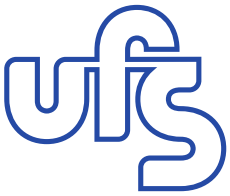

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Análise Exploratória e Prática da Aplicação de NER em Notícias sobre Saúde

Dissertação de Mestrado

Samuel Santana de Almeida

São Cristóvão – Sergipe

2026

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Samuel Santana de Almeida

Análise Exploratória e Prática da Aplicação de NER em Notícias sobre Saúde

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Dr. Methanias Colaço Júnior
Coorientador(a): M.e. Raphael Silva Fontes

São Cristóvão – Sergipe

2026

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

"A tecnologia é apenas uma ferramenta. No que se refere a motivar as crianças e ajudá-las a trabalhar juntas, o professor é o recurso mais importante."(Bill Gates)

Agradecimentos

A trajetória que culmina nesta dissertação foi percorrida ao lado de pessoas fundamentais, sem as quais eu não teria alcançado a maturidade pessoal e profissional que hoje possuo. Guardo com carinho a memória dos meus avós, Rosa Maria e Geraldo Santana, embora não estejam mais presentes, seus legados ecoam na responsabilidade de me tornar o primeiro mestre da minha família. Esta conquista remete às minhas lembranças de infância, época em que o fascínio pela computação já se manifestava como uma vocação para a vida toda.

Meus sinceros agradecimentos ao meu orientador, Methanias Colaço Júnior, que com sabedoria e bom humor soube guiar este trabalho em meio aos desafios de conciliar a minha vida acadêmica e profissional. Sua condução foi fundamental para que o processo de escrita fluísse com leveza e qualidade. Estendo minha gratidão ao coorientador Raphael Silva Fontes e a toda a equipe do grupo de pesquisa, em especial aos colegas João Alysson dos Santos Guimarães e Helder Prado Santos, pela colaboração essencial.

No âmbito profissional, agradeço ao SergipeTec, instituição que permitiu a adequação dos meus horários, viabilizando esta formação. Deixo meu agradecimento especial ao meu chefe de setor, Vitor Vaz, pelo apoio institucional, e à Professora Rita de Cassia, cuja orientação inicial me mostrou o poder transformador da educação. Jamais esquecerei o acolhimento recebido em meu primeiro dia como docente, transformando o receio inicial em propósito.

Agradeço à minha família, em especial à minha mãe e à minha irmã, pelo suporte constante. Por fim, dedico um agradecimento especial à minha noiva, Lorena Maciel, companheira incansável que compartilhou comigo as angústias e alegrias destes quase três anos de jornada.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

A447a

Almeida, Samuel Santana de
Análise exploratória e prática da aplicação de NER em notícias sobre saúde / Samuel Santana de Almeida ; orientador Methanias Colaço Rodrigues Júnior - São Cristóvão, 2026.
67 f.; il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2026.

1. Processamento de linguagem natural (Computação). 2. Auditoria interna. 3. Sistema Único de Saúde (Brasil). 4. Redes neurais (Computação). I. Rodrigues Júnior, Methanias Colaço orient. II. Título.

CDU 004:61

Resumo

Contexto: O setor público, especificamente no âmbito das auditorias do Ministério da Saúde e do Sistema Único de Saúde (SUS), enfrenta gargalos operacionais decorrentes de processos manuais de análise de dados. Essa ineficiência resulta em morosidade e custos elevados, comprometendo o combate à corrupção e a garantia do direito universal à saúde. **Objetivos:** Este estudo buscou caracterizar o estado da arte das arquiteturas de Named Entity Recognition (NER) aplicadas à saúde e identificar a abordagem mais eficaz para a extração de entidades e classificação de textos em notícias do setor. O foco recai sobre a otimização da auditoria do SUS, comparando o desempenho dos modelos BERT, BERT-CRF e ModBERTBr. **Metodologia:** A pesquisa empregou um Mapeamento Sistemático da Literatura (MSL), com a análise de 310 estudos de um universo inicial de 5.863, seguido por um experimento controlado. O experimento utilizou um pipeline de Processamento de Linguagem Natural (PLN) aplicado a um corpus de 800 notícias de saúde para o treinamento e avaliação das tarefas de NER e classificação. **Resultados:** O MSL revelou a hegemonia de modelos de Deep Learning baseados em Transformers, sendo o BERT a técnica mais frequente (215 estudos). No experimento prático de NER, o BERT-CRF destacou-se com os melhores índices de recall (0,880), precisão (0,855) e F1-score (0,860), enquanto o BERT obteve a maior acurácia (0,900). Na tarefa de classificação, o BERT superou o ModBERTBr em todas as métricas. Quanto à eficiência, o BERT foi superior em tempo de execução para NER (8min 10s), ao passo que o BERT-CRF foi mais ágil na classificação (7min 10s). **Conclusão:** A eficácia do modelo é contingente à tarefa: o BERT-CRF é superior para a detecção precisa em sequências complexas (como relatórios de auditoria), enquanto o BERT é mais indicado para a triagem célere de grandes volumes documentais. Conclui-se que a implementação de um sistema híbrido possui elevado potencial para otimizar a seleção de conteúdos auditáveis no SUS, fortalecendo a integridade e a investigação nos processos públicos.

Palavras-chave: Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Auditoria do SUS, BERT, Experimentação.

Abstract

Context: Context: The public sector, specifically regarding audits within the Ministry of Health and the Unified Health System (SUS), faces operational bottlenecks due to manual data analysis processes. This inefficiency leads to delays and high costs, hindering the fight against corruption and the assurance of the universal right to health. **Objectives:** This study aimed to characterize the state-of-the-art in Named Entity Recognition (NER) architectures applied to healthcare and identify the most effective approach for entity extraction and text classification in health news. The focus is on optimizing SUS auditing by comparing the performance of BERT, BERT-CRF, and ModBERTBr models. **Methodology:** The research employed a Systematic Literature Mapping (SLM), analyzing 310 studies from an initial pool of 5,863, followed by a controlled experiment. The experiment utilized a Natural Language Processing (NLP) pipeline applied to a corpus of 800 health news articles for the training and evaluation of NER and classification tasks. **Results:** The SLM revealed the dominance of Transformer-based Deep Learning models, with BERT being the most frequent technique (215 studies). In the practical NER experiment, BERT-CRF excelled with the highest recall (0.880), precision (0.855), and F1-score (0.860), while BERT achieved the highest accuracy (0.900). In the classification task, BERT outperformed ModBERTBr across all metrics. Regarding efficiency, BERT was superior in execution time for NER (8min 10s), whereas BERT-CRF was faster in classification (7min 10s). **Conclusion:** Model effectiveness is task-dependent: BERT-CRF is superior for precise detection in complex sequences (such as audit reports), while BERT is better suited for the rapid screening of large document volumes. It is concluded that the implementation of a hybrid system has high potential to optimize the selection of auditable content in SUS, strengthening integrity and investigation within public processes.

Keywords: Natural Language Processing, Named Entity Recognition, SUS Auditing, BERT, Experimentation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
PROCC	Pós-Graduação em Ciência da Computação
UFS	Universidade Federal de Sergipe
LAIS	Laboratório de Inovação Tecnológica em Saúde
ACM	Association for Computing Machinery
AI	Inteligência Artificial
API	Application Programming Interface
BI	Business Intelligence
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CSV	Comma-Separated Values
DATASUS	Departamento de Informática do Sistema Único de Saúde
biLSTM	bidirectional Long Short-Term Memory
GB	Gigabyte
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
I/O	Input/Output
ER	Extração de Relações
IEEE	Institute of Electrical and Electronics Engineers
JSON	JavaScript Object Notation
LLM	Large Language Model
ML	Aprendizado de Máquina
REN	Reconhecimento de Entidades Nomeadas
MSL	Mapeamento Sistemático da Literatura
NAVI	Núcleo Avançado de Inovação Tecnológica

CRF	Conditional Random Fields
ELMo	Embeddings from Language Model
PICO	População, Intervenção, Controle e Resultado
PLN	Processamento de Linguagem Natural
RAM	Random Access Memory
CoNLL	Computational Natural Language Learning
CNN	Convolutional Neural Networks
TF-IDF	Term Frequency-Inverse Document Frequency
RNN	Recurrent Neural Networks
XAI	IA Explicável
SNA	Sistema Nacional de Auditoria

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: Samuel Santana de Almeida

Em 28 dias do mês de janeiro do ano de dois mil e vinte seis, com início às 10:00hs, realizou-se na Sala de Seminários do PROCC da Universidade Federal de Sergipe, na Cidade Universitária Prof. José Aloísio de Campos, a Sessão Pública de Defesa de Dissertação de Mestrado do candidato **Samuel Santana de Almeida** que desenvolveu o trabalho intitulado: **" Análise Exploratória e Prática da Aplicação de NER em Notícias sobre Saúde",** sob a orientação do Prof. Dr. **Methanias Colaço Rodrigues Júnior.** A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, o Dr. **André Luis Meneses Silva (UFS)** como membro externo ao programa, em seguida,Dr**. Juciano de Sousa Lacerda (UFRN)** como membro externo à instituição e**,** em seguida, Dr. **Leonardo Nogueira Matos** (PROCC/UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando (a) aprovado "*(aprovado/reprovado)"*. Atendidas as exigências da Instrução Normativa 05/2019/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), e da Resolução nº 04/2021/CONEPE que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos",* 28 de janeiro de 2026.

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Prof. Dr. Leonardo Nogueira Matos
(PROCC/UFS)
Examinador Interno

Prof. Dr. André Luis Meneses Silva
(UFS)
Examinador Externo ao programa

Prof. Dr. Juciano de Sousa Lacerda
(UFRN)
Examinador Externo ao programa

Samuel Santana de Almeida
Candidato

Sumário

1 Introdução

Este capítulo introdutório estabelece as bases conceituais e estruturais desta dissertação. Seu propósito é apresentar o contexto da aplicação do NER no setor de saúde e delinear a problemática central que fundamenta a pesquisa: a persistência de processos manuais, ineficientes e de alto custo na coleta e análise de dados para as auditorias do Sistema Único de Saúde (SUS). A partir dessa delimitação, o capítulo detalha a justificativa para a proposição de um modelo experimental e as contribuições esperadas para o setor público, com ênfase nas auditorias do SUS. Em seguida, são definidas as questões de pesquisa que guiaram a investigação e os objetivos que foram traçados. Por fim, é descrita a metodologia de pesquisa construtiva adotada para o desenvolvimento e a demonstração do artefato tecnológico proposto.

1.1 Contextualização

No Brasil, o SUS foi instituído em 1988 com o objetivo de garantir acesso universal e equitativo à assistência médica (PENSESUS - Portal de Informações de Saúde Pública da Fiocruz, s.d.a; PENSESUS - Portal de Informações de Saúde Pública da Fiocruz, s.d.b). Apesar de sua relevância, o SUS enfrenta desafios significativos, como subfinanciamento, desigualdades regionais e a necessidade de uma abordagem mais integrada que contemple a prevenção e a promoção da saúde (BRITO-SILVA; BEZERRA; TANAKA, 2012). Para assegurar a qualidade dos serviços e a utilização eficiente dos recursos, a atuação de órgãos de controle e fiscalização, como o Departamento Nacional de Auditoria do SUS (DENASUS), é crucial (Brasil, 1993; Brasil, 2023).

A fase analítica da auditoria, voltada ao planejamento estratégico, fundamenta-se na coleta de informações de diversas fontes, incluindo a mídia, para subsidiar a identificação de possíveis irregularidades. No entanto, o processo atual do DENASUS, considerando o alto volume de solicitações e a limitação das equipes, resulta em um tempo excessivo para a conclusão

das auditorias (AQUINO, 2022). Corroborando essa perspectiva, um estudo (COSTA et al., 2021) desenvolvido por pesquisadores em Natal, Rio Grande do Norte, evidencia que a eficácia operacional do Sistema Nacional de Auditoria (SNA) enfrenta sérios obstáculos estruturais que comprometem a agilidade processual. O artigo destaca o papel central do Componente Federal (DENASUS/SNA), que além de demandar o maior volume de ações representando 33,12% do total no estado, exerce funções essenciais de organização e orientação de toda a estrutura sistêmica.

Nesse cenário, observa-se que o sistema padece de um número insuficiente de auditores, com uma distribuição geográfica desigual e concentrada na capital potiguar, o que limita a capacidade de resposta frente ao volume crescente de demandas. A análise detalhada dos métodos indica que o tempo de execução varia drasticamente conforme a complexidade do serviço auditado: enquanto auditorias em serviços de alta complexidade são finalizadas em média em 25 dias, os trabalhos na Atenção Básica demandam cerca de 127 dias.

Essa disparidade é explicada pelo fato de que as verificações na atenção primária envolvem análises mais amplas de repasses financeiros e redes municipais, exigindo maior tempo de execução intrínseco nas fases operativa e analítica. Tal ineficiência é ainda agravada pela ausência de componentes de auditoria implantados na maioria dos municípios do Rio Grande do Norte, o que sobrecarrega as unidades centrais e dificulta a gestão oportuna dos recursos do SUS.

O Processamento de Linguagem Natural (PLN), um dos campos da Inteligência Artificial (IA), permite que sistemas computacionais compreendam, interpretem e gerem textos de forma análoga à humana. Uma tarefa fundamental no PLN é o Named Entity Recognition (NER), que identifica e classifica entidades como nomes de pessoas, organizações, condições médicas e medicamentos em textos (LI; XU; YUAN, 2020). Essa capacidade aprimora a recuperação de informações, a medicina personalizada e os sistemas de apoio à decisão clínica.

Figura 1 – Named Entity Recognition (NER) , adaptado de (Analytics Vidhya,).

A Figura 1 ilustra o NER em textos de notícias utilizando a linguagem de programação

Python e o recurso NER do spaCy. O spaCy NER é uma funcionalidade essencial da biblioteca spaCy, especializada em processamento de linguagem natural (Analytics Vidhya,).

1.2 Análise do Problema

Métodos tradicionais de NER, como sistemas baseados em regras, word embeddings, por exemplo, Word2Vec, GloVe (KULSHRETHA; LODHA, 2023) e modelos de marcação de sequência por exemplo, CRFs, HMMs, (XING; GAO; CHEN, 2014; FENG; MANMATHA; MCCALLUM, 2006), têm dificuldades em capturar a complexidade e as nuances dos textos médicos, levando a baixa precisão e inflexibilidade. Além disso, esses métodos exigem grandes conjuntos de dados rotulados, que são difíceis de obter.

Foi formulada a seguinte questão: Qual modelo (BERT-CRF, BERT ou ModBERTBr) apresenta melhor desempenho para NER em notícias de saúde voltadas à auditoria do SUS, e qual sua eficácia para classificação de texto neste contexto?

1.3 Justificativa

Recentemente, modelos de linguagem avançados, particularmente os modelos baseados em transformadores como o Bidirectional Encoder Representations from Transformers (BERT) Figura 3, demonstraram alta performance em tarefas de NER (KIM et al., 2020). De acordo com (VASWANI et al., 2017), a arquitetura Transformer foi introduzida como um modelo neural baseado em mecanismos de atenção, eliminando a necessidade de camadas recorrentes ou convolucionais presentes em arquiteturas anteriores. Conhecido por sua eficiência no processamento de sequências de dados, como texto, o Transformer tem sido amplamente adotado em diversas tarefas de PLN. O funcionamento desta arquitetura pode ser observado na Figura 2.

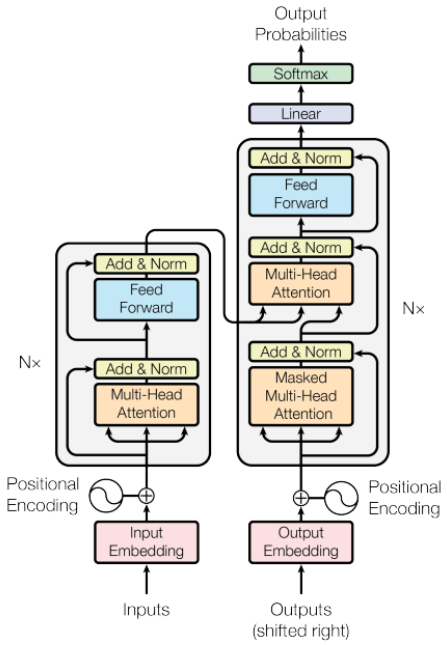

Figura 2 – Arquitetura do modelo Transformer, adaptada de (VASWANI et al., 2017).

A capacidade desses modelos de capturar dependências semânticas complexas e nuances linguísticas é crucial para o processamento preciso de textos médicos, a integração dessas técnicas com o SUS promete avanços significativos na organização automática de informações médicas, melhoria da vigilância de doenças, otimização da gestão de recursos hospitalares, integração de dados de diferentes sistemas e suporte a decisões baseadas em evidências (MILNE-IVES; COCK; LIM, 2020).

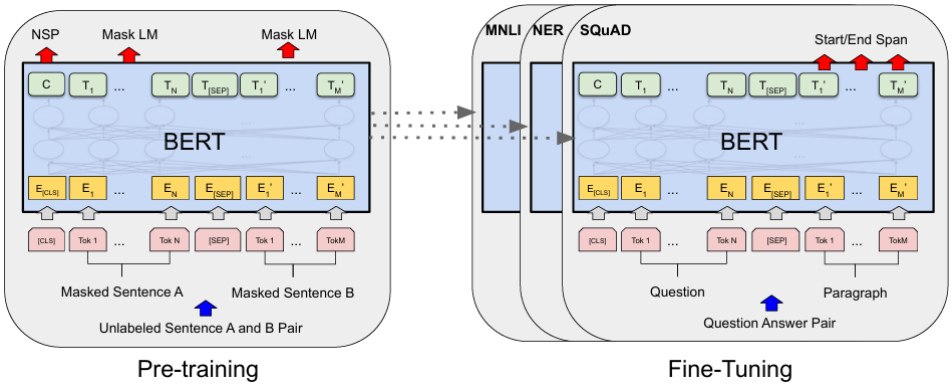

Figura 3 – Procedimentos gerais de pré-treinamento e ajuste fino para o BERT. Além das camadas de saída, as mesmas arquiteturas são usadas tanto no pré-treinamento quanto no ajuste fino. Os mesmos parâmetros do modelo pré-treinado são usados para inicializar modelos para diferentes tarefas posteriores. Durante o ajuste fino,todos os parâmetros são ajustados. [CLS] é um símbolo especial adicionado na frente de cada exemplo de entrada, e [SEP] é um token separador especial (por exemplo, separando perguntas/respostas), adaptada de (DEVLIN et al., 2018).

O BERT é considerado um dos exemplos notáveis de LLMs (Large Language Models),

juntamente com o GPT (Generative Pré-trained Transformer) da OpenAI, como o GPT-3, e modelos desenvolvidos por outras organizações, como o XLNet da Google/CMU, entre outros. LLMs constituem uma classe de modelos de inteligência artificial projetados para entender e gerar texto em larga escala. Treinados em grandes quantidades de dados textuais, esses modelos aprendem padrões na linguagem natural e são capazes de realizar uma variedade de tarefas relacionadas ao processamento de linguagem natural, como tradução, sumarização, geração de texto e classificação de texto (NVIDIA Blog, 2023).

O método estatístico Conditional Random Fields (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) é uma arquitetura discriminativa projetada para a rotulagem de dados sequenciais. Diferente de classificadores independentes, o CRF modela dependências estruturais onde a predição de um rótulo é condicionada ao contexto global da sequência. Para a representação das entidades, adotou-se o esquema de marcação IOB (RAMSHAW; MARCUS, 1995), que segmenta os tokens em: Início (B - Beginning), Continuidade (I - Inside) e Externo (O - Outside). A integração do CRF com o esquema IOB permite a otimização das transições de estado, impondo restrições lógicas que garantem a coerência sintática; por exemplo, o modelo aprende que um rótulo I-ENTITY não pode suceder um rótulo O, exigindo obrigatoriamente um prefixo B ou I da mesma categoria para validar a fronteira da entidade.

Essa integração entre o SUS promete avanços significativos tanto na área da computação quanto na área da saúde, alinhando-se aos princípios fundamentais do SUS, que visam proporcionar acesso universal e equitativo ao sistema de saúde. O uso do NER no SUS permite a organização automática de informações médicas, o aprimoramento da vigilância de doenças, a otimização da gestão de recursos hospitalares, a integração de dados de diferentes sistemas e o suporte a decisões baseadas em evidências. Isso resulta em atendimento mais rápido, políticas públicas mais eficientes e melhor qualidade de atendimento à população. Após a revisão, o objetivo é contribuir para a área da computação com o desenvolvimento de técnicas de inteligência artificial, aplicando a inteligência artificial à extração de dados dentro de grandes conjuntos de informações, favorecendo uma computação mais acessível e inclusiva.

1.4 Objetivos

1.4.1 Objetivo Geral

Projetar, implementar e apresentar evidências da viabilidade de um modelo NER em português, modular, escalável e replicável, com o propósito de gerenciar e otimizar o processo de engenharia de dados e o ciclo de vida completo de modelos de Inteligência Artificial voltados para a auditoria no setor de saúde.

1.4.2 Objetivos Específicos

Para a consecução do objetivo geral, foram traçados os seguintes objetivos específicos:

• Executar um Mapeamento Sistemático da Literatura (MSL) para mapear e sintetizar o estado da arte, identificando e caracterizando as abordagens arquiteturais, os conceitos fundamentais e os softwares open-source mais proeminentes na construção de plataformas de NER em textos de saúde.

• Projetar um modelo NER em português, fundamentado nos achados do MSL e orientada por princípios de modularidade, replicabilidade.

1.5 Metodologia

A metodologia desta pesquisa foi delineada em duas fases complementares, que visaram estabelecer tanto a base teórica quanto a validação empírica das propostas. A fase inicial consistiu na realização de um Mapeamento Sistemático da Literatura (MSL), conduzido com base no rigoroso protocolo de (KITCHENHAM; CHARTERS, 2007). Este MSL, cujos resultados foram submetidos à revista [1],Frontiers in Artificial Intelligence, estabeleceu a fundação teórica e empírica do trabalho. Seu objetivo primário foi explorar sistematicamente a literatura científica para identificar padrões, ferramentas consolidadas e as lacunas existentes na implementação de arquiteturas de NER no setor de saúde. Os achados dessa análise aprofundada estão detalhados no Capítulo 2 e informaram criticamente as decisões de projeto para a fase subsequente.

A segunda fase da pesquisa consistiu em uma abordagem experimental, comparativa e controlada, focada no desenvolvimento e teste de três modelos de NER. Este experimento foi concebido a partir dos padrões identificados no MSL e implementado em ambiente de nuvem (Google Colab), seguindo um pipeline completo de PLN desde a tokenização até a análise estatística.

A condução desses experimentos seguiu rigorosamente as diretrizes práticas para experimentação com IA e os passos propostos por (Colaço Júnior, 2025). Essa metodologia estruturou-se nas etapas de Definição do Objetivo, Planejamento, Operação, Análise e Interpretação de Dados, além da Descrição de Ameaças à Validade.

Dentro dessa estrutura, a etapa de Planejamento englobou a seleção do contexto, definição de variáveis e formulação de hipóteses. Já a fase de Operação desdobrou-se na preparação, execução e validação dos dados. Tais processos, incluindo o detalhamento do subprocesso operacional, estão ilustrados nas Figuras 4, 5 e 6. Os resultados desta fase subsidiaram um segundo artigo científico, submetido à revista Array.

[1] https://www.frontiersin.org/journals/artificial-intelligence

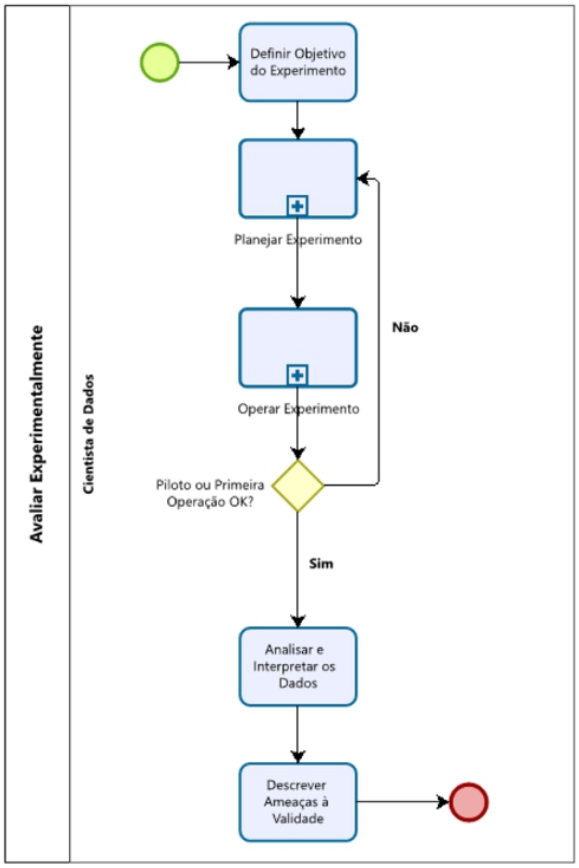

Figura 4 – Atividades do Processo Avaliar Experimentalmente (Colaço Júnior, 2025).

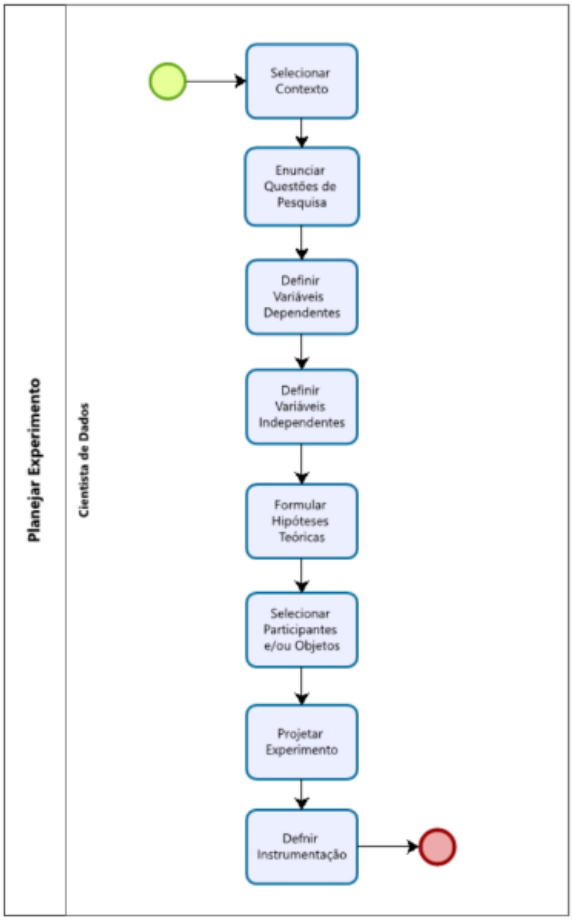

Figura 5 – Atividades do Subprocesso Planejar Experimento (Colaço Júnior, 2025).

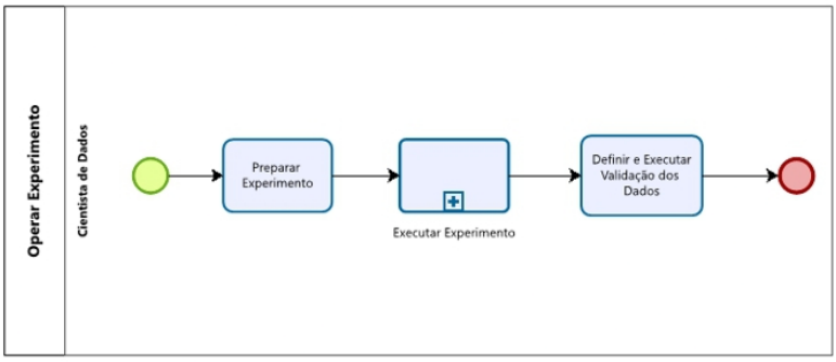

Figura 6 – Atividades do Subrocesso Operar Experimento (Colaço Júnior, 2025).

1.6 Organização da Dissertação

Este documento está organizado de acordo com a Instrução Normativa Nº 05/2019/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com Qualis". São 5 capítulos que fornecem uma base conceitual para o entendimento sistêmico:

- O Capítulo 1 apresenta esta Introdução, contextualizando o problema, a justificativa e os objetivos do trabalho.
- O Capítulo 2 apresenta parte do artigo de Mapeamento Sistemático da Literatura que investiga arquiteturas de NER em textos de notícias de saúde. Este artigo foi publicado na revista Frontiers in Artificial Intelligence.
- O Capítulo 3 apresenta parte do artigo que detalha o experimento realizado. Este artigo foi aceito para publicação na revista Array.
- O Capítulo 4 realiza uma Discussão unificada, conectando os achados do mapeamento sistemático com as decisões do experimento e as lições aprendidas durante sua implementação.
- O Capítulo 5 apresenta a Conclusão geral do trabalho, sintetizando as contribuições, as limitações e apontando os trabalhos futuros, que agora são viabilizados pelo modelo desenvolvido.

2

Artificial Intelligence in Healthcare Text Processing: A Review Applied to Named Entity Recognition

Este capítulo apresenta os resultados de um artigo publicado em periódico científico internacional, cujo objetivo foi mapear sistematicamente o estado da arte das arquiteturas de Reconhecimento de Entidades Nomeadas aplicadas ao domínio da saúde. O artigo foi publicado na revista Frontiers in Artificial Intelligence, e fornece a fundamentação empírica para o projeto e a implementação do artefato tecnológico apresentado no Capítulo 3.

2.1 Planejamento do Mapeamento Sistemático

2.1.1 Objetivo

Este mapeamento teve como objetivo identificar e caracterizar os algoritmos utilizados para construir uma arquitetura de NER que possa ser replicada especificamente para uma auditoria em um sistema de saúde, com ênfase em desempenho e precisão na previsão de entidades.

2.1.2 Questões de Pesquisa

Com o objetivo de analisar e avaliar técnicas de PLN para extração de entidades nomeadas utilizadas na área da saúde, o método de MSL foi inicialmente escolhido. Este processo teve início em 25 de abril de 2024. O Mapeamento Sistemático da Literatura surge como uma ferramenta valiosa quando a demanda não é por respostas aprofundadas a questões específicas, mas sim pela obtenção de uma visão abrangente e holística de um determinado domínio ou área do conhecimento (KITCHENHAM; CHARTERS, 2007).

Ao contrário de abordagens mais específicas, o objetivo principal do mapeamento sistemático é compreender e organizar sistematicamente o corpo de literatura existente. Para isso,

é necessário definir a orientação da pesquisa, a estratégia de busca e os critérios de seleção dos artigos.

A Tabela 1 abaixo ilustra o modelo PICO (População, Intervenção, Controle, Resultados) utilizado neste estudo. Utilizamos a estrutura PICO para formular as questões de pesquisa. A ideia de estruturar as questões clínicas em quatro componentes foi originalmente proposta por (RICHARDSON et al., 1995). Ao formular uma pergunta de pesquisa utilizando o modelo PICO, os pesquisadores podem estruturar suas investigações de forma clara e específica (SANTOS; PIMENTA; NOBRE, 2007). Essa abordagem é útil para delimitar o escopo da pesquisa, identificar variáveis-chave e facilitar a busca por evidências relevantes na literatura.

Tabela 1 – Questões estruturadas no formato PICO

Sigla	Categoria	Descrição
P	População	Publicações que abordam a extração de entidades nomeadas em documentos de saúde.
I	Intervenção	Contexto das técnicas de pré-treinamento de linguagem bidirecional, transformadores ou grandes modelos de linguagem usados na extração de entidades nomeadas em documentos relacionados à saúde.
C	Controle	Abordagens convencionais que não utilizam essas técnicas avançadas de pré-treinamento de linguagem para extração de entidades nomeadas em documentos da área da saúde.
O	Resultados	A eficácia da extração de entidades nomeadas e o nível de automação alcançado na extração de entidades de registros médicos.

Com base na definição de PICO, a revisão foi orientada pelas seguintes questões de pesquisa.

- Q1: Quais são as principais técnicas utilizadas?
- Q2: Quais técnicas específicas apresentam melhor e pior desempenho (Linguagem de base de avaliação e Tipo de Aprendizagem)?
- Q3: Como as publicações relacionadas ao uso de técnicas de pré-treinamento de linguagem bidirecional, Transformers ou LLMs na extração de entidades nomeadas de documentos relacionados à saúde são distribuídas ao longo dos anos?
- Q4: Quais países contribuíram mais significativamente para publicações no contexto de técnicas de pré-treinamento de linguagem bidirecional, Transformers ou grandes modelos de linguagem usados na extração de entidades nomeadas de documentos relacionados à saúde?

2.1.3 Estratégia de Busca e Seleção

Para executar a busca de artigos, foram selecionadas as bases de dados responsáveis pela publicação dos principais periódicos na área de Ciência da Computação. Entre elas, estão SCOPUS, IEEE Xplore, ACM, Web of Science, Springer e ScienceDirect. Queríamos abranger todas as bases de dados que indexam os principais artigos na área de computação, uma área-chave para NER na área da saúde. Para o processo de busca, foram utilizadas ferramentas de filtragem fornecidas por cada base de dados, com foco em título, resumo e palavras-chave. O acesso às bases de dados foi concedido por meio do portal de periódicos da (CAPES, 2021), utilizando assinaturas institucionais, sem restrições de acesso aos artigos.

Iniciamos a lista com a Scopus, uma grande base de dados que contém muitos periódicos importantes nas ciências e é confiável em todos os tipos de áreas. A IEEE Xplore, gerenciada pelo IEEE, é gerenciada pelo IEEE. (O eminente instituto de tecnologia elétrica) e ciências relacionadas contribuem para a base desta pesquisa, ambas as plataformas em grande medida. Outro recurso incrível é a Association for Computing Machinery, sendo a Biblioteca Digital da ACM uma das maiores bibliotecas eletrônicas na área de computação e tecnologia da informação. Com uma ampla variedade de periódicos, conferências e artigos técnicos, a Biblioteca Digital da ACM é muito importante para que os pesquisadores tenham as informações mais recentes e de primeira classe sobre ciência da computação, IA e Desenvolvimento de Software/SD prontamente disponíveis.

A base de dados Web of Science da Clarivate Analytics contribuiu para a pesquisa. Esta plataforma é amplamente reconhecida por seu rigor e qualidade na indexação de periódicos de alta qualidade em diversos domínios. Ela oferece dados abrangentes sobre citações e a influência de publicações científicas na Web of Science por meio de ferramentas analíticas de ponta para ajudar os pesquisadores a avaliar pesquisas e tendências em áreas científicas. A Elsevier, com seu pacote complementar ScienceDirect e seu extenso banco de dados de artigos revisados por pares, a plataforma abrange áreas como saúde, ciências biológicas, ciências físicas e engenharia, e também utiliza seu acesso a conteúdo de alta qualidade para apoiar o compartilhamento de conhecimento científico e incentivar todos os aspectos da atividade acadêmica em todas as áreas.

Por fim, mas não menos importante, a Springer, uma das melhores editoras científicas, também ocupa uma posição significativa no meio acadêmico mundial. Publica em diversas áreas (ciência da computação, inteligência artificial, etc.). As seções de engenharia da Springer contam com diversos artigos, livros e anais de congressos. A Springer é uma das plataformas reconhecidas para o avanço e a disseminação de novas descobertas científicas, oferecendo aos pesquisadores recursos temáticos altamente apropriados e academicamente úteis.

A combinação dos recursos oferecidos pelo Scopus [1], IEEE Xplore [2], ACM Digital

[1] https://www.scopus.com
[2] https://ieeexplore.ieee.org

Library [3], Web of Science [4], Springer [5] e ScienceDirect [6], plataformas cada uma com suas próprias especializações e pontos fortes, é fundamental para construir uma base sólida e confiável para a pesquisa científica, promovendo o progresso contínuo em seus respectivos campos e contribuindo significativamente para o avanço do conhecimento global. Para realizar a busca em bases de dados digitais, foi definida uma sequência de busca composta por termos e sinônimos em inglês relacionados ao reconhecimento de entidades nomeadas em textos relacionados à saúde. Os termos foram identificados com base nas funções definidas no modelo PICO, descritas na Tabela 1. A Tabela 2 apresenta os termos adaptados para utilização ideal da sequência de busca, e os termos posteriormente refinados são mostrados na Tabela 3.

Tabela 2 – Categorias e termos do modelo PICO identificados para a pesquisa bibliográfica.

Categoria	Palavras-chave
População	saúde*, clínica*, médica*
Interveção	LLM*, Large Languague Model*, BERT*, Bidirectional Encoder Representations from Transformers, NER, Named Entity Recognition
Controle	CNN, RNN
Resultados	Métricas de desempenho, avaliação de precisão, avaliação de modelos.

Tabela 3 – String após refinamento.

População	Intervenção	Resultados
Health	BERT	NER
	LLMs	

A razão para não utilizar modelos sequenciais como Redes Neurais Convolucionais e Redes Neurais Recorrentes (RNNs) na tarefa de Reconhecimento de Entidades de Nomes em textos médicos é que CNNs e RNNs são modelos poderosos em tarefas que envolvem foco em sequência, incluindo exemplos de padrões em imagens (RODRIGUES, 2018) e séries temporárias, respectivamente, mas a NER requer um conceito de contextualização e semântica entre palavras profundas. BERT e LLMs foram escolhidos com base em sua capacidade de capturar dependências de longo alcance e representar as nuances da linguagem natural de forma mais profunda, o que é crucial para identificar entidades em textos médicos; BERT, por exemplo, é bidirecional, ou seja, considera o contexto à esquerda e à direita da palavra.

Com base nas considerações acima, a seguinte sequência de caracteres de busca foi desenvolvida, juntamente com sequências de caracteres específicas adaptadas para cada banco de dados:

[3] https://dl.acm.org
[4] https://www.webofscience.com
[5] https://link.springer.com
[6] https://www.sciencedirect.com

STR01 - (llm OR ”large language model” OR BERT OR ”Bidirectional Encoder Representations from Transformer”) AND (NER OR ”Named Entity Recognition”) AND (health OR medical OR clinical) Database-Specific Search Strings Scopus (ABS(llm OR ”Large Language Model” OR BERT OR ”Bidirectional Encoder Representations from Transformer”) AND ABS(NER OR ”Named Entity Recognition”) AND ABS(health OR medical OR clinical))

IEEE Xplore Digital Library (”Abstract”:llm OR ”Abstract”:”Large Language Model” OR ”Abstract”:BERT OR ”Abstract”:”Bidirectional Encoder Representations from Transformer”) AND (”Abstract”:NER OR ”Abstract”:”Named Entity Recognition”) AND (”Abstract”:health OR ”Abstract”:medical OR ”Abstract”:clinical)

ACM Digital Library [[Abstract: llm] OR [Abstract: ”Large Language Model”] OR [Abstract: BERT] OR [Abstract: ”Bidirectional Encoder Representations from Transformer”]] AND [[Abstract: health] OR [Abstract: medical] OR [Abstract: clinical]] AND [[Abstract: NER] OR [Abstract: ”Named Entity Recognition”]]

Web of Science (llm OR ”Large Language Model” OR BERT OR ”Bidirectional Encoder Representations from Transformer”) AND (NER OR ”Named Entity Recognition”) AND (health OR medical OR clinical)

ScienceDirect (LLM OR ”Large Language Model” OR BERT OR ”Bidirectional Encoder Representations from Transformer”) AND (NER OR ”Named Entity Recognition”) AND (health OR medical OR clinical)

Springer (ABS(llm OR ”Large Language Model” OR BERT OR ”Bidirectional Encoder Representations from Transformer”) AND ABS(NER OR ”Named Entity Recognition”) AND ABS(health OR medical OR clinical))

2.1.4 Critérios de Seleção de Fontes

Critérios de inclusão e exclusão são utilizados para garantir que apenas estudos ou dados apropriados sejam apresentados. Serão aceitos estudos com certa relevância/adequação à sua questão de pesquisa e com viés mínimo na seleção de estudos/informações. Critérios de inclusão: definem os critérios específicos de inclusão e dados que um estudo ou conjunto de dados deve atender para inclusão na análise atual. Por outro lado, os critérios de exclusão indicam quais estudos ou dados precisam ser excluídos devido à falta de informações diretas ou à baixa qualidade dos estudos na análise. Abaixo, podemos ver os critérios analisados nos artigos em inglês:

Critérios de Inclusão

- Artigos recentes (a partir de 2019);
- Artigos de periódicos científicos, originais ou de pesquisa.

Critérios de Exclusão

- Artigos duplicados;
- Estudos secundários ou terciários;
- Trabalhos não relacionados ao objeto de estudo;
- Trabalhos que não detalharam experimentos práticos conduzidos para testar suas hipóteses. Critérios de Avaliação da Qualidade
- O estudo visa se especializar em um novo modelo (ajuste fino)?

2.1.5 Estratégia para Extrair Informações

A extração de informações é uma estratégia que formula um pacote muito detalhado de métodos e técnicas para identificar e recuperar partes do texto de um documento ou da entrega de dados não estruturados. Essa estratégia é comum nas áreas de pesquisa de processamento de linguagem natural e mineração de texto, e nosso objetivo é extrair informações valiosas que possam ser direcionadas a análises bastante enriquecedoras.

A etapa inicial crucial no processo de extração de informações é definir quais são os objetivos e as informações-alvo a serem extraídas. Definir o que se pretende obter neste estudo inicia o processo de ter um esboço claro de como as seguintes seções de coleta e relacionamento de dados se encaixam. A coleta pode ser realizada manualmente, por meio de revisão e anonimização de documentos, ou de forma automatizada, utilizando métodos de coleta de dados e APIs capazes de obter quantidades consideráveis de dados de diversas fontes, permitindo aos pesquisadores extrair insights valiosos e identificar padrões relevantes para a pesquisa em questão (JOSEPHSON et al., 2019). A presente pesquisa foi semiautomatizada. A Tabela 4 apresenta o método de extração utilizado

Tabela 4 – Formulário de Extração.

1.	Que tipo de estudo foi realizado?	[Aplicação prática, estudo de caso, prova de conceito, experimento controlado]
2.	Quais são os principais objetivos do artigo?	
3.	Que modelos foram investigados no artigo?	[BERT, LLMs]
4.	O estudo possui alguma avaliação experimental?	[Sim, Não]
5.	Quais foram os principais resultados obtidos no NER?	
6.	Foram declaradas ameaças à validade?	[Sim, Não]

2.2 Condução do Mapeamento Sistemático

Execução do Protocolo

1. Acessar as bases de dados de busca e realizar a busca utilizando as respectivas strings de busca;
2. Aplicar os filtros de inclusão;
3. Os pesquisadores analisam os títulos, resumos, palavras-chave e metodologia, removendo os trabalhos que não atendem aos critérios estabelecidos;
4. Utilizar o método de revisão às cegas, cujo objetivo principal é garantir que a avaliação seja realizada de forma anônima, de forma que os revisores não conheçam a identidade dos autores e vice-versa. A plataforma Rayyan servirá como ferramenta para a execução dessa atividade;
5. Avaliar, entre os pares, se há empate na seleção dos trabalhos, discutindo a inclusão ou exclusão do artigo de acordo com os critérios estabelecidos;
6. Os artigos selecionados serão revisados para coleta de metadados referentes aos critérios de avaliação da qualidade e quaisquer características relevantes.

A seguir, apresentamos a Etapa 1 do protocolo de execução deste mapeamento sistemático, juntamente com os resultados relacionados aos critérios de inclusão e exclusão aplicados.

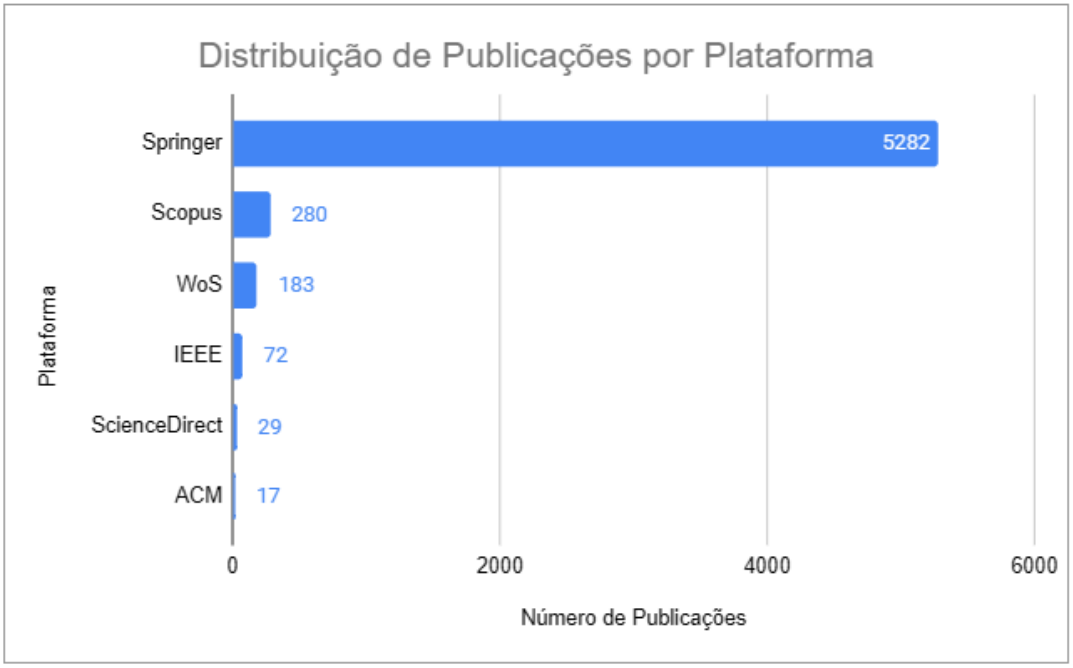

Figura 7 – Primeira etapa do protocolo de execução - Busca de Strings em Bibliotecas Digitais.

Os resultados obtidos, conforme mostrado na Figura 7 acima, foram os seguintes: 72 artigos do IEEE Xplore, 280 do Scopus, 29 do ScienceDirect, 5.282 do Springer, 17 da Biblioteca Digital ACM e 183 da Web of Science. Esses números indicam que a base de dados Springer contribuiu com a maioria dos artigos em relação ao total, com aproximadamente 90,07%, seguida

pela Scopus com 4,77%, Web of Science com 3,12%, IEEE Xplore com 1,23%, ScienceDirect com 0,49% e Biblioteca Digital ACM com 0,29%.

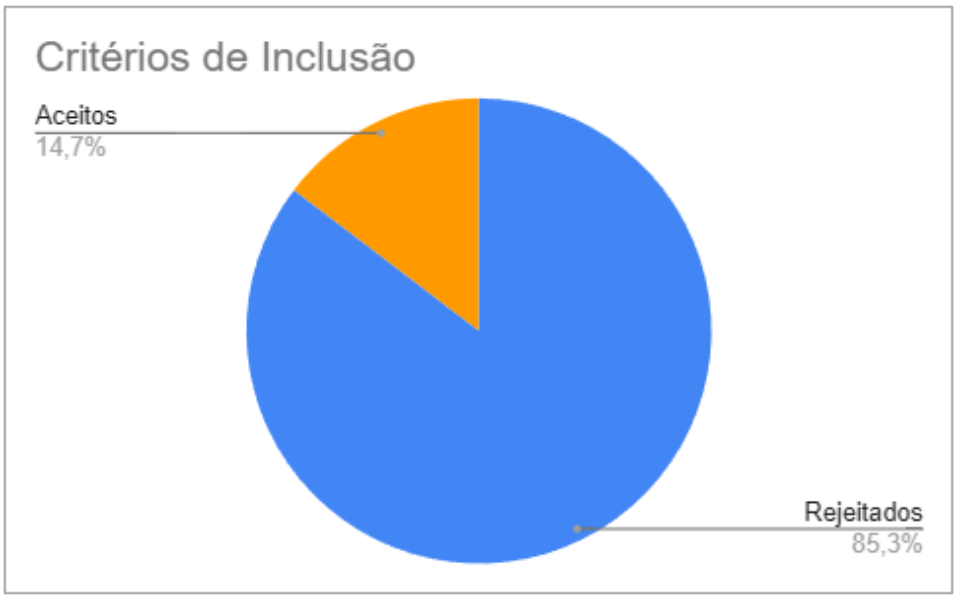

Figura 8 – Primeira Etapa de Seleção de Artigos.

Após a recuperação dos artigos das bases de dados, iniciou-se o processo de filtragem, com base nos critérios de inclusão definidos. Cada artigo foi classificado como Aceito ou Rejeitado. A Figura 8 retrata que das 5863 publicações analisadas, 5004 (85%) não atenderam aos critérios de inclusão. O critério de inclusão 1 significou que restaram 280 artigos da Scopus, 29 da Science Direct, 72 da IEEE e 17 da ACM (nenhum foi removido), enquanto 4615 foram removidos da Springer (restando 667) e 1 artigo da Web of Science (restando 182).

O critério de inclusão 2 significou que foram retornados 104 artigos da Scopus, 25 da Science Direct, 597 da Springer, 2 da IEEE, 130 da Web of Science e 1 da ACM. Após a remoção desses artigos, foi realizada uma leitura superficial dos trabalhos restantes, analisando o título, o resumo e as palavras-chave. Ao final desta etapa, 549 artigos (64% do total) estavam fora do escopo deste mapeamento e foram classificados como Rejeitados. A Figura 9 apresenta um resumo desta etapa. Por fim, para análise detalhada, os artigos restantes foram classificados como aceitos, onde será realizada a etapa de avaliação da qualidade.

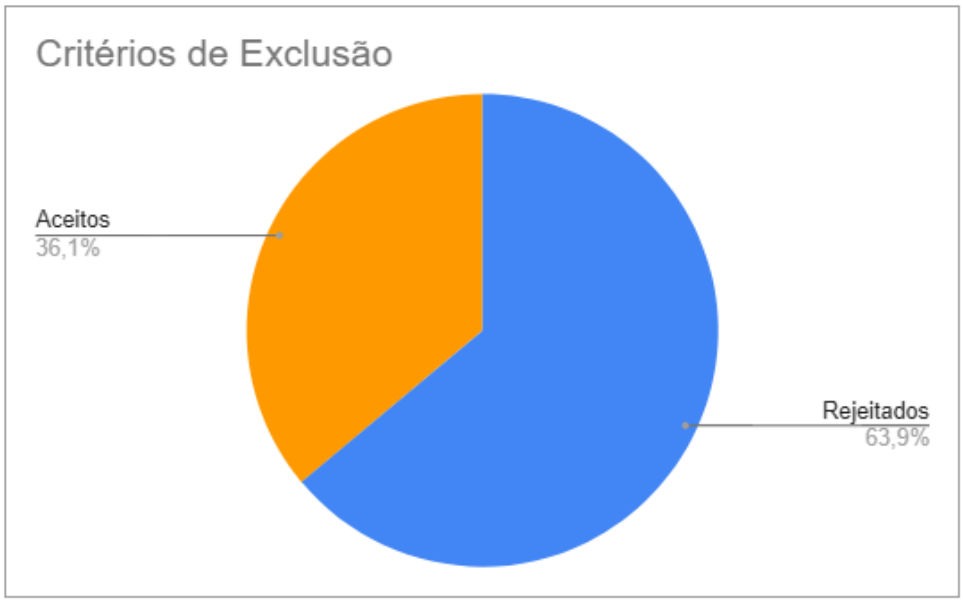

Figura 9 – Segunda Etapa de Seleção de Artigos.

2.3 Síntese e Apresentação de Resultados

Nesta seção, os resultados do mapeamento sistemático são apresentados. A Figura 10 apresenta um fluxograma descrevendo o processo de extração dos artigos obtidos em cada etapa do processo e, ela não apenas organiza sistematicamente esse mapeamento, mas também serve como um recurso essencial para documentar e comunicar claramente os resultados da pesquisa, em seguida, as questões do estudo são respondidas de acordo com os dados extraídos.

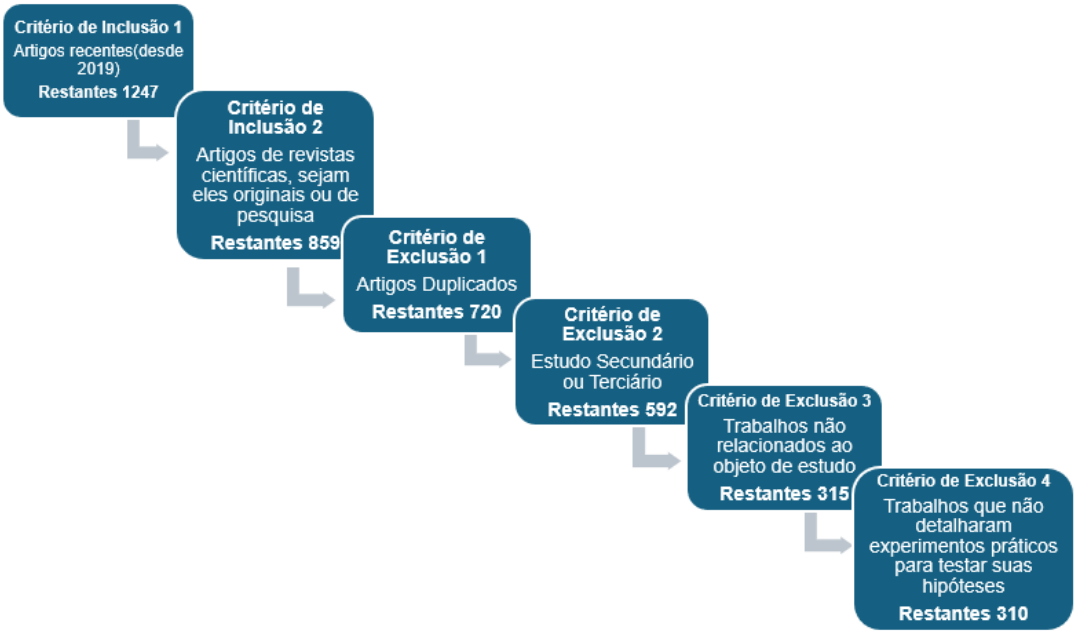

Figura 10 – Processo metodológico (Dados numéricos).

2.3.1 Quais são as principais técnicas usadas

A Figura 11 apresenta as principais técnicas utilizadas, incluindo BERT, com 215 ocorrências devido à sua capacidade de capturar o contexto completo de uma palavra em uma sequência, tornando-a essencial para tarefas de PLN como NER. BiLSTM-CRF, com 60 ocorrências, combina redes LSTM bidirecionais com CRFs para rotulagem de sequências, alavancando o contexto em ambas as direções e capturando dependências entre tags. Outras técnicas importantes incluem BiLSTM (22 ocorrências), usada em análise de sentimento e classificação de texto, e GCN (19 ocorrências), aplicada a dados estruturados, como redes sociais. CNN, com 10 ocorrências, também é usada em PLN, principalmente para classificação de texto. Assim, o foco está em modelos de aprendizado profundo que capturam relacionamentos complexos em dados sequenciais.

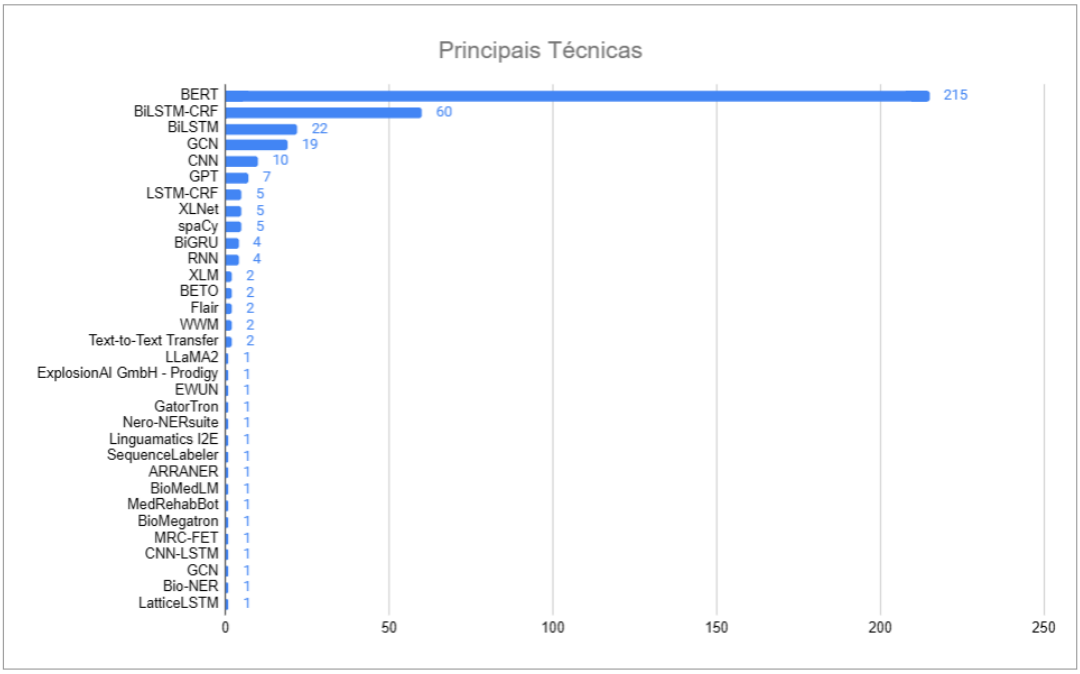

Figura 11 – Técnicas utilizadas em pesquisas ou aplicações.

2.3.2 Quais técnicas específicas apresentam melhor e pior desempenho (avaliação baseada no tipo de linguagem e aprendizagem)?

As técnicas apresentadas e os estudos empíricos como os mais eficazes para NER em textos sobre saúde utilizando IA, vistos na Figura 12, foram identificados na literatura por meio da leitura dos resultados e discussão de cada artigo. Muitos desses resultados estavam em imagens, exigindo uma análise mais cuidadosa. Esses estudos frequentemente também traziam resultados de outras atividades realizadas além da NER, como por exemplo, a análise de sentimentos. A comparação entre os cinco melhores modelos revela desempenhos excepcionais em termos de pontuação F1, recall e precisão, todos acima de 97%. É importante notar que este estudo não

comparou a tarefa NER entre idiomas. É possível que o mesmo modelo LLM possa produzir resultados diferentes de pontuação F1 (Lee et al., 2024), levando em consideração a quantidade, organização e clareza dos dados para um idioma específico.

O BERT lidera com uma pontuação F1 de 99,56%, recall de 99,90% e precisão de 99,85%, sendo o modelo com o desempenho mais equilibrado e consistente em todos os aspectos. Logo em seguida, o MCN-BERT-AdamP também apresenta resultados impressionantes, com uma pontuação F1 de 99,13%, Recall de 99,28% e Precisão de 99,18%. Embora ligeiramente inferior ao BERT, mantém alta Precisão e Recall, demonstrando excelente robustez. Continuando a análise, o TinyBERT, com uma pontuação F1 de 98,91%, Recall de 99,13% e Precisão de 98,70%, oferece desempenho sólido, mas um pouco inferior aos dois modelos mais populares. Ainda assim, continua sendo uma alternativa eficiente, especialmente em cenários com recursos computacionais limitados. O Clinical-BERT atinge uma pontuação F1 de 98,20%, Recall de 98,50% e Precisão de 97,80%. Embora inferior aos três modelos mais populares, continua sendo uma opção eficiente devido ao seu treinamento específico para compreender e processar textos médicos.

Por fim, o ClinicalDistilBERT, com uma pontuação F1 de 97,75%, completa o top 5. Recall e Precision não foram relatados no estudo. Seu desempenho é semelhante ao do Clinical-BERT, sugerindo que ambas as arquiteturas, pré-treinadas especificamente com dados médicos, apresentam resultados próximos ao modelo base do BERT puro, possivelmente até melhores.

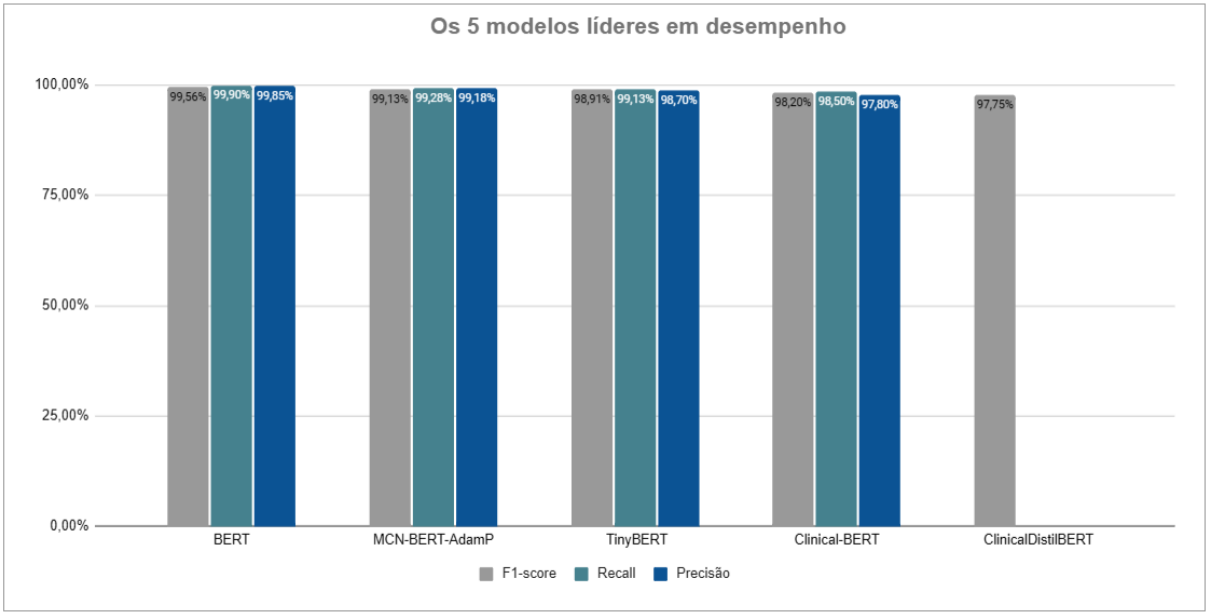

Figura 12 – As 5 melhores técnicas.

Abaixo, na Tabela 5, você pode ver o idioma da base de avaliação e o tipo de aprendizagem dos 10 artigos mais importantes, incluindo os apresentados acima. Você também encontrará o

nome do respectivo artigo e a base de avaliação utilizada.

Classificação	Medida F	Título do Artigo	Base de Avaliação	Idioma da base de Avaliação	Modelo	Tipo de Aprendizagem
1º	99,56%	Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records.	Pathology reports (Korea University Hospital)	English	BERT	Supervised learning and Fine-tuning
2º	99,13%	Optimizing classification of diseases through language model analysis of symptoms	Symptom2Disease and Twitter Drug	English	MCN-BERT and BiLSTM	Supervised learning and Fine-tuning
3º	98,91%	An Efficient Method for Deidentifying Protected Health Information in Chinese Electronic Health Records: Algorithm Development and Validation	EHRs (local hospitals in Chongqing city)	Chinese	TinyBERT	Supervised learning and Fine-tuning
4º	98,20%	A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit	Singapore Civil Defense Force parameedic reports.	English	BERT-base-uncased and Clinical-BERT	Weakly-supervised
5º	97,75%	Lightweight transformers for clinical natural language processing	Public pools (MedNLI and i2b2) and an internal pool (ICN)	English	BioDistilBERT, ClinicalDistilBERT and others based on BioClinicalBERT	Supervised learning and Fine-tuning
6º	96,96%	Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models	EHRs - eHOP CDW Medical Records	French	mBERT, CamemBERT, FlauBERT and Flair + Bi-LSTM-CRF and FastText	Supervised learning
7º	96,80%	A Chinese NER Model Based on BERT with Multi Knowledge Graph Fusion and Embedding	MSRA-NER and Medical-NER	Chinese	BERT	Supervised learning and Fine-tuning
8º	96,29%	Research on Named Entity Identification of Tibetan Medical Ancient Books Based on Hybrid Deep Learning	The Four Medical Tantras	Chinese	ALBERT and BiLSTM-CRF	Few-shot learning
9º	96,27%	An offline English optical character recognition and NER using LSTM and adaptive neuro-fuzzy inference system	EHRs	English	ANFIS-BERT-CRF	Supervised learning and Fine-tuning
10º	96,27%	A large language model for electronic health records trained from scratch	i2b2 (2010, 2012), n2c2 (2018, 2019), MedNLI and emrQA	English	scaled BERT trained from scratch	Self-supervised pre-training and Supervised fine-tuning

Tabela 5 – Top 10 melhores

Após analisar os modelos com melhor desempenho, é igualmente importante considerar os modelos com os piores resultados. A Figura 13 abaixo mostra um resumo dos cinco modelos que obtiveram os piores resultados em cada indicador (F1-Score, Recall e Precision). Esses modelos estão ordenados em ordem decrescente: o F1-Score foi de 64%, a Precision foi de 63,73% e o Recall foi de 66,25%; quanto ao modelo menos ruim (BERT), o modelo LatticeLSTM obteve um F1-Score de 56%, uma Precision de 57% e um Recall de 55%. Outros modelos também foram utilizados neste estudo, a saber, o BERT pré-treinado e o BiLSTM-CRF. O modelo chinês ROBERTa-CRF obteve um F1-score de 55,45%, enquanto o valor do Recall foi menor (55%) e o da Precision (58%). Logo atrás veio o BioBERT, que obteve uma pontuação F1 de 41,30% e, nas outras duas métricas, valores de 58,10% para Precisão e 32,10% para Recall. Por fim, o modelo MED obteve uma pontuação F1 de 21,70%, o pior desempenho entre todos os modelos identificados. Precisão e Recall não são reportados.

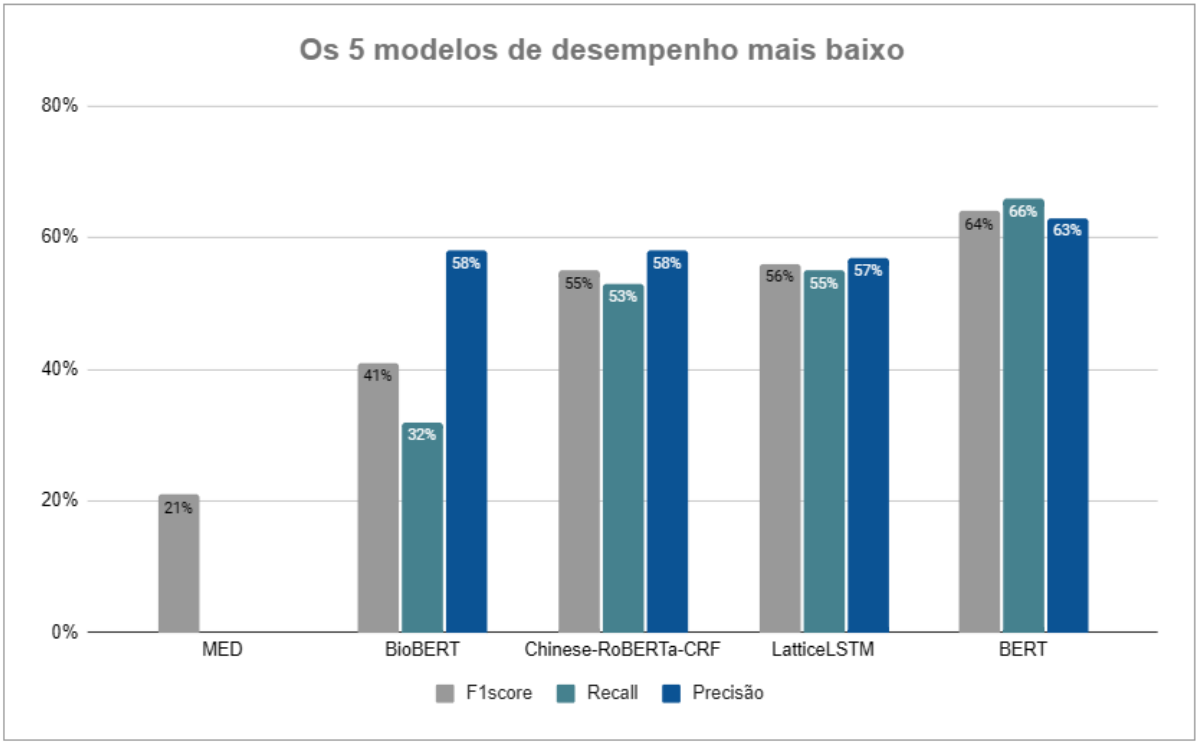

Figura 13 – As 5 piores técnicas.

Abaixo, na Tabela 6, você pode ver o idioma da base de avaliação e o tipo de aprendizagem dos 10 artigos com os menores retornos métricos, incluindo os mostrados acima. Você também encontrará o nome do respectivo artigo e a base de avaliação utilizada.

Classificação	Medida F	Título do Artigo	Base de Avaliação	Idioma da base de Avaliação	Modelo	Tipo de Aprendizagem
1º	21.70%	Knowledge grounded medical dialogue generation using augmented graphs	MedDialog(EN) and Covid Dataset - UMLS	English	MED - BioBERT	Supervised learning (fine-tuning)
2º	41.30%	Large-scale protein-protein post-translational modification extraction with distant supervision and confidence calibrated BioBERT	IntAct database and PubMed abstracts	English	BioBERT	Distant supervision
3º	55.45%	Subsequence and distant supervision based active learning for relation extraction of Chinese medical texts	CMeIE (Chinese Medical Information Extraction)	Chinese	Chinese-RoBERTa-CRF	Active learning
4º	56%	A Chinese telemedicine-dialogue dataset annotated for named entities	haodf.com - Chinese Telemedicine Platform, IMCS-NER and MedDialog-CN	Chinese	BiLSTM-CRF, BERT and LatticeLSTM	Traditional supervised
5º	64.97%	A Unified Knowledge Extraction Method Based on BERT and Handshaking Tagging Scheme	CMeEE	Chinese	BERT	Supervised learning (fine-tuned)
6º	68.01%	We are not ready yet: limitations of state-of-the-art disease named entity recognizers	NCBI and BC5CDR	English	BioBERT	Transfer learning (fine-tuning)
7º	70%	An evaluation of GPT models for phenotype concept recognition	HPO-GS(Human Phenotype Ontology) and BIOC-GS	English	GPT-3.5-turbo and GPT-4.0	Zero-shot/few-shot learning through in-context learning
8º	76%	Machine Reading Comprehension Model in Domain-Transfer Task	NEREL and NEREL-BIO	Russian	RuBERT	Few-shot/zero-shot learning and transfer learning
9º	79%	Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations	Abstracts of scientific articles on RCTs	English	BioBERT, SciBERT and RoBERTa	Few-shot (fine-tuned)
10º	91.80%	Survey of transformers and towards ensemble learning using transformers for natural language processing	Tweets, SQuAD 1.1, CNN/Daily Mail, Disaster, Groningen Meaning Bank	English	(BERT, XLNet, RoBERTa, GPT-2 and ALBERT)	Supervised learning (fine-tuned)

Tabela 6 – Top 10 piores

Analisando as tabelas acima, é possível perceber a predominância do modelo BERT e suas variantes nos melhores resultados, a língua inglesa foi a mais utilizada em termos de bases de testes e avaliações, o tipo de aprendizagem mais recorrente é a aprendizagem supervisionada com ajuste fino, o ajuste fino é o ajuste do modelo pré-treinado à tarefa específica, que no caso do nosso estudo é o NER em textos da área da saúde, podemos observar também na tabela de piores modelos avaliados que são utilizados o zero-shot e o few-shot, outros tipos de aprendizagem também são mais diversos, então podemos concluir que utilizar o BERT com aprendizagem supervisionada e ajuste fino é considerado uma opção que certamente trará bons resultados, levando em consideração também o conjunto de dados que será utilizado.

2.3.3 Em quais anos foram publicados a maioria dos artigos sobre o uso de pré-treinamento bidirecional, transformadores ou modelos de linguagem ampla para extração de entidades nomeadas em documentos de saúde?

A Figura 14 mostra a distribuição dos estudos selecionados por ano de publicação. Observa-se que a maioria dos estudos foi publicada em 2023. O BERT foi introduzido em 2019 e, desde então, tem havido um interesse crescente no uso de LLMs para tarefas de PNL. A pesquisa científica geralmente leva tempo, e pode haver um atraso entre a coleta de dados, a análise e a publicação dos resultados. Estudos iniciados em 2019 podem ter levado até 2023 para serem concluídos e publicados em periódicos revisados por pares.

Figura 14 – Artigos selecionados por ano de publicação.

2.3.4 Quais países contribuíram mais significativamente com publicações no contexto de técnicas de pré-treinamento de linguagem bidirecional, transformadores ou modelos de linguagem ampla utilizados na extração de entidades nominadas em documentos relacionados à saúde?

A Figura 15 apresenta os países que publicaram pesquisas sobre o tema abordado neste mapeamento. O país que mais se destacou foi a China, com o maior número de publicações, seguida pelos Estados Unidos e pela Coreia do Sul.

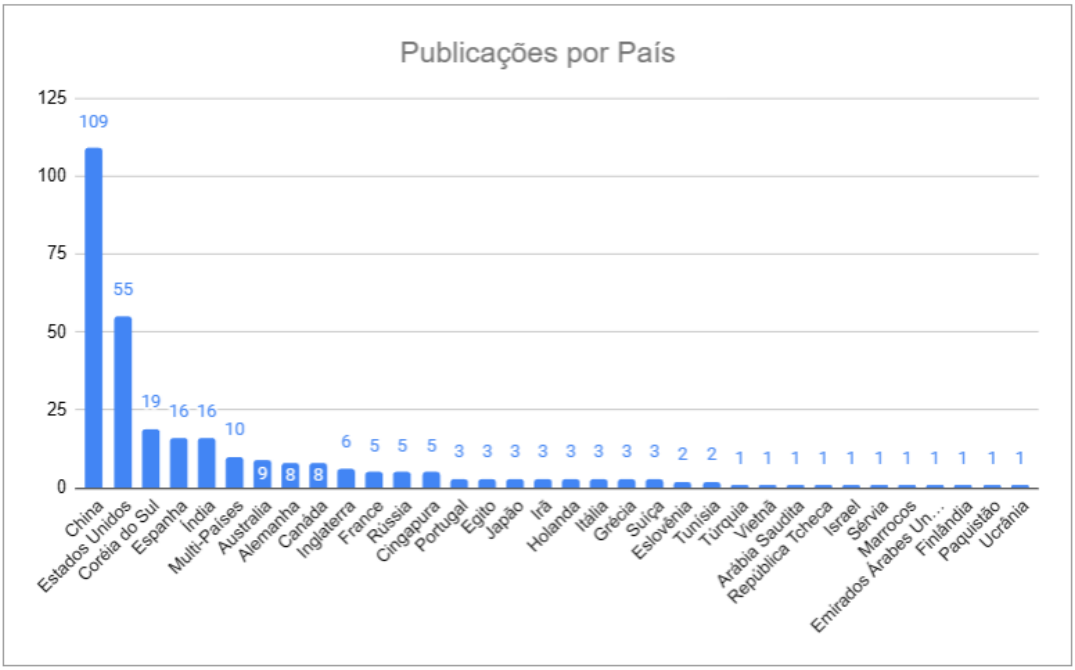

Figura 15 – Publicações por país.

2.3.5 Contexto e Síntese da Literatura

O autor (RABINOWITZ, 1987) fornece uma perspectiva detalhada sobre como a síntese narrativa ocorre dentro de um mapeamento sistemático da literatura. A síntese concentra-se em apresentar dados relevantes para a análise proposta de forma resumida. Além de simplificar e tornar acessível um extenso corpo de literatura, a síntese também ajuda os investigadores a formular novas questões e identificar áreas de investigação promissoras.

Várias abordagens e técnicas inteligentes têm sido aplicadas ao NER, com a maioria dos artigos sobre o tema publicados em 2023, indicando que este campo de pesquisa ainda está em crescimento. Os resultados mostram que as publicações sobre este tema abrangem vários países, evidenciando uma preocupação global em encontrar técnicas eficazes para auxiliar no contexto de textos relacionados à saúde.

2.3.5.1 Modelos e Arquiteturas Predominantes

A maioria dos estudos utiliza o modelo BERT e suas variantes, como BioBERT, ClinicalBERT e PubMedBERT, reconhecidos por sua alta eficácia em tarefas de NER e extração de relações (ER), frequentemente alcançando pontuações F1 acima de 90%. As variantes são ajustadas para domínios específicos usando dados de fontes como PubMed, aprimorando sua compreensão e processamento de textos biomédicos. Outro modelo amplamente utilizado, o BiLSTM-CRF, combina redes neurais recorrentes bidirecionais com CRFs para capturar efetivamente dependências sequenciais e contextuais, embora seu desempenho varie dependendo do contexto e dos dados. Além disso, arquiteturas híbridas como BERT-CNN-BiLSTM-CRF (ZHANG et al., 2021)integram diferentes técnicas para melhorar a captura do contexto e das relações semânticas nos dados. Esses modelos são avaliados em uma variedade de corpora, incluindo dados clínicos (MIMIC-III e i2b2), dados biomédicos (BC5CDR e NCBI) e dados de redes sociais. As melhorias de desempenho dependem da complexidade e do tipo de dados, com BioBERT e ClinicalBERT apresentando avanços consistentes nas tarefas NER e ER. Técnicas como aumento de dados e anotações manuais também contribuem para um desempenho superior.

2.3.5.2 Desempenho e Inovações na Pesquisa

Inovações como adversarial learning (GUO; ZHANG, 2023) e prompt tuning (HE et al., 2024) são empregadas para melhorar o desempenho em tarefas específicas.

- Avanços e Precisão: Modelos como BioELECTRA e BioALBERT demonstraram melhorias nas tarefas BioNLP (NASEEM et al., 2022), avançando em precisão e recuperação. Arquiteturas como transformadores com camadas de atenção e redes de memória são conhecidas por sua capacidade de capturar relações complexas dentro dos dados. Aplicados a uma ampla gama de tarefas, esses modelos abordam tudo, desde a identificação de entidades biomédicas até a extração de relações em dados clínicos, oferecendo um impacto significativo na pesquisa biomédica e na tomada de decisões clínicas. Por exemplo, um estudo relatou uma pontuação F1 de 98,2% (WANG et al., 2021), demonstrando a robusta precisão desses modelos.

- Variação e Eficácia: Modelos como BiLSTM-CRF e BERT-BiLSTM-CRF demonstram eficácia variável dependendo do corpus, com pontuações F1 flutuando com base no conjunto de dados utilizado. Isso destaca a importância do ajuste fino e da seleção de modelos específicos para cada tarefa. o Por exemplo, o BiLSTM-CRF (CHENG et al., 2021) alcançou uma pontuação F1 média de 91,07% nos conjuntos de dados CCKS2017 e CCKS2018 e 87,05% no conjunto de dados privado FCCd. o A nossa análise sugere que o BERT oferece uma precisão ligeiramente superior e um desempenho sólido em tarefas de processamento em tempo real. o Variantes especializadas, como o BioClinicalBERT (SHYR et al., 2024), também mostraram desempenho superior em tarefas específicas,

como a identificação de doenças raras e sinais clínicos, com pontuações F1 de 0,778 e 0,725, respetivamente.

2.4 Ameaças à Validade

Uma ameaça à validade de um estudo é qualquer fator que possa afetar a validade interna ou externa dos resultados obtidos. Assim, o estudo identifica os seguintes riscos à validade:

Viés de Seleção: As publicações incluídas neste estudo não refletem a população total de estudos primários dos últimos cinco anos. Por estar diretamente relacionado a critérios específicos, não inclui a diversidade existente de estudos primários disponíveis.

Viés de Exclusão: Publicações relevantes que podem ter sido excluídas pelos critérios de exclusão neste estudo podem levar à subestimação ou superestimação dos efeitos observados.

Viés de Idioma: A limitação de estudos no idioma selecionado pode limitar a generalização dos resultados para a população ou contexto de outro idioma.

Viés de Tempo: A validade dos resultados não se limitaria apenas ao último ano, visto que a prática, a tecnologia ou o método de pesquisa mudariam substancialmente.

3

SUS Audit Aided by Natural Language Processing: A Comparative Evaluation of BERT Models in the Analysis of Health News

Este capítulo apresenta os resultados de um artigo em fase de publicação após aceite na revista Array, no qual é conduzida uma avaliação experimental controlada de modelos baseados em BERT aplicados à análise de notícias de saúde. O estudo operacionaliza, em ambiente experimental, as recomendações e lacunas identificadas no Mapeamento Sistemático da Literatura apresentado no Capítulo 2.

3.1 ModBERTBr

Para explorar a fronteira tecnológica, foi incluído o ModernBERT, um modelo de encoder de última geração introduzido na comunidade científica em dezembro de 2024 por (Answer.AI; LightOn, 2024) e (LightOn, 2024). O ModernBERT incorpora melhorias arquitetônicas significativas em relação ao BERT, como maior capacidade de contexto e eficiência. O modelo central testado é o ModBERTBr, que é uma derivação do ModernBERT, especializado na língua portuguesa brasileira. Este modelo foi pré-treinado especificamente para o escopo linguístico do português do Brasil, utilizando dados do Corpus BrWac (FILHO et al., 2018) e do subconjunto em português do conjunto de dados da Wikipédia (WALLACE, 2024).

Neste contexto, o trabalho (FONTES; COLAÇO JÚNIOR; PRADO, 2023) classificou 56.053 registros como evidências para auditoria em saúde, de mais de 2 milhões de materiais. Este estudo, ao focar na aplicação e avaliação de modelos BERT para NER e classificação de texto em notícias de saúde para auditoria do SUS, complementa e aprofunda as capacidades de análise textual propostas pelo Sussurro, buscando otimizar ainda mais a seleção de conteúdo para auditores.

3.2 Coleção de Banco de Dados

O experimento foi conduzido utilizando notícias de saúde para auditoria. Os dados foram extraídos de dois arquivos principais: dataset.conll para a tarefa de NER e Amostra.csv para a classificação de texto. Os arquivos de dados foram acessíveis no ambiente Google Colab. O conjunto de dados para o experimento, compreendendo aproximadamente 800 notícias, foi previamente extraído, pré-processado e classificado. A amostra foi considerada estatisticamente representativa com base no cálculo para população infinita, um método apropriado quando a população é significativamente grande, como é o caso das 60.000 notícias, das quais as 800 representam apenas 1,33%.

O cálculo de amostragem foi realizado utilizando a fórmula padrão para populações grandes. Adotamos um nível de confiança de 95% (com Z = 1.96) e uma proporção conservadora de P = 0.5, que maximiza o tamanho da amostra e garante a maior segurança estatística. Ao utilizar 800 notícias, alcançamos uma margem de erro de aproximadamente 3,5% (0.035). Dessa forma, confirmamos que 800 notícias é um tamanho amostral estatisticamente robusto e suficiente. Este número nos permitiu obter uma precisão superior à margem de erro usualmente aceita em pesquisas (5%), garantindo alta confiabilidade nos resultados do experimento, conforme demonstrado pelo cálculo em (Colaço Júnior, 2025).

3.2.1 Seleção de notícias relacionadas à saúde

As notícias coletadas foram selecionadas com base na relevância para a auditoria do SUS e categorizadas nas classes: 'Saúde', 'Genérica' e 'Fraude/Má Gestão'. A etapa de anotação manual foi conduzida na plataforma Label Studio, resultando na geração de dois conjuntos de dados distintos: o arquivo dataset.conll, destinado à tarefa de NER, e o amostra.csv, voltado para a classificação documental.

Para a tarefa de NER, que consiste em classificar e delimitar tokens pertencentes a categorias específicas, adotou-se o esquema de etiquetagem IOB. Os experimentos computacionais foram processados na infraestrutura do Google Colab. A taxonomia completa das entidades utilizadas, divididas por áreas de interesse, é apresentada na Tabela 7, logo em seguida na Tabela 8 poderá encontrar exemplos de notícias e suas categorizações.

Tabela 7 – Taxonomia de Entidades Nomeadas por Área

Entidade	**Tag**	**Descrição**
	Entidades Gerais	
Pessoa	PER	Nome de indivíduos
Organização	ORG	Instituições públicas, privadas, ONGs
Localização	LOC	Países, cidades, estados, hospitais
Data	DATE	Datas absolutas ou relativas
Valor	VAL	Valores financeiros, contratos, propinas
	Entidades Específicas de Notícias de Saúde	
Doença / Condição	DIS	Nome de doenças e quadros clínicos
Tratamento / Produto	MED	Medicamentos, vacinas, equipamentos hospitalares
Sintoma	SYM	Sintomas relatados
Fabricante / Marca	FAB	Empresas farmacêuticas ou fornecedoras
	Entidades Indicativas de Corrupção	
Agente Público	FUNC	Pessoas envolvidas em cargos públicos com poder decisório
Tipo de Irregularidade	IRREG	Natureza do ato ilícito
Contrato / Licitação	LIC	Nome ou identificação de contratos públicos
Empresa Envolvida	EMP_FRAUDE	Empresas investigadas ou condenadas
Órgão Investigativo	ORG_INVEST	Instituições de controle, fiscalização e justiça
Operação Policial	OP	Nome de operações que investigam corrupção
Processo / Inquérito	PROC	Números ou nomes de processos legais
Instrumento Legal	LEGAL_INST	Leis, decretos e regulamentações

Notícia	**Categoria**
CPI ouve ex-diretor da Saúde acusado de pedir propina de US$ 1 por vacina.	Fraude
UCB oferece serviços médicos gratuitos para a população de Ceilândia.	Genérico
De salame a refrigerante: veja como ultraprocessados detonam sua saúde.	Saúde

Tabela 8 – Exemplos de notícias e suas respectivas categorias

Para exemplificar a aplicação prática da taxonomia apresentada, a Figura 16 demonstra o resultado da predição do modelo desenvolvido. No exemplo, é possível visualizar o funcionamento da camada de NER, que identifica sintomas (SYM), doenças (DIS) e produtos médicos (MED) seguindo a estrutura de marcação IOB. Adicionalmente, a figura exibe o resultado da classificação de texto, onde a sentença é rotulada dentro das categorias pré-definidas.

```
Sentença de Teste 2: 'Estou com dor de barriga e com febre, a covid-19 deve ter sido o culpado, tenho medo de morrer desta doença, mas tomei vacina.'
Entidades Previstas:
  Palavra: 'Estou' | Tag: 'O'
  Palavra: 'com' | Tag: 'O'
  Palavra: 'dor' | Tag: 'B-SYM'
  Palavra: 'de' | Tag: 'I-DIS'
  Palavra: 'barriga' | Tag: 'I-DIS'
  Palavra: 'e' | Tag: 'O'
  Palavra: 'com' | Tag: 'O'
  Palavra: 'febre' | Tag: 'B-SYM'
  Palavra: ',' | Tag: 'O'
  Palavra: 'a' | Tag: 'O'
  Palavra: 'covid' | Tag: 'B-DIS'
  Palavra: '-' | Tag: 'I-DIS'
  Palavra: '19' | Tag: 'I-DIS'
  Palavra: 'deve' | Tag: 'O'
  Palavra: 'ter' | Tag: 'O'
  Palavra: 'sido' | Tag: 'O'
  Palavra: 'o' | Tag: 'O'
  Palavra: 'culpado' | Tag: 'O'
  Palavra: ',' | Tag: 'O'
  Palavra: 'tenho' | Tag: 'O'
  Palavra: 'medo' | Tag: 'O'
  Palavra: 'de' | Tag: 'O'
  Palavra: 'morrer' | Tag: 'B-DIS'
  Palavra: 'desta' | Tag: 'O'
  Palavra: 'doença' | Tag: 'O'
  Palavra: ',' | Tag: 'O'
  Palavra: 'mas' | Tag: 'O'
  Palavra: 'tomei' | Tag: 'O'
  Palavra: 'vacina' | Tag: 'B-MED'
  Palavra: '.' | Tag: 'O'
                                    Categoria Prevista: Saúde
```

Figura 16 – Saída do modelo: Classificação de tokens e predição de categoria.

3.2.2 Métricas de avaliação (acurácia, recall, precisão, pontuação F1, TMT)

A acurácia (accuracy) representa o percentual de instâncias que foram classificadas de forma correta, sendo definida por:

$$acurrarcy = \frac{TP + TN}{TP + TN + FP + FN}$$

A revocação (recall), também conhecida como a taxa de verdadeiros positivos ou sensibilidade, é o percentual de instâncias positivas que foram classificadas corretamente:

$$recall = \frac{TP}{TP + FN}$$

A precisão (precision) é a razão entre as instâncias classificadas como "verdadeiro positivo" e todas as instâncias classificadas como positivas:

$$precision = \frac{TP}{TP + FP}$$

A medida-F1 (F1-Score) é a métrica que combina dois indicadores de desempenho, sendo a expressão da média harmônica da precisão e da revocação:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

3.3 Avaliação Experimental

3.3.1 Objetivo

Este estudo teve como objetivo aplicar técnicas de NER e classificação de texto em notícias de saúde com foco em auditoria para o SUS. Foram comparados três modelos: BERT com classificador, BERT com CRF (BERT-CRF) e ModBERTBr. Buscou-se identificar qual abordagem apresentou melhor desempenho na identificação de entidades e categorização das notícias.

3.3.2 Planejamento

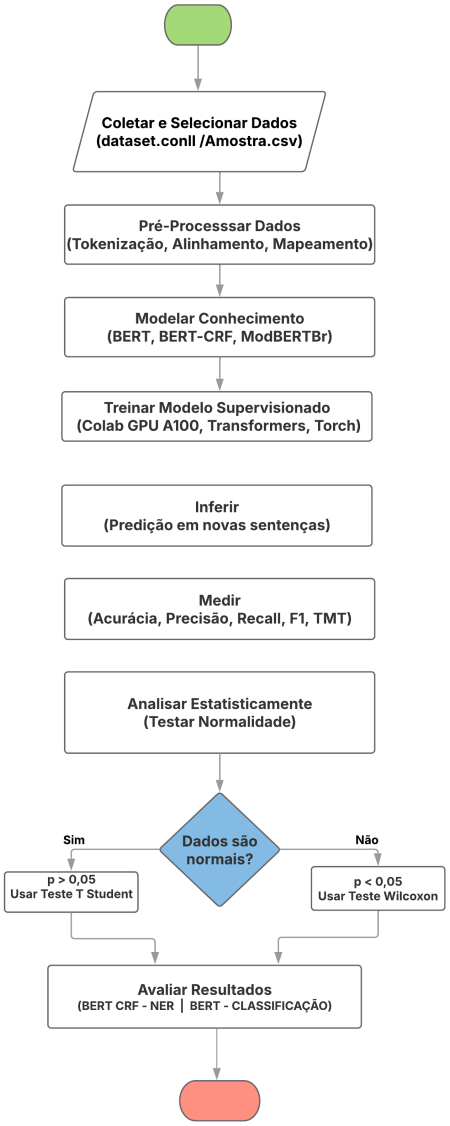

Figura 17 – Fluxograma da metodologia de pesquisa que demonstra o processo de avaliação de modelos de linguagem (BERT, BERT-CRF e ModBERTBr) para as tarefas de Reconhecimento de Entidades Nomeadas (NER) e Classificação de Texto.

O processo de pesquisa inicia-se na fase de planejamento, que estabelece o contexto (notícias de saúde para auditoria no SUS), a questão e as hipóteses de pesquisa, a definição das variáveis (dependentes e independentes), e a seleção dos objetos de estudo (modelos e categorias de notícias) e da instrumentação necessária (Label Studio, Google Colab A100, Python e biblioteca específica como Torch (PASZKE et al., 2019)).

Na sequência, inicia-se a fase de execução do experimento, estruturada pelas atividades sequenciais do fluxograma. A primeira etapa é a de Coletar e Selecionar Dados, utilizando os arquivos dataset.conll para a tarefa de NER e Amostra.csv para a Classificação de Texto. Em seguida, ocorre o Pré-Processar Dados, que engloba a Tokenização, o Alinhamento de rótulos e o Mapeamento de categorias.

Na etapa de Modelar Conhecimento, são aplicados os modelos BERT, BERT-CRF e ModBERTBr. Estes modelos são posteriormente submetidos ao Treinar Modelo Supervisionado em ambiente Google Colab com GPU A100, utilizando as bibliotecas Transformers e Torch. Após o treinamento, o modelo realiza a Inferir (Predição em novas sentenças), e os resultados são submetidos à etapa de Medir, onde o desempenho é quantificado através de métricas como Acurácia, Precisão, Recall, F1-score e Tempo Médio de Treinamento (TMT).

A fase final do processo é a de análise estatística, que se inicia com o Analisar Estatisticamente (Testar Normalidade). Na etapa de decisão (Dados são normais?), se o valor-p do teste de normalidade Shapiro-Wilk (SHAPIRO; WILK, 1965) for superior a 0,05 (p>0,05), indica-se que Sim, os dados são normais, e aplica-se o Teste T de Student. Caso o valor-p seja inferior a 0,05 (p<0,05), indica-se que Não, e emprega-se o teste não paramétrico de Teste Wilcoxon (ou Wilcoxon-Mann-Whitney). Por fim, a etapa de Avaliar Resultados apresenta as conclusões, destacando a performance superior do BERT-CRF para NER e do BERT para Classificação.

3.3.3 Pergunta de pesquisa e hipótese

Foi formulada a seguinte questão: Qual modelo (BERT-CRF, BERT ou ModBERTBr) apresenta melhor desempenho para NER em notícias de saúde voltadas à auditoria do SUS, e qual sua eficácia para classificação de texto neste contexto?

- **Hipótese nula (H0):** Não há diferença significativa no desempenho para a tarefa de NER e Classificação entre o modelo BERT-CRF, o modelo BERT padrão e o ModBERTBr em notícias de saúde voltadas à auditoria do SUS.
- **Hipótese Alternativa (H1):** O modelo BERT-CRF, por sua arquitetura intrinsecamente projetada para tarefas de sequenciamento como NER, apresentou um desempenho superior na identificação de entidades nomeadas e classificação de texto em comparação com o modelo BERT padrão, e obteve resultados comparáveis ou melhores que o ModBERTBr.

3.3.4 Variáveis dependentes

- Métricas: Acurácia, Recall, Precisão, F1-score, TMT.

3.3.5 Variáveis independentes

- Tipo de modelo (BERT, BERT-CRF, ModBERTBr).
- Conjunto de dados (dataset.conll e Amostra.csv).
- Hiperparâmetros (learning rate, batch size, número de épocas, weight decay, etc.).

3.3.5.1 Procedimento de Validação do Modelo

- K-fold cross-validation (com k=10)

3.3.6 Seleção de objetos

Os objetos de estudo foram os modelos de linguagem (BERT-CRF, Classificador de Texto, ModBERTBr) e as notícias de saúde para auditoria. As notícias foram categorizadas em saúde, genérica e fraude.

3.3.7 Projeto Experimental

Foi elaborado um experimento comparativo e controlado, com um pipeline completo: leitura e tokenização dos dados, alinhamento de rótulos, treinamento supervisionado, avaliação, inferência e análise estatística.

3.3.8 Instrumentação

As ferramentas utilizadas incluíram Label Studio, Excel, Python (com bibliotecas como transformers, torch, torcherf, pandas, json, os, datasets, sklearn.metrics, evaluate (seqeval), numpy, partial), Google Colab, BERT-base, CRF e ModBERTBr.

3.4 Operação do experimento

3.4.1 Preparação

Instalaram-se as bibliotecas necessárias e carregaram-se os dados. Realizou-se a tokenização e o alinhamento de rótulos para NER, e o mapeamento de categorias para classificação. Os modelos BertCRFForNER, AutoModelForTokenClassification e AutoModelForSequenceClassification foram inicializados. Argumentos de treinamento e funções para cálculo de métricas

(F1-score, Precisão, Recall, Acurácia) foram definidos. Objetos Trainer foram configurados para as tarefas de NER e classificação de texto.

3.4.2 Execução

Treinaram-se os modelos BertCRFForNER e AutoModelForSequenceClassification em uma máquina Colab A100. Avaliaram-se ambos usando métricas padrão. Após o treinamento, os modelos salvos e seus tokenizadores foram carregados e configurados para o modo de avaliação, sendo então utilizados para realizar inferência (predições) em novas frases, com funções específicas para NER e classificação de categoria.

3.4.3 Ambiente

O experimento foi realizado no Google Colab, com uso de GPU (se disponível), Python 3.10+, Hugging Face (transformers, datasets, evaluate, seqeval, pytorch-crf) e bibliotecas auxiliares como torch, pandas, json e os. Os arquivos de dados (dataset.conll, Amostra.csv) foram acessíveis no ambiente Colab, e os modelos e logs treinados foram salvos em diretórios locais.

3.5 Resultado

3.5.1 Validação de dados

A normalidade dos dados foi verificada usando o teste de Shapiro-Wilk, uma técnica recomendada para tamanhos de amostra pequenos a médios, de acordo com a literatura especializada. O processamento e a análise estatística foram conduzidos usando o software jamovi (The JAMOVI project, 2025). A decisão sobre a normalidade foi baseada em um nível de significância de 0,05.

Quando o valor p do teste de Shapiro-Wilk estava abaixo desse limite, a hipótese nula de que os dados seguiam uma distribuição normal foi rejeitada. Para variáveis com distribuição normal, usamos o teste t de Student, uma técnica paramétrica para comparar médias entre grupos. Por outro lado, para dados que não apresentaram normalidade, usamos o teste não paramétrico de Wilcoxon-Mann-Whitney, que compara distribuições com base nas posições (classificações) dos valores observados. Ambos os testes foram realizados em comparações aos pares, destacando as diferenças ou semelhanças entre os modelos BERT, BERT-CRF e ModBERTBr.

3.5.2 Análise e interpretação de dados

3.5.2.1 Tarefa NER

Para responder às questões de pesquisa listadas na Seção 4.2.2, a etapa de execução foi conduzida, e os resultados de classificação, bem como o tempo médio, foram obtidos para as

tarefas de NER e Classificação de Categoria. As métricas de desempenho para a tarefa de ER são apresentadas na Figura 18, enquanto o tempo médio para ambas as tarefas é detalhado na Figura 19.

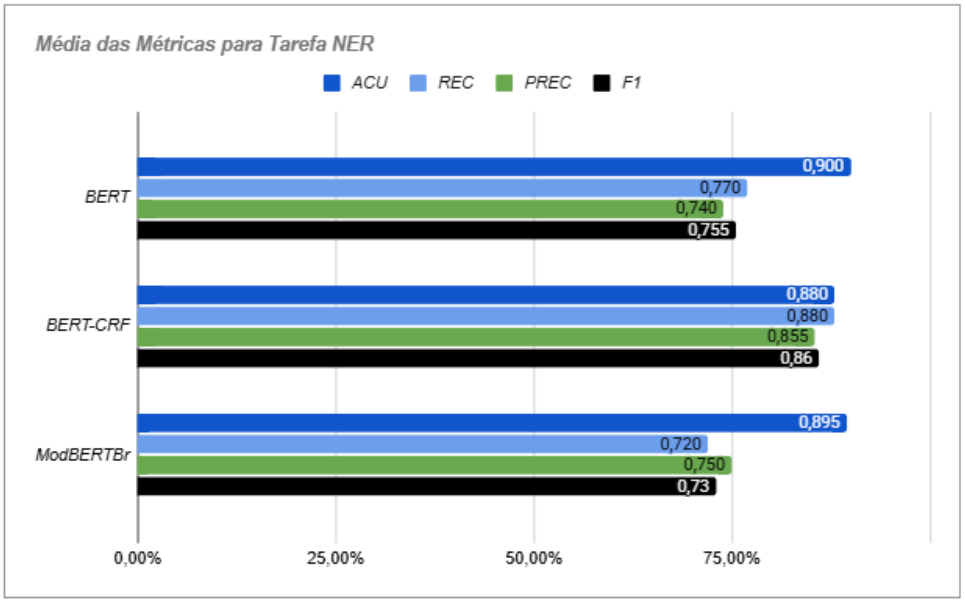

Figura 18 – Comparação das médias das métricas para tarefa NER.

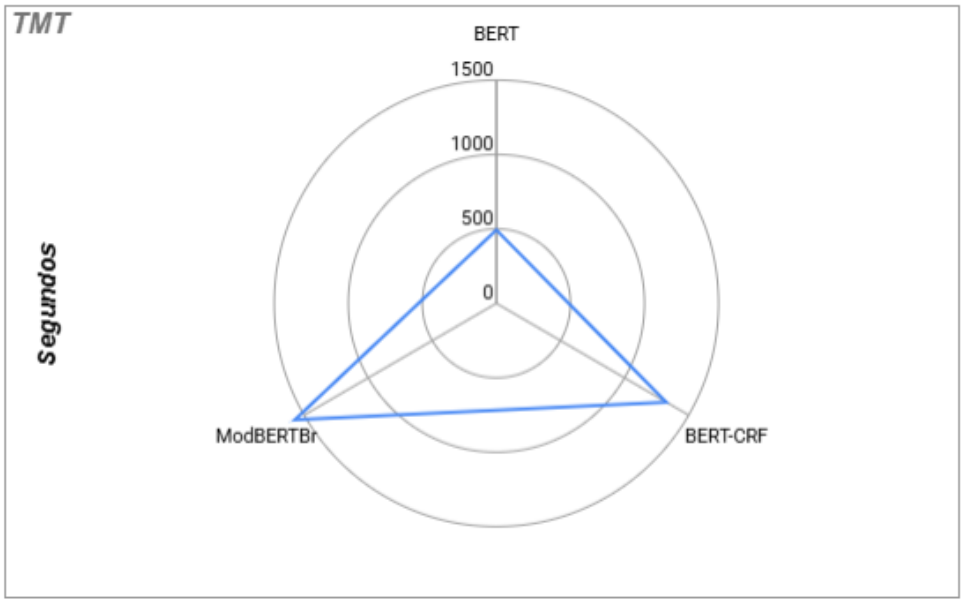

Figura 19 – Tempo médio em segundos para cada modelo na Tarefa NER.

Os três algoritmos BERT, BERT-CRF e ModBERTBr apresentaram médias de acurácia (ACU) semelhantes, variando de 0,880 a 0,900. Em termos de desempenho, o BERT-CRF se destaca com os melhores resultados nas métricas de recall (0,880), precisão (0,855) e F1 (0,860). O algoritmo BERT obteve a melhor acurácia (0,900). Já o ModBERTBr apresentou os menores

valores de acurácia, recall e F1-score. Em relação ao tempo de treinamento (TMT), o BERT foi o mais rápido, com um tempo de (8 minutos e 10 segundos). O BERT-CRF e o ModBERTBr tiveram tempos de treinamento mais longos, com (22 minutos e 5 segundos) e (26 minutos e 10 segundos), respectivamente.

Embora os algoritmos tenham demonstrado boa acurácia para a tarefa, a imagem por si só não fornece evidências estatísticas conclusivas para afirmar qual deles é o melhor, pois não inclui testes de significância ou valores de p, portanto foi estabelecido um nível de significância (ÿ) de 0,05 para todo o experimento. Quando o teste de Shapiro-Wilk foi aplicado para analisar a normalidade da distribuição dos dados, foram obtidos os valores de p mostrados na Tabela 9.

Tabela 9 – Resultados do teste de Shapiro-Wilk para analisar a normalidade dos dados do NER.

Algoritmo	**ACU (p-value)**	**REC (p-value)**	**PREC (p-value)**	**F1 (p-value)**
BERT	0,359	0,273	0,101	0,438
BERT-CRF	0,238	0,238	0,520	0,583
ModBERTBr	0,008	0,133	0,258	0,307

A análise da normalidade dos dados revela que o modelo BERT demonstra a distribuição mais próxima de uma curva normal para a maioria das métricas avaliadas. Com valores p de 0,359 para ACU, 0,273 para REC e 0,438 para F1, o modelo BERT consistentemente superou os demais, indicando que não há evidência para rejeitar a hipótese de normalidade. Em seguida, o modelo BERT-CRF se posiciona como o segundo melhor em termos de normalidade, com destaque para a métrica PREC, onde obteve o valor p mais elevado (0,520). Por outro lado, o modelo ModBERTBr apresentou valores p extremamente baixos, como 0,008 para ACU, sugerindo que os dados de desempenho associados a este modelo não seguem uma distribuição normal. Em suma, com um nível de significância de 0,05, todos os modelos, com exceção do ModBERTBr na métrica ACU, indicam que os dados são compatíveis com uma distribuição normal.

Em complemento à metodologia de validação dos dados, os testes de comparação par a par, conforme visualizado na Tabela 10, foram realizados para identificar quais pares de modelos apresentaram diferenças de desempenho estatisticamente significativas. Essa abordagem granular foi crucial para validar as conclusões do estudo, permitindo ir além da simples constatação de que os modelos eram diferentes entre si. A estratégia de comparação adotada focou em testar a superioridade do modelo BERT, que se destacou nos testes de normalidade. As comparações foram conduzidas confrontando o BERT com os demais modelos, seguindo a ordem: BERT vs. BERT-CRF e BERT vs. ModBERTBr.

Tabela 10 – Teste de W de Wilcoxon e T de Student, dois por dois.

	BERT - BERT-CRF			**BERT - ModBERTBr**		
Métrica	Diferença	**p-value**	**Teste**	Diferença	**p-value**	**Teste**
ACU	0,0260	0,001	Teste t de Student	-0,0100	0,066	Teste de Wilcoxon
REC	-0,0990	0,001	Teste t de Student	0,0600	0,001	Teste t de Student
PREC	-0,1030	0,001	Teste t de Student	-0,0010	0,895	Teste t de Student
F1	-0,0940	0,001	Teste t de Student	0,0310	0,010	Teste t de Student

A análise estatística de comparação par a par revelou diferenças de desempenho significativas entre os modelos. Os testes de t de Student indicaram que o modelo BERT diferia do BERT-CRF de forma estatisticamente significativa em todas as métricas (ACU, REC, PREC e F1), embora o BERT-CRF tenha apresentado desempenho superior nas métricas REC, PREC e F1. Por outro lado, a comparação entre BERT e ModBERTBr, utilizando uma combinação de testes t de Student e W de Wilcoxon, mostrou que as diferenças foram significativas apenas para as métricas REC e F1, onde o BERT se sobressaiu. Para ACU e PREC, as diferenças de desempenho entre os dois modelos foram consideradas estatisticamente irrelevantes, sugerindo que suas performances são similares nessas métricas.

3.5.2.2 Tarefa de Classificação de Categoria

A análise de normalidade dos dados para a tarefa de classificação, baseada nos valores p e apresentada na Tabela 11, revela que todos os modelos demonstram uma distribuição compatível com a normal para a tarefa de classificação. Em um nível de significância de 0,05, não há evidência para rejeitar a hipótese de normalidade para nenhum dos modelos em qualquer métrica.

O modelo ModBERTBr apresentou a maior compatibilidade geral, com o valor p mais elevado para a métrica ACU (0,279), seguido por REC (0,075), PREC (0,124) e F1 (0,124). O modelo BERT-CRF, por sua vez, destacou-se com o valor p mais alto para a métrica PREC (0,254). Por fim, o modelo BERT obteve p-values consistentes, variando de 0,147 a 0,190, indicando também uma distribuição normal para todas as métricas.

Tabela 11 – Resultados do teste de Shapiro-Wilk para analisar a normalidade dos dados da Classificação de Categoria.

Algoritmo	**ACU (p-value)**	**REC (p-value)**	**PREC (p-value)**	**F1 (p-value)**
BERT	0,153	0,153	0,190	0,147
BERT-CRF	0,092	0,092	0,254	0,092
ModBERTBr	0,279	0,075	0,124	0,124

Complementando a análise de normalidade, as métricas de desempenho para a tarefa de classificação mostraram variações notáveis entre os modelos. O modelo BERT demonstrou o

melhor desempenho geral, com uma média de 0,928 nas métricas de ACU, REC, PREC e F1, e um pico em PREC (0,933). O BERT-CRF apresentou um desempenho muito similar, com uma média geral de 0,926, confirmando sua robustez. Em contrapartida, o ModBERTBr consistentemente obteve valores inferiores, com pontuações variando de 0,903 a 0,910. Essa análise demonstra que os modelos baseados em BERT superaram significativamente o ModBERTBr em todas as métricas de performance. O tempo médio de execução para cada modelo será detalhado na Figura 20.

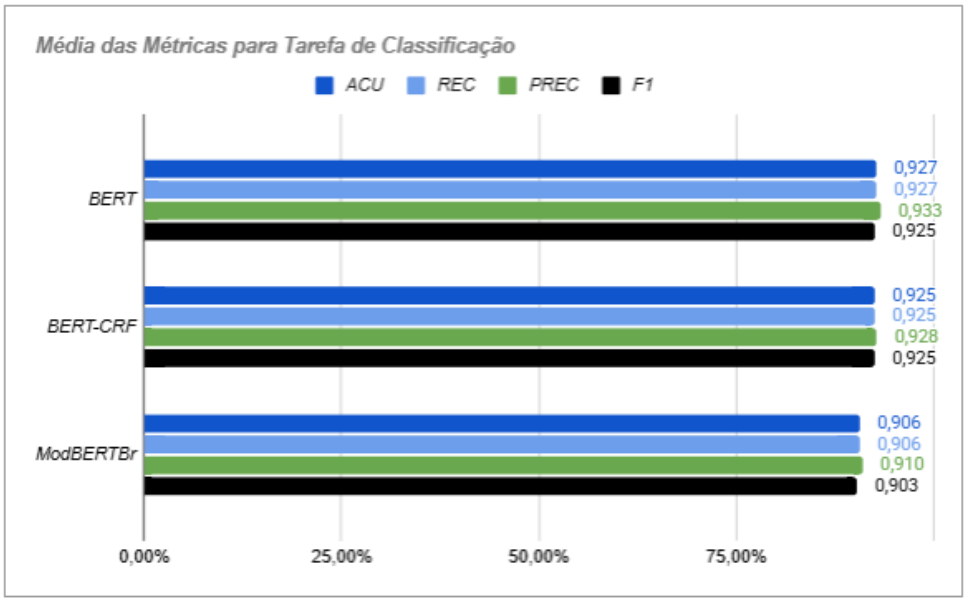

Figura 20 – Comparação das médias das métricas para tarefa de Classificação de Categoria.

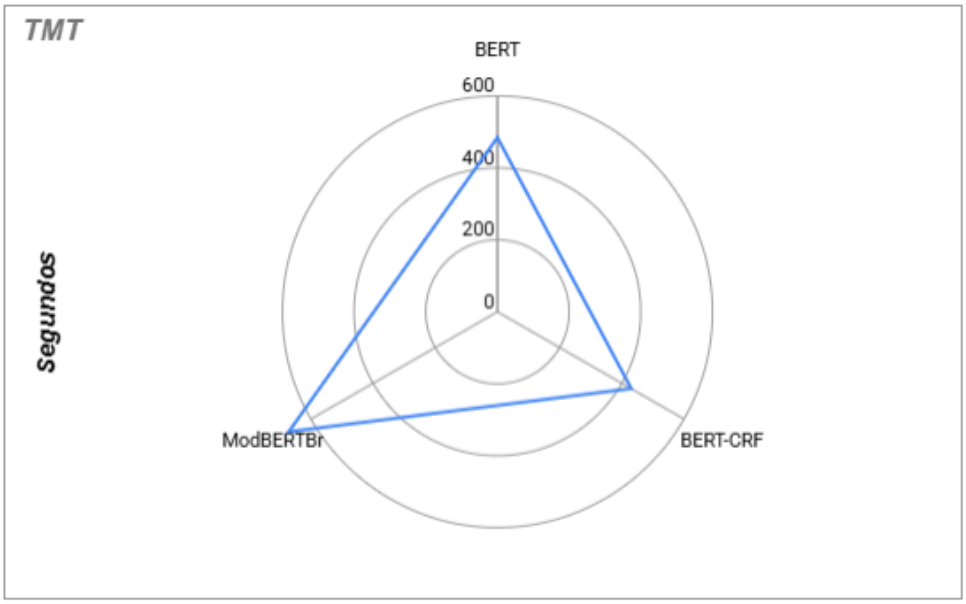

Figura 21 – Tempo médio em segundos para cada modelo na Tarefa de Classificação de Categoria.

A análise da eficiência computacional dos modelos revela uma clara distinção no tempo

de execução. O modelo BERT-CRF demonstrou ser o mais rápido, com um tempo médio de 430 segundos (7 minutos e 10 segundos). O modelo BERT apresentou um tempo ligeiramente superior, com 485 segundos (8 minutos e 5 segundos), mas ainda assim manteve-se como uma opção de alta eficiência. Em contrapartida, o ModBERTBr foi o modelo menos eficiente, com um tempo médio de execução de 670 segundos (11 minutos e 10 segundos), o que representa um tempo substancialmente maior que os demais. Esses resultados indicam que, em cenários onde a velocidade de processamento é um fator crítico, os modelos baseados em BERT são consideravelmente superiores. Com base na normalidade dos dados para a tarefa de classificação de categoria, o Teste t de Student foi empregado para determinar a significância das diferenças de desempenho entre os modelos. Os resultados dessa análise, apresentados na Tabela 4, mostram conclusões distintas para cada par de modelos comparados.

Tabela 12 – Resultados do teste t de Student para comparação aos pares entre modelos

Métrica	**Dif.**	**p-value**	**Teste**	**Dif.**	**p-value**	**Teste**
	BERT - BERT-CRF			**BERT - ModBERTBr**		
ACU	0,0020	0,591	t de Student	0,0210	0,011	t de Student
REC	0,0020	0,591	t de Student	0,0210	0,011	t de Student
PREC	0,0050	0,138	t de Student	0,0230	0,003	t de Student
F1	$1{,}11\times10^{-17}$	1,000	t de Student	0,0220	0,006	t de Student

Para a comparação entre BERT e BERT-CRF, os valores p foram consistentemente altos, variando de 0,138 a 1,000. Como todos esses valores estão acima do nível de significância de 0,05, não há evidência estatística para rejeitar a hipótese nula. Isso indica que a diferença média de desempenho entre os modelos BERT e BERT-CRF para as métricas avaliadas não é estatisticamente significativa.

Por outro lado, a comparação entre BERT e ModBERTBr apresentou valores p extremamente baixos, com todos os resultados abaixo de 0,05 (variando de 0,003 a 0,011). Isso demonstra que as diferenças observadas no desempenho entre esses dois modelos são estatisticamente significativas. Em suma, o modelo BERT supera o ModBERTBr de forma estatisticamente confiável.

4

Discussão

Este capítulo estabelece a articulação entre as evidências consolidadas no MSL, as decisões de projeto que orientaram a construção do experimento e os resultados empíricos obtidos. Diferentemente de uma mera descrição de métricas, a discussão proposta busca interpretar os achados à luz da literatura, explicitar convergências e divergências entre evidência teórica e comportamento experimental, e derivar implicações concretas para o desenho de sistemas de apoio à auditoria no âmbito do SUS.

O MSL apresentado no Capítulo 2 indicou uma predominância de arquiteturas baseadas em Transformers para tarefas de NER, com destaque para abordagens que integram mecanismos de modelagem sequencial, como o CRF. A literatura sugere que tais arquiteturas tendem a apresentar maior robustez na identificação de limites de entidades e na preservação da coerência estrutural das sequências, especialmente em domínios caracterizados por textos longos e semanticamente densos, como documentos administrativos e de saúde pública. O experimento conduzido neste trabalho operacionaliza essas recomendações e permite avaliá-las em um cenário aplicado, baseado em notícias relacionadas à auditoria do SUS.

Os resultados experimentais confirmam, em grande medida, as tendências apontadas pelo MSL. O modelo BERT-CRF apresentou desempenho superior na tarefa de NER, superando as demais arquiteturas em recall (0,880), precisão (0,855) e pontuação F1 (0,860). Embora o modelo BERT independente tenha alcançado a maior acurácia global (0,900), a análise estatística evidenciou diferenças significativas em favor do BERT-CRF nas métricas diretamente associadas à extração estruturada de entidades. Esse achado reforça a distinção, já sugerida pela literatura, entre métricas globais de acerto e métricas sensíveis à qualidade da segmentação sequencial.

Em contrapartida, o desempenho inferior do ModBERTBr em todas as tarefas avaliadas indica limitações de generalização para as especificidades linguísticas e semânticas do domínio de auditoria em saúde. Embora treinado para o português, o modelo mostrou menor capacidade de adaptação a variações terminológicas, estruturas sintáticas complexas e entidades compostas,

características recorrentes nos textos analisados. Esse resultado dialoga com lacunas identificadas no MSL quanto à escassez de avaliações controladas de modelos nacionais em contextos aplicados específicos.

Sob a perspectiva de eficiência computacional, os resultados revelam um trade-off relevante para o desenho de sistemas em escala. O BERT apresentou o menor tempo de treinamento na tarefa de NER (8 minutos e 10 segundos), enquanto o BERT-CRF demandou maior custo computacional. Esse comportamento se inverte parcialmente na tarefa de Classificação por Categoria, na qual o BERT-CRF alcançou o menor tempo de treinamento (7 minutos e 10 segundos). Esses achados indicam que a eficiência não é uma propriedade intrínseca do modelo, mas depende da interação entre arquitetura, tarefa e objetivo operacional do sistema.

4.1 Conditional Random Field

O desempenho superior do BERT-CRF na tarefa de NER pode ser atribuído à incorporação da camada CRF, responsável por modelar explicitamente as dependências entre rótulos adjacentes. Essa capacidade é particularmente relevante em sequências longas e semanticamente interdependentes, como aquelas presentes em documentos relacionados à auditoria do SUS, nos quais entidades como nomes de instituições, procedimentos médicos, datas e valores financeiros apresentam fortes restrições estruturais. Ao reduzir previsões inconsistentes e reforçar a coerência das sequências rotuladas, a camada CRF contribui diretamente para uma extração mais confiável de informações críticas para a análise de irregularidades.

A análise qualitativa dos erros complementa as métricas quantitativas ao revelar que a ambiguidade semântica e a variação vocabular constituem as principais fontes de falha dos modelos. Casos envolvendo sinônimos de procedimentos médicos, abreviações regionais e terminologia administrativa específica evidenciam desafios inerentes ao processamento de linguagem natural em documentos heterogêneos. Essas limitações reforçam a necessidade de estratégias complementares, como enriquecimento do corpus, refinamento do esquema de anotação e uso de técnicas adaptativas.

Na tarefa de Classificação por Categoria, o BERT demonstrou elevada robustez, superando significativamente o ModBERTBr em todas as métricas avaliadas. Embora o desempenho do BERT e do BERT-CRF tenha sido comparável em termos de acurácia, a maior eficiência do BERT o torna particularmente adequado para cenários de triagem rápida de grandes volumes de documentos. Em contraste, o BERT-CRF mostrou-se mais apropriado para etapas que demandam extração precisa e contextualizada de entidades.

Essas evidências convergem para uma implicação direta de desenho de sistema: para auditorias baseadas em grandes coleções textuais, a adoção de um único modelo não atende plenamente às exigências operacionais. Os resultados deste estudo sustentam a proposta de uma arquitetura híbrida, na qual modelos baseados em BERT são empregados na triagem inicial

e classificação em larga escala, enquanto arquiteturas BERT-CRF são acionadas em etapas subsequentes de extração detalhada de entidades. Tal arranjo combina eficiência computacional e precisão informacional, alinhando-se às demandas práticas de auditoria no contexto do SUS.

5
Conclusão

Este trabalho investigou como técnicas modernas de PLN, em especial arquiteturas baseadas em Transformers, podem contribuir para a otimização da auditoria do SUS a partir da análise automatizada de grandes volumes de documentos textuais. Ao articular evidência científica consolidada, experimentação controlada e implicações de engenharia de sistemas, a pesquisa avança da análise acadêmica para recomendações operacionais concretas, alinhadas às demandas reais de órgãos de controle e fiscalização.

5.1 Contribuições

As contribuições deste trabalho são integradas e materializam-se em dois artigos científicos com diferentes níveis de maturidade, refletindo etapas complementares do processo de pesquisa. O primeiro artigo, publicado na revista *Frontiers in Artificial Intelligence*, apresenta um Mapeamento Sistemático da Literatura que consolida o estado da arte das arquiteturas de Reconhecimento de Entidades Nomeadas aplicadas ao domínio da saúde. Esse estudo fornece evidência científica robusta sobre tendências, lacunas e limitações da literatura, servindo como base para as decisões metodológicas adotadas nesta dissertação.

O segundo artigo, em fase de revisão na revista *Array*, operacionaliza empiricamente essas evidências por meio de uma comparação controlada entre modelos BERT, BERT-CRF e ModBERTBr, aplicada a um corpus específico de notícias relacionadas à auditoria do SUS. Essa avaliação experimental permite verificar, em um cenário aplicado, quais recomendações da literatura se confirmam e quais se mostram contingentes ao contexto.

A partir desses estudos, emergem três contribuições centrais: (i) a consolidação do estado da arte sobre NER em saúde por meio de evidência sistematizada; (ii) a análise comparativa controlada de arquiteturas em um domínio realista e pouco explorado; e (iii) a formulação de uma recomendação operacional explícita para o uso de arquiteturas híbridas BERT/BERT-CRF

em fluxos de auditoria.

5.2 Resposta à Questão Principal de Pesquisa

A questão central desta pesquisa buscou identificar quais arquiteturas de aprendizado profundo são mais adequadas para apoiar auditorias do SUS baseadas em análise textual. Os resultados demonstram que não existe um modelo universalmente superior para todas as etapas do processo. O BERT destacou-se pela eficiência e robustez na classificação de documentos em larga escala, enquanto o BERT-CRF apresentou desempenho superior na extração estruturada de entidades em textos longos e complexos.

Portanto, para auditorias do SUS baseadas em notícias e documentos textuais, a melhor estratégia não é a escolha de um único modelo, mas a combinação de abordagens complementares. A integração de modelos BERT para triagem inicial e priorização de documentos, com modelos BERT-CRF para extração precisa de entidades relevantes, mostrou-se a solução mais alinhada às exigências operacionais e informacionais do contexto analisado.

5.3 Recomendações Práticas

Os achados desta pesquisa permitem derivar recomendações claras, orientadas à tomada de decisão institucional. Quando a prioridade é a triagem rápida de grandes volumes de documentos, com foco em escalabilidade e eficiência computacional, o uso do BERT se mostra mais adequado. Em cenários que demandam extração precisa de informações estruturadas em sequências complexas, como identificação de instituições, procedimentos, datas e valores, o BERT-CRF apresenta vantagens significativas.

Para ambientes reais de auditoria, nos quais coexistem restrições de tempo, recursos computacionais e necessidade de precisão, recomenda-se a adoção de um pipeline híbrido. Nesse arranjo, o BERT atua na filtragem inicial e categorização dos documentos, enquanto o BERT-CRF é acionado em etapas subsequentes, voltadas à análise aprofundada de documentos de maior risco. Essa estratégia equilibra desempenho, eficiência e confiabilidade, maximizando o valor informacional entregue aos auditores.

5.4 Limitações

Apesar dos resultados promissores, este estudo apresenta limitações que devem ser consideradas na interpretação dos achados. O corpus utilizado é composto por notícias relacionadas à auditoria do SUS, o que difere de relatórios técnicos internos e documentos administrativos oficiais, potencialmente mais complexos e heterogêneos. Além disso, o esquema de anotação

adotado e a definição das entidades influenciam diretamente o desempenho dos modelos, podendo limitar a generalização dos resultados.

Outra limitação refere-se à extrapolação dos achados para outros domínios, regiões ou variedades do português. Diferenças terminológicas, normativas e culturais podem impactar o comportamento dos modelos, exigindo adaptações adicionais para uso em contextos distintos.

5.5 Trabalhos Futuros

Como continuidade natural desta pesquisa, propõe-se a aplicação do pipeline desenvolvido em relatórios reais de auditoria do SUS, permitindo avaliar o desempenho das arquiteturas em documentos institucionais completos. Adicionalmente, o uso de técnicas de *active learning* surge como uma estratégia promissora para reduzir o custo de anotação de dados, direcionando o esforço humano para exemplos mais informativos.

Por fim, considerando os requisitos de governança, transparência e prestação de contas no setor público, trabalhos futuros devem investigar a incorporação de técnicas de Inteligência Artificial Explicável (XAI). A avaliação da interpretabilidade dos modelos é essencial para promover confiança institucional, viabilizar auditorias dos próprios sistemas automatizados e apoiar a adoção responsável dessas tecnologias em ambientes críticos de decisão.

Referências

Analytics Vidhya. *NLP Application:(NER) in Python with Spacy*. Blog post. Disponível em: <https://www.analyticsvidhya.com/blog/2021/06/nlp-application-named-entity-recognition-ner-in-python-with-spacy/>. Citado 3 vezes nas páginas 7, 16 e 17.

Answer.AI; LightOn. *Finally, a replacement for BERT: Introducing ModernBERT*. 2024. Answer.AI Blog. Disponível em: <https://www.answer.ai/posts/2024-12-19-modernbert.html>. Citado na página 42.

AQUINO, M. Auditorias no farmácia popular demorariam 20 anos com o atual efetivo [audits in the popular pharmacy would take 20 years with the current staff]. *Folha de S.Paulo*, 2022. In Portuguese. Disponível em: <https://www1.folha.uol.com.br/poder/2022/10/com-atual-efetivo-auditorias-no-farmacia-popular-demorariam-20-anos.shtml>. Citado na página 16.

Brasil. *Lei n. 8.689, de 27 de julho de 1993 [Law no. 8.689, of July 27, 1993]*. 1993. Diário Oficial da União. In Portuguese. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l8689.htm>. Citado na página 15.

Brasil. *Decreto n. 11.358, de 1° de janeiro de 2023 [Decree no. 11.358, of January 1, 2023]*. 2023. Diário Oficial da União. In Portuguese. Disponível em: <https://www.in.gov.br/en/web/dou/-/decreto-n-11.358-de-1-de-janeiro-de-2023-455829623>. Citado na página 15.

BRITO-SILVA, K.; BEZERRA, A. F. B.; TANAKA, O. Y. Direito à saúde e integralidade: uma discussão sobre os desafios e caminhos para sua efetivação [right to health and integrality: a discussion about the challenges and paths to its effectiveness]. *Interface - Comunicação, Saúde, Educação*, v. 16, n. 40, p. 249–260, Jan 2012. In Portuguese. Citado na página 15.

CAPES. *Portal de periódicos CAPES/MEC*. 2021. <https://www.periodicos.capes.gov.br>. Acesso em: 03 abril 2024. Citado na página 25.

CHENG, M. et al. Multi-task learning for chinese clinical named entity recognition with external knowledge. *BMC Medical Informatics and Decision Making*, v. 21, n. 372, 2021. Disponível em: <https://doi.org/10.1186/s12911-021-01717-1>. Citado na página 40.

Colaço Júnior, M. *AI for everyone: agents and experimental innovation without code [IA para a galera toda: agentes e inovação experimental sem código]*. [S.l.]: Kindle Direct Publishing, 2025. Em português. ISBN 978-65-01-24603-1. Citado 4 vezes nas páginas 7, 20, 21 e 43.

COSTA, T. D. et al. Análise do perfil das ações de auditoria realizadas a partir do sistema de auditoria do sistema único de saúde. *Revista de Administração em Saúde*, v. 21, n. 83, p. e290, 2021. Disponível em: <http://dx.doi.org/10.23973/ras.83.290>. Citado na página 16.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Stroudsburg,

PA, USA: Association for Computational Linguistics, 2018. p. 4171–4186. Citado 2 vezes nas páginas 7 e 18.

FENG, S.; MANMATHA, R.; MCCALLUM, A. Exploring the use of conditional random field models and hmms for historical handwritten document recognition. In: *Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL)*. Washington, DC, USA: IEEE Computer Society, 2006. p. 278–287. Disponível em: <http://www.cs.umass.edu/~mccallum/papers/handwritten-crf-feng06.pdf>. Citado na página 17.

FILHO, J. A. W. et al. The brWaC corpus: A new open resource for Brazilian Portuguese. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1686/>. Citado na página 42.

FONTES, R. S.; COLAÇO JÚNIOR, M.; PRADO, H. Sussurro - web detection of auditable events that represent risks to public health [sussurro - detecção na web de eventos auditáveis que representam riscos à saúde pública]. In: *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2023. p. 211–217. ISSN 2763-8987. Em português. Citado na página 42.

GUO, R.; ZHANG, H. Chinese medical named entity recognition based on roberta and adversarial training. *Huadong Ligong Daxue Xuebao/Journal of East China University of Science and Technology*, v. 49, n. 1, p. 144–152, 2023. Citado na página 40.

HE, J. et al. Prompt tuning in biomedical relation extraction. *Journal of Healthcare Informatics Research*, v. 8, p. 206–224, 2024. Disponível em: <https://doi.org/10.1007/s41666-024-00162-9>. Citado na página 40.

JOSEPHSON, B. W. et al. Uncle sam rising: Performance implications of business-to-government relationships. *Journal of Marketing*, v. 83, n. 1, p. 51–72, 2019. Disponível em: <https://doi.org/10.1177/0022242918814254>. Citado na página 28.

KIM, Y. et al. Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records. *Scientific Reports*, v. 10, n. 20265, 2020. Disponível em: <https://doi.org/10.1038/s41598-020-77258-w>. Citado na página 17.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. v. 2, 2007. Citado 2 vezes nas páginas 20 e 23.

KULSHRETHA, S.; LODHA, L. Performance evaluation of word embedding algorithms. *International Journal of Innovative Science and Research Technology*, Jaipur National University, v. 8, n. 12, p. 1555–1561, December 2023. Citado na página 17.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 282–289. Citado na página 19.

LI, G.; XU, A.; YUAN, L. Named entity recognition based on bi-lstm and crf-cel. In: IEEE. *2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. Xi'an, China, 2020. p. 337–341. Citado na página 16.

LightOn. *Better, faster, stronger knowledge retrieval and classification with ModernBERT*. 2024. LightOn Blog. Disponível em: <https://www.lighton.ai/lighton-blogs/better-faster-stronger-knowledge-retrieval-and-classification-with-modernbert>. Citado na página 42.

MILNE-IVES, M.; COCK, C. de; LIM, E. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, p. e20346, oct 2020. Disponível em: <https://www.jmir.org/2020/10/e20346/>. Citado na página 18.

NASEEM, U. et al. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC Bioinformatics*, v. 23, p. 144, 2022. Disponível em: <https://doi-org.ez20.periodicos.capes.gov.br/10.1186/s12859-022-04688-w>. Citado na página 40.

NVIDIA Blog. *What are large language models used for?* 2023. Blog post. Disponível em: <https://blogs.nvidia.com/blog/what-are-large-language-models-used-for/>. Citado na página 19.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. p. 8024–8035. Disponível em: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>. Citado na página 47.

PENSESUS - Portal de Informações de Saúde Pública da Fiocruz. *Direito à Saúde [Right to health]*. s.d. PENSESUS - Portal de Informações de Saúde Pública da Fiocruz. In Portuguese. Disponível em: <https://pensesus.fiocruz.br/direito-a-saude>. Citado na página 15.

PENSESUS - Portal de Informações de Saúde Pública da Fiocruz. *SUS*. s.d. PENSESUS - Portal de Informações de Saúde Pública da Fiocruz. In Portuguese. Disponível em: <https://pensesus.fiocruz.br/sus>. Citado na página 15.

RABINOWITZ, P. J. *Before Reading: Narrative Conventions and the Politics of Interpretation*. [S.l.]: Cornell University Press, 1987. ISBN 0801420105. Citado na página 39.

RAMSHAW, L. A.; MARCUS, M. P. *Text Chunking using Transformation-Based Learning*. 1995. Disponível em: <https://arxiv.org/abs/cmp-lg/9505040>. Citado na página 19.

RICHARDSON, W. S. et al. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, v. 123, n. 3, p. A12–A13, Nov-Dec 1995. Citado na página 24.

RODRIGUES, D. A. *Deep Learning e Redes Neurais Convolucionais: Reconhecimento Automático de Caracteres em Placas de Licenciamento Automotivo*. [S.l.: s.n.], 2018. Trabalho de Conclusão de Curso de Graduação, Universidade Federal da Paraíba. Citado na página 26.

SANTOS, C. M. d. C.; PIMENTA, C. A. d. M.; NOBRE, M. R. C. The pico strategy for the research question construction and evidence search. *Revista Latino-Americana de Enfermagem*, Escola de Enfermagem de Ribeirão Preto / Universidade de São Paulo, v. 15, n. 3, p. 508–511, Jun 2007. Disponível em: <https://doi.org/10.1590/S0104-11692007000300023>. Citado na página 24.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, Oxford University Press, v. 52, n. 3/4, p. 591–611, 1965. Citado na página 47.

SHYR, C. et al. Identifying and extracting rare diseases and their phenotypes with large language models. *Journal of Healthcare Informatics Research*, v. 8, p. 438–461, 2024. Disponível em: <https://doi.org/10.1007/s41666-023-00155-0>. Citado na página 40.

The JAMOVI project. *jamovi (Version 2.6) [Software]*. 2025. Statistical Software. Disponível em: <https://www.jamovi.org>. Citado na página 49.

VASWANI, A. et al. Attention is all you need. *arXiv*, abs/1706.03762, 2017. Manuscript submitted for publication. Disponível em: <https://arxiv.org/abs/1706.03762>. Citado 3 vezes nas páginas 7, 17 e 18.

WALLACE. *ModBERTBr*. 2024. Hugging Face Repository. Language model. Disponível em: <https://huggingface.co/wallacelw/ModBERTBr>. Citado na página 42.

WANG, H. et al. A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. *International Journal of Environmental Research and Public Health*, v. 18, n. 15, 2021. Cited by: 5; All Open Access, Gold Open Access, Green Open Access. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110631762&doi=10.3390%2fijerph18157776&partnerID=40&md5=4dc3539652f9060da47c12aadef93a62>. Citado na página 40.

XING, E. P.; GAO, Q.; CHEN, S. *12: Conditional Random Fields*. 2014. Lecture Notes for 10-708: Probabilistic Graphical Models, Carnegie Mellon University, Spring 2014. Disponível em: <https://www.cs.cmu.edu/~epxing/Class/10708-14/scribe_notes/scribe_note_lecture12.pdf>. Citado na página 17.

ZHANG, Q. et al. Named entity recognition method in health preserving field based on bert. *Procedia Computer Science*, v. 183, p. 212–220, 2021. ISSN 1877-0509. Proceedings of the 10th International Conference of Information and Communication Technology. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050921006189>. Citado na página 40.

Apêndices

APÊNDICE A – ESTRUTURA DO REPOSITÓRIO EXPERIMENTAL

A seguir, apresenta-se a estrutura de diretórios e arquivos gerada durante a execução da fase experimental. Esta organização reflete o processo de validação cruzada (k-fold) aplicado aos modelos de NER e ao classificador de texto.

```
dataset.conll                  (arquivo de entrada NER)
Amostra.csv                    (arquivo de entrada do classificador)

ner_model_artifacts/
|-- tag_to_id.json
|-- id_to_tag.json
|-- bert_crf_ner_model_fold_1/
|   |-- model.safetensors
|   |-- config.json
|   |-- tokenizer_config.json
|   `-- ...
|-- bert_crf_ner_model_fold_2/
|   `-- ...
`-- bert_crf_ner_model_fold_10/
    |-- model.safetensors
    |-- config.json
    |-- tokenizer_config.json
    `-- ...

ner_crf_output_fold_1/
|-- checkpoint-XXXX/
`-- ...

ner_crf_logs_fold_1/
`-- runs/
    `-- ...

text_classifier_output_fold_0/
|-- checkpoint-XXXX/
`-- ...

text_classifier_logs_fold_0/
`-- runs/
    `-- ...

my_text_classifier_model/
|-- model.safetensors
|-- config.json
|-- tokenizer_config.json
|-- special_tokens_map.json
|-- vocab.txt
|-- id_to_label.json
`-- ...
```

Figura 22 – Estrutura de diretórios do pipeline experimental