estrutura consistente de categorias semânticas (como *Disease or Syndrome*, *Sign or Symptom*, entre outros) e os relacionamentos entre elas, servindo como base para a organização e interpretação de conceitos biomédicos [U.S. National Library of Medicine 2025].

zados com modelos locais através do Ollama[8]. No entanto, a limitação computacional foi um fator decisivo, uma vez que as configurações do computador utilizado não eram compatíveis com os modelos que possuíam um maior número de parâmetros. Entre os modelos testados nessa fase (llama3.1:8b, gemma3:4b, mistral:7b, qwen3:4b), nenhum apresentou desempenho satisfatório na tarefa de revisão, cometendo muitas alucinações e fugindo do contexto restrito dos casos clínicos apresentados.

Buscando a utilização de LLMs mais robustos, a ferramenta Groq foi a escolhida, pois disponibiliza gratuitamente modelos de produção acessíveis via API [Groq 2025]. Os modelos testados foram o llama-3.3-70b-versatile (Meta) e o openai/gpt-oss-120b (OpenAI). Para simplificação, eles serão referidos neste estudo como Llama e GPT-OSS, respectivamente. A Tabela 1 apresenta informações mais detalhadas a respeito dos modelos, trazendo a janela de contexto (*Context Window*) e o número máximo de tokens de saída (*Max Output Tokens*).

Tabela 1. Janela de contexto e número máximo de tokens de saída dos modelos Llama e GPT-OSS. Fonte: a autora.

Modelo	Groq ID	Janela de Contexto	Max. Tokens de Saída
Llama	llama-3.3-70b-versatile	131.072	65.536
GPT-OSS	openai/gpt-oss-120b	131.072	32.768

Nesta etapa, aplicou-se o mesmo algoritmo de estratificação descrito no Subtópico 4.1 para segmentar o conjunto de validação, reduzindo a quantidade de arquivos necessários para a validação manual, mas preservando a proporção original das entidades. Inicialmente, dos 100 arquivos disponíveis, 12 foram selecionados para compor o conjunto de validação. Além disso, optou-se por remover três entidades da lista original, mantendo assim as seguintes categorias: *Diagnostic Procedure, Disease or Syndrome, Finding, Health Care Activity, Laboratory Procedure, Medical Device, Negation, Sign or Symptom* e *Therapeutic or Preventive Procedure*.

Em seguida, o novo conjunto de validação foi submetido ao modelo híbrido, cuja arquitetura está representada na Figura 5. Inicialmente, o texto é processado pelo ClinptCy, responsável por anotar as entidades clínicas e retornar um objeto Doc. A seguir, esse objeto Doc é convertido para o formato JSON, armazenando o texto original e suas respectivas anotações. Por fim, esses dados servem de entrada para o LLM, que atua como revisor e gera a versão final das entidades anotadas, também em formato JSON.

A fim de selecionar o LLM com melhor desempenho para a tarefa de revisão das anotações, o mesmo experimento foi aplicado tanto ao Llama quanto ao GPT-OSS. Os conjuntos de anotações gerados foram comparados com o *dataset* original do SemClinBr, considerado o *gold standard* (padrão-ouro) de referência para a avaliação no presente estudo.

A partir dessa comparação, foram calculadas as métricas de *Precision* (P), *Recall* (R) e *F1-Score* (F1), cujos resultados foram apresentados nas Tabelas 2 e 3. A Tabela 2

[8]Ollama é uma ferramenta que permite executar modelos de linguagem localmente em diferentes sistemas operacionais, oferecendo suporte a diversos modelos e facilitando sua integração e aplicações [Ollama 2025].

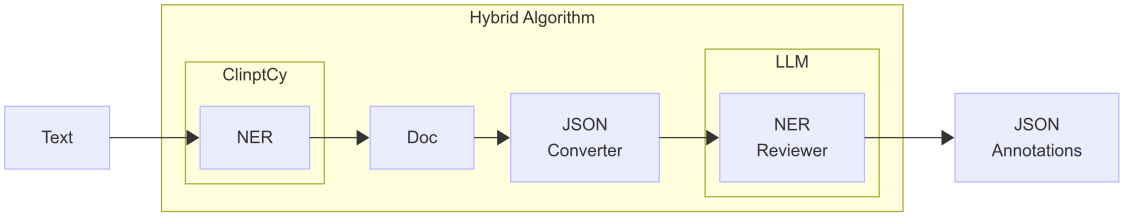

Figura 5. Diagrama demonstrando o funcionamento do novo algoritmo híbrido. Fonte: a autora.

apresenta os valores de F1 por entidade, evidenciando que o Llama obteve melhor desempenho em todas as 9 categorias analisadas. Já a Tabela 3 reúne as métricas globais de P, R e F1, demonstrando que o Llama superou o GPT-OSS em 22,8 pontos percentuais em precisão, 8,5 em revocação e 15,6 em *F1-score*.

Tabela 2. Valores de *F1-score* por entidade para os modelos Llama e GPT-OSS. Fonte: a autora.

Modelo	DProced	DisSynd	Find	Health	LProced	MDev	Neg	Sign	TProced
GPT-OSS	0,545	0,683	0,576	0,511	0,1540	0,538	0,824	0,610	0,563
Llama	**0,737**	**0,889**	**0,688**	**0,632**	**0,571**	**0,667**	**0,848**	**0,747**	**0,634**

Tabela 3. Métricas globais de *Precision* (P), *Recall* (R) e *F1-Score* (F1) para os modelos Llama e GPT-OSS. Fonte: a autora.

Modelo	P	R	F1
GPT-OSS	0,557	0,621	0,587
Llama	**0,785**	**0,706**	**0,743**

Em razão do desempenho superior do Llama em relação ao GPT-OSS, ele foi o LLM selecionado para compor o algoritmo híbrido. Durante o processo de validação, foram identificados dois padrões de classificação incorreta: medicamentos com posologia eram classificados como *Therapeutic or Preventive Procedure*, enquanto traumatismos e lesões eram anotados como *Sign or Sympton*. Buscando aprimorar os resultados, o prompt foi ajustado para reinserir as três entidades anteriormente removidas, resultando em uma lista final composta pelas 12 categorias originais apresentadas no Subtópico 4.1.

Por fim, o escopo de validação foi ampliado com a aplicação do mesmo algoritmo de estratificação, desta vez considerando 33 dos 100 arquivos disponíveis. O cálculo das métricas seguiu o mesmo procedimento já estabelecido, e os resultados foram apresentados no Capítulo 5. Para fins didáticos, a partir desse ponto o novo algoritmo híbrido será referido como ClinptHyb (*Hybrid Clinical Portuguese Model*).

5. Experimentos e Resultados

Este capítulo apresenta os resultados dos experimentos realizados com os modelos desenvolvidos no presente estudo. Inicialmente, foram comparados o ClinptCy e o ClinptHyb,

com destaque para os ganhos de desempenho proporcionados pela abordagem híbrida. Em seguida, os valores de *F1-score* de ambos os modelos foram confrontados com resultados da literatura, indicando avanços relevantes e o estabelecimento de novos padrões de desempenho para algumas categorias do SemclinBr. Por fim, foram apresentados achados pontuais sobre a consistência do *dataset*, destacando casos de falta de padronização em algumas anotações e omissão de entidades relevantes.

Nas subseções seguintes, são apresentados os resultados comparativos entre o modelo spaCy e o modelo híbrido para verificar a melhor performance entre eles. Também são analisados os resultados de ambos perante aos modelos estado da arte do SemClinBr.

5.1. spaCy e Abordagem Híbrida

O primeiro experimento teve como objetivo comparar os dois modelos desenvolvidos neste estudo, avaliando o ganho de desempenho ao substituir o modelo spaCy isolado pelo algoritmo híbrido proposto. Para o cálculo das métricas do ClinptCy, foram utilizados todos os 100 arquivos disponíveis para validação, em um processo automatizado por meio de uma instrução de linha de comando disponível na biblioteca spaCy. Já o ClinptHyb contou com o cálculo manual das métricas. Conforme descrito no capítulo anterior, foi utilizada uma amostra de 33 relatos clínicos do conjunto de validação, selecionados por meio de estratificação a fim de preservar a proporção original de entidades.

Os resultados apresentados na Tabela 4 destacam a superioridade do ClinptHyb em todas as categorias quando comparado com o ClinptCy. O principal destaque foi a categoria *Negation*, que alcançou um F1-score de 97%. Além disso, as categorias *Chemicals and Drugs, Diagnostic Procedure, Disease or Syndrome* e *Medical Device* também obtiveram desempenho expressivo, com valores F1 acima de 85%. Já a Tabela 5 reforça o ganho de desempenho do ClinptHyb, mostrando melhorias consistentes em todas as métricas avaliadas: *Precision*, *Recall* e *F1-score*.

Tabela 4. Valores de *F1-score* por entidade para os modelos ClinptCy e ClinptHyb. Fonte: a autora.

Modelo	Chem	DProced	DisSynd	Disor	Find	Health	LResult	LProced	MDev	Neg	Sign	TProced
ClinptCy	0,822	0,682	0,738	0,530	0,716	0,604	0,562	0,388	0,700	0,872	0,611	0,616
ClinptHyb	**0,905**	**0,867**	**0,866**	**0,696**	**0,783**	**0,726**	**0,713**	**0,545**	**0,873**	**0,972**	**0,767**	**0,773**

Tabela 5. Métricas globais de *Precision* (P), *Recall* (R) e *F1-Score* (F1) para os modelos ClinptCy e ClinptHyb. Fonte: a autora.

Modelo	P	R	F1
ClinptCy	0,703	0,680	0,691
ClinptHyb	**0,875**	**0,752**	**0,809**

5.2. Análise de Resultados frente ao Estado da Arte

O segundo experimento avaliou o desempenho dos modelos desenvolvidos neste estudo frente às variações do BERT apresentadas por [Schneider et al. 2020], que até então representavam o melhor resultado para o corpus SemClinBr. Conforme demonstrado na

Tabela 6. Valores de *F1-score* por entidade dos modelos deste estudo (ClinptCy e ClinptHyb) em comparação com modelos da literatura (BERT e BioBERTpt). Fonte: a autora.

Modelo	Chem	DProced	DisSynd	Disor	Find	Health	LResult	MDev	Sign	TProced
ClinptCy	0,822	0,682	0,738	0,530	0,716	0,604	0,562	0,700	0,611	0,616
ClinptHyb	0,905	**0,867**	**0,866**	0,696	**0,783**	**0,726**	**0,713**	**0,873**	**0,767**	**0,773**
BERT multilingual uncased	0,903	0,546	0,562	0,787	0,503	0,374	0,378	0,559	0,519	0,487
BERT multilingual cased	0,901	0,519	0,538	0,782	0,505	0,412	0,417	0,593	0,537	0,486
Portuguese BERT large	0,782	0,504	0,575	0,625	0,526	0,336	0,404	0,514	0,552	0,489
Portuguese BERT base	0,904	0,556	0,540	0,784	0,500	0,346	0,422	0,537	0,538	0,471
BioBERTpt (bio)	0,894	0,551	0,575	0,785	0,526	0,459	0,398	0,604	0,534	0,501
BioBERTpt (clin)	**0,911**	0,560	0,583	0,781	0,521	0,406	0,453	0,562	0,544	0,459
BioBERTpt (all)	0,904	0,548	0,564	**0,791**	0,517	0,404	0,440	0,555	0,566	0,513

Tabela 6, o ClinptHyb superou todos os modelos em 8 das 10 categorias avaliadas, estabelecendo uma nova referência de desempenho para as entidades *Diagnostic Procedure, Disease or Syndrome, Finding, Health Care Activity, Laboratory or Test Result, Medical Device, Sign or Syptom* e *Therapeutic or Preventive Procedure*.

Além disso, destaca-se que o modelo spaCy isolado apresentou resultados robustos, superando as variações do BERT em diversas categorias. Esse resultado reforça sua relevância como base consistente para o desenvolvimento do algoritmo híbrido proposto.

5.3. Achados relacionados à consistência do *dataset*

Durante o processo de cálculo manual das métricas do ClinptHyb, foram identificadas duas inconsistências nas anotações originais do SemclinBr, cujas implicações serão discutidas a seguir.

A primeira diz respeito a falta de padronização (discordância aparente entre os anotadores) em algumas anotações, com casos em que uma mesma entidade nomeada apresenta limites diferentes dentro do próprio dataset. Considere o seguinte trecho extraído de um dos relatos:

> "Às 01:38: Lucido, nao verbalizando, traqueostomizado, **eupneico** em aa, afebril, apresenta fixador externo em MMSS . . . " [Oliveira et al. 2022]

Nessa sentença, a palavra "**eupneico**" foi anotada como *Finding*. Já em outro arquivo, em contexto muito semelhante, a entidade anotada como *Finding* foi "**eupneico em aa**":

> "Às 14:20 hrs: paciente consciente, comunicativo, **eupneico em aa**, em repouso no leito, sem queixas algicas." [Oliveira et al. 2022]

Entre os 33 arquivos avaliados, foram observadas cerca de 24 ocorrências semelhantes. Esse padrão não invalida o *dataset*, mas pode introduzir incertezas para o modelo, que passa a lidar com exemplos contraditórios. Em tarefas de NER, essas variações nos limites da entidade podem gerar confusão no aprendizado, impactando a consistência das previsões.

Já a segunda inconsistência refere-se a omissão de algumas entidades relevantes. Em um dos relatos, por exemplo, a negação "sem" aparecia três vezes, mas as anotações originais registraram apenas duas. O ClinptHyb, por sua vez, foi capaz de reconhecer de forma consistente todas as três ocorrências. No total, entre os 33 arquivos de validação, foram identificadas cerca de 60 anotações ausentes.

A confirmação desses achados foi feita por meio de uma comparação cruzada (perguntar ao Carlos) dentro do próprio SemclinBr: em trechos equivalentes de outros arquivos, a anotação omitida estava presente, o que sugere falhas pontuais no processo de anotação. Essas omissões não comprometem a validade geral do corpus, mas podem levar a pequenas distorções na avaliação dos modelos, já que o conjunto original de anotações não contempla todas as entidades mencionadas nos relatos.

Dessa forma, os resultados apontam que a abordagem híbrida proposta neste estudo possui potencial para atuar como ferramenta de apoio à anotação manual, atuando como um revisor automático capaz de sinalizar inconsistências e recuperar entidades que passaram despercebidas.

6. Considerações Finais

Este estudo apresentou o desenvolvimento e avaliação do ClinptHyb, um algoritmo híbrido para o reconhecimento de entidades nomeadas em relatos de casos clínicos em português.

Os experimentos realizados demonstraram que o ClinptHyb superou os modelos de referência introduzidos por [Schneider et al. 2020] em 8 das 10 categorias analisadas, estabelecendo novos padrões de desempenho para essas entidades no corpus SemClinBr. Observou-se ainda que o modelo spaCy isolado obteve melhor desempenho em algumas categorias quando comparados às variações do BERT, o que reforça sua relevância como uma base consistente para o desenvolvimento do modelo híbrido.

Destaca-se, em particular, o desempenho alcançado na categoria de negação (*Negation*), que atingiu *F1-score* de 97%. Esse resultado é especialmente relevante para o domínio clínico, uma vez que a identificação correta de negações é essencial para evitar interpretações equivocadas sobre o estado de saúde do paciente. Estudos na área de extração de informação médica indicam que falhas nesse processo podem comprometer diagnósticos, gerar custos indevidos e até implicações legais [Fabregat et al. 2023], o que reforça a relevância prática da solução proposta.

Além das métricas apresentadas, o ClinptHyb evidenciou um potencial adicional ao identificar inconsistências no próprio corpus SemClinBr, incluindo variações de limites de entidades e omissões pontuais de anotações. Esse achado indica a que o modelo pode atuanão apenas como ferramenta de extração automática, mas também como recurso de apoio à anotação manual e à curadoria de *datasets* clínicos em português, contribuindo para a melhoria contínua da qualidade da informação.

Sob a perspectiva da OIPT, os resultados indicam que o ClinptHyb contribui para ampliar a capacidade de processamento da informação clínica em contextos caracterizados por narrativas textuais não estruturadas. O modelo combina o uso do spaCy para o processamento padronizado das entidades com a atuação de um LLM como revisor semântico. Essa integração mostrou-se eficaz na redução de incertezas informacionais

e contribuiu para a melhoria da qualidade, consistência e completude das anotações clínicas.

Embora os resultados sejam promissores, ainda há espaço para , ainda há espaço para aprimora, especialmente em categorias com desempenho inferior, como *Disorders* e *Laboratory or Test Result*. Esses casos evidenciam a necessidade de refinar o modelo e investigar estratégias complementares de treinamento e validação. Nesse sentido, trabalhos futuros tornam-se fundamentais para ampliar o alcance e a aplicação prática do ClinptHyb, consolidando sua contribuição para a área de Sistemas de Informação em saúde.

7. Trabalhos Futuros

Como trabalho futuro, propõe-se a aplicação do ClinptHyb a todo o corpus SemClinBr, com o objetivo de identificar inconsistências e omissões de forma sistemática, possibilitando a criação de uma versão revisada do *dataset*. Esse processo também permite a ampliação do conjunto de entidades contempladas pelo modelo, abrangendo todas as categorias disponíveis no corpus e, consequentemente, tornando as anotações mais completas.

Outra linha de investigação consiste na integração de técnicas de Recuperação de Conhecimento (*Retrieval-Augmented Generation - RAG* [9]), permitindo a utilização de fontes complementares, como bulas de medicamentos e prontuários eletrônicos. Esse avanço potencializaria a capacidade do modelo de produzir anotações mais assertivas e contextualizadas.

Além disso, considera-se ampliar o escopo do ClinptHyb para além do reconhecimento de entidades nomeadas, incorporando a extração de relações. Esse recurso possibilitaria a identificação de interações medicamentosas, correlações entre sintomas e diagnósticos e outras conexões relevantes, enriquecendo a análise clínica.

Por fim, considera-se relevante o desenvolvimento de uma ferramenta de anotação com interface web, na qual o modelo poderia ser utilizado tanto para a pré-anotação automática quanto para a revisão das anotações humanas. Dessa forma, seria possível otimizar fluxos de trabalho e contribuir para uma maior consistência nos dados anotados.

Conclui-se que este trabalho representa uma contribuição relevante para o avanço do reconhecimento de entidades nomeadas em textos clínicos em português. A abordagem híbrida proposta, além de proporcionar ganhos expressivos em desempenho, abriu possibilidades para novas pesquisas e aplicações, consolidando um caminho promissor para o desenvolvimento de soluções em NLP voltadas ao domínio da saúde.

Referências

Caseli, H. d. M., Nunes, M. d. G. V., and Pagano, A. (2024). O que é pln? In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 1. BPLN, 2 edition.

[9]Técnica que combina modelos de linguagem com sistemas de recuperação de informações para acessar bases externas e enriquecer a análise de textos.

Claro, D. B., Santos, J., Souza, M., Vieira, R., and Pinheiro, V. (2024). Extração de informação. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 20. BPLN, 2 edition.

Fabregat, H., Duque, A., Martinez-Romo, J., and Araujo, L. (2023). Negation-based transfer learning for improving biomedical named entity recognition and relation extraction. *Journal of Biomedical Informatics*, 138.

Groq (2025). Groq documentation: Overview. `https://console.groq.com/docs/overview`. Acesso em: 2025-07-16.

Jurafsky, D. and Martin, J. H. (2024). Large language models. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, chapter 7. Prentice Hall PTR, USA, 3rd edition. Draft, August 24, 2025.

Kmetz, J. L. (2018). *The Information Processing Theory of Organization: Managing Technology Accession in Complex Systems*. Routledge.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mello, C. E. R., Schneider, E. T. R., Oliveira, L. E. S. e., Nascimento, J. N. d., Gumie, Y. B., Araújo, I. F. d., and Moro, C. (2024). Avaliação de grandes modelos de linguagem na extração de informações clínica. *Journal of Health Informatics*, 16(Especial):1–14. Apresentado no XX Congresso Brasileiro de Informática em Saúde (CBIS'24), Belo Horizonte, MG, Brasil, 8-11 out. 2024.

Nadeau, D. (2007). *Semi-supervised named entity recognition: learning to recognize 100 entity types with little supervision*. PhD thesis, University of Ottawa, [s.l.].

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., and Zweigenbaum, P. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):1–13.

Oliveira, L. E. S., Peters, A. C., Pucca da Silva, A. M., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(13).

Ollama (2025). Ollama documentation. `https://docs.ollama.com/`. Acesso em: 2025-07-16.

Pagano, A., Moro, C., Schneider, E. T. R., Cintho, L. M. M., and Gumiel, Y. (2024). Pln na saúde. In Caseli, H. M. and Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 25. BPLN, 2 edition.

Pavanelli, L., Gumiel, Y. B., Ferreira, T., Pagano, A., and Laber, E. (2023). Bete: A brazilian portuguese dataset for named entity recognition and relation extraction in the diabetes healthcare domain. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent*

Systems. BRACIS 2023. Lecture Notes in Computer Science, volume 14197 of *Lecture Notes in Computer Science*, pages 256—-267. Springer, Cham.

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. M. (2020). BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Sechidis, K., Tsoumakas, G., Vlahavas, I., and updated by Trent-B (2024). iterative-stratification: Iterative stratification algorithms for multi-label data. `https://github.com/trent-b/iterative-stratification`. Acesso em: 22 set. 2025.

spaCy (2025). spacy 101: Everything you need to know · spacy usage documentation. `https://spacy.io/usage/spacy-101`. Acesso em: 2025-07-16.

Torres, A. M. N. (2024). Desenvolvimento de modelo ner para a extração de informações em relatos de casos clínicos. Orientador: Cristiano da Silveira Colombo.

U.S. National Library of Medicine (2025). Umls semantic network. `https://uts.nlm.nih.gov/uts/umls/semantic-network/root`. Acesso em: 2025-09-18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 6000–6010, Long Beach, CA, USA. Curran Associates Inc.

Wulff, P., Kubsch, M., and Krist, C. (2025). Natural language processing and large language models. In Wulff, P., Kubsch, M., and Krist, C., editors, *Applying Machine Learning in Science Education Research: When, How, and Why?*, Springer Texts in Education, pages 117–142. Springer.