UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Doutorado Acadêmico

Diego Pinheiro da Silva

Reconhecimento de Entidades Nomeadas e Extração de Relações de registros de prontuários médicos para população de ontologia

São Leopoldo, 2023

Diego Pinheiro da Silva

RECONHECIMENTO DE ENTIDADES NOMEADAS E EXTRAÇÃO DE RELAÇÕES DE REGISTROS DE PRONTUÁRIOS MÉDICOS PARA POPULAÇÃO DE ONTOLOGIA

Tese apresentada como requisito parcial para a obtenção do título de Doutor, pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos – UNISINOS

Orientador:
Prof. Dr. Sandro José Rigo

Coorientador:
Profa. Dra. Renata Vieira

São Leopoldo
2023

S586r Silva, Diego Pinheiro.
Reconhecimento de entidades nomeadas e extração de relações de registros de prontuários médicos para população de ontologia / Diego Pinheiro Silva. – 2023.
171 f. : il. ; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2023.
"Orientador: Prof. Dr. Sandro José Rigo
Coorientadora: Profa. Dra. Renata Vieira."

1. Deep Learning. 2. EHR. 3. Processamento de linguem natural. 4. Reconhecimento de entidades nomeadas. 5. Extração de relações. I. Título.

CDU 004.4

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Silvana Dornelles Studzinski – CRB 10/2524)

Dedico esta pesquisa a minha família, professores, colegas e amigos, por todo apoio e motivação que mantiveram meus sonhos vivos nestes quatro anos de doutorado.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, irmãos, sogros e cunhados, heróis que me deram amor, incentivo e um apoio sem igual.

Ao Guilherme Frozza, por todo auxilio e companheirismo que me manteve firme nessa caminhada e que continua ao meu lado em todos os momentos.

A Luna, Pixie, Tibbers e Amora, por um amor incondicional do qual não há palavras no idioma que o descreva.

Ao Sandro Rigo, por todo ensinamento, por me guiar em todo caminho e mostrar que eu posso mais, exemplo de pessoa para mim.

A Marta Bez, por todo o seu exemplo, amor e amizade que me inspira a ser melhor, que eu possa um dia ser metade da pessoa que és.

A Blanda Mello, por juntos termos superado todos os obstáculos, um apoiando ao outro, sem perder a alegria e a inspiração para fazer ciência.

A Luana Rockenback, por toda parceria, conversa e admiração, que não mede forças para estar ao meu lado em todos os momentos.

A Interprocess, em especial ao Marco e José, pelo voto de confiança, apoio e crença em meu potencial.

Aos amigos e colegas, em especial aos que integram o grupo de pesquisa ”Computação Aplicada”, dos quais fizeram parte da minha formação e que sempre vão estar presentes em minha vida.

A todos os envolvidos, muito obrigado.

RESUMO

Existe um grande aumento do número de *Electronic Health Records* (EHRs) que acomodam dados não estruturados, tais como textos e observações em linguagem natural. Consequentemente, cresce o interesse em utilizar tais dados para promover melhorias na saúde. A análise manual desses dados não é viável, devido ao grande volume existente, cuja tendência é o aumento contínuo. Sendo assim, há necessidade de uma abordagem que possibilite que essas informações sejam automaticamente estruturadas para que possam auxiliar profissionais da saúde na análise de dados, recomendação de tratamentos, diagnósticos de doenças, entre outros. Uma avaliação da literatura na área permitiu identificar demandas para o tratamento deste problema na língua portuguesa, bem como a existência de um número ainda reduzido de trabalhos com dados reais da área de saúde. Também foi identificado como oportunidade de pesquisa a utilização de recursos baseados na arquitetura *Transformers* e uso dos resultados para a estruturação de dados em ontologias. Neste contexto, este trabalho tem como objetivo o desenvolvimento de um modelo para o processamento de dados não estruturados de EHRs para apoiar a atividade de atualização de uma ontologia. As contribuições dessa pesquisa estão presentes em dois aspectos relacionados. Um deles situa-se no apoio ao desenvolvimento de aplicações em sistemas EHR para oncologia, através da ampliação da capacidade desses sistemas para o uso dos dados não estruturados. O segundo aspecto de contribuição está relacionado com a experimentação e proposta de avanços na computação em abordagens para reconhecimento de entidades e extração de relações, bem como sua integração com uma ontologia. Este trabalho foi realizado no contexto de um estudo de caso em uma empresa que atua na área de Oncologia. Foram efetuadas análises detalhadas de um sistema amplamente utilizado em EHRs de clínicas de oncologia. A partir desta análise foi gerado um dos diferenciais do trabalho, através da composição de *datasets* inéditos de entidades e relações de evoluções médicas, contendo 1.622 documentos anotados, sendo 146.769 entidades e 111.716 relações. Outro diferencial do trabalho está relacionado com a adaptação de uma ontologia de domínio para representar os dados estruturados deste estudo de caso. Por fim, foram conduzidos experimentos com abordagens para extrair entidades e relações em texto, alcançando resultados como 78,24% de precisão no domínio de exames e 72,87% no domínio de diagnósticos. Além disso, uma ontologia com foco em oncologia foi construída e integrada ao modelo, englobando aproximadamente 181 classes, 14 propriedades de dados, 12 propriedades de objetos e mais de 200 indivíduos. Especialistas da área de saúde avaliaram o modelo, obtendo uma taxa de acerto de 73,52% em relação a análise deles, e a pesquisa de usabilidade mostrou uma excelente aceitação. Destaca-se como diferencial do trabalho o treinamento de modelos com uso de dados reais de oncologia e a construção de uma base de conhecimento através da ontologia.

Palavras-chave: Deep Learning. Processamento de Linguagem Natural. Reconhecimento de Entidades Nomeadas e Extração de Relações. Ontologia. EHR.

ABSTRACT

There has been a significant increase in the number of Electronic Health Records (EHRs) that accommodate unstructured data, such as text and natural language observations. Consequently, there is a growing interest in using this data to promote improvements in health. Manual analysis of these data is not feasible due to the large volume, which continues to increase. Therefore, there is a need for an approach that automatically structures this information, enabling it to assist health professionals in data analysis, treatment recommendations, disease diagnoses, among other applications.An evaluation of the literature in this area has identified demands for addressing this problem in Portuguese. However, there are still a limited number of studies with real data from the health sector. A research opportunity identified is the use of resources based on the Transformers architecture and the application of the results for data structuring in ontologies.In this context, this work aims to develop a model for processing unstructured data from EHRs to support the activity of updating an ontology. The contributions of this research are present in two related aspects. Firstly, it aims to support the development of applications in EHR systems for oncology by enhancing their capacity to utilize unstructured data. Secondly, the research focuses on experimenting and proposing advances in computing approaches for entity recognition and relations extraction, as well as integrating them with an ontology. The study was carried out as a case study in a company operating in the field of Oncology. Detailed analyses of a widely used system in EHRs of oncology clinics were conducted. As a result of this analysis, one of the distinctive features of the work is the creation of unpublished datasets of entities and relations of medical evolutions, containing 1,622 annotated documents, comprising 146,769 entities and 111,716 relations. Another unique aspect of the work is the adaptation of a domain ontology to represent the structured data of this case study. Finally, experiments were conducted with approaches to extract entities and relations in text, achieving results such as 78.24% accuracy in the exams domain and 72.87% in the diagnostics domain. In addition, an ontology focused on oncology was built and integrated into the model, encompassing approximately 181 classes, 14 data properties, 12 object properties, and over 200 individuals. Healthcare specialists evaluated the model, obtaining a 73.52% accuracy rate in relation to their analysis, and the usability research showed excellent acceptance. The training of models using real oncology data and the construction of a knowledge base through ontology stands out as a differential of the work.

Keywords: Deep Learning. Natural Language Processing. Named Entity Recognition and Relation Extraction. Ontology. EHR.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ADAM	*Adaptive Moment Estimation*
API	*Application Programming Interface*
BERT	*Bidirectional Encoder Representations from Transformer*
BILSTM	*Bidirectional Long Short-Term Memory*
BILSTM-CRF	*Bidirectional Long Short-Term Memory - Conditional Random Field*
BIRCH	*Balanced Iterative Reducing and Clustering using Hierarchies*
BO-LSTM	*Back-off - Long Short-Term Memory*
BoW	*Bag-of-Words*
BRACIS	*Brazilian Conference on Intelligent Systems*
C	Concordo
CBM	*Computers in Biology and Medicine*
CMR	Classificação multirrótulo
CIE	Critérios de Inclusão/Exclusão
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CT	Concordo Totalmente
D	Discordo
DGO	*Dataset* Gemed Onco
DGO-CD	*Dataset* Gemed Onco - Características do Diagnóstico
DGO-E	*Dataset* Gemed Onco - Exames
DAI	Doutorado Acadêmico para Inovação
DICOM	*Digital Imaging and Communications in Medicine*
DL	*Deep Learning*
DDI	*Drug-Drug Interaction*
DT	Discordo Totalmente
EHR	*Electronic Health Records*
ER	Extração de Relações
FAU	*Universidade Friedrich-Alexander de Erlangen-Nürnberg*
GLN	Geração de Linguagem Natural
GNN	*Graph Neural Network*
HL7	*Health Level Seven International*
IA	Inteligência Artificial
ICD	*International Classification of Diseases*
ICENCO	*International Computer Engineering Conference*
ID	Identificador
IOB	*Beginning, Inside, Outside*

JSON	*JavaScript Object Notation*
KNN	*K-Nearest Neighbors*
LN	Linguagem Natural
LSTM	*Long Short-Term Memory*
LGPD	Lei Geral de Proteção de Dados
MDS	*Multidimensional Scaling*
ML	*Machine Learning*
MLP	*Multilayer Perceptron neural network*
MAE	*Multi-document Annotation Environment*
NDC	Nem Discordo nem Concordo
PG	Perguntas Gerais
PF	Perguntas Focalizadas
PCA	*Principal Component Analysis*
PLN	Processamento de Linguagem Natural
RDF	*Resource Description Framework*
REN	Reconhecimento de Entidades Nomeadas
RN	Rede Neural
RNC	Rede Neural Convolucional
RNR	Rede Neural Recorrente
RNP	Rede Neural Profunda
SGD	*Stochastic Gradient Descent*
SGO	Sistema Gemed Onco
SL	*Super Learner*
SQL	*Structured Query Language*
SVM	*Support Vector Machines*
UFCSPA	Universidade Federal de Ciências da Saúde de Porto Alegre
UEVORA	*Universidade de Évora*
UMLS	*Unified Medical Language System*
UNISINOS	Universidade do Vale do Rio dos Sinos
US	Ultrassom
WKNN	*Weighted K-nearest Neighbor*
W3C	*World Wide Web Consortium*
XML	*Extensible Markup Language*
W2V	*Word2Vec*

SUMÁRIO

1 INTRODUÇÃO

Os equívocos no diagnóstico médico não são um problema recente, como descrito já na pesquisa realizada há mais de uma década por Bhasale (1998). A maioria dos erros de diagnóstico clínico estão ligados aos erros de julgamento, particularmente na formação e avaliação de hipóteses diagnósticas, em parte devido à redução do acesso às informações necessárias (ZURYNSKI et al., 2017) (SUTHERLAND et al., 2020). O prontuário do paciente é um documento essencial para garantir uma assistência integral e continuada ao enfermo, armazenando o histórico da sua saúde (BENÍCIO, 2020). Conforme Shickel et al. (2017), os registros médicos foram projetados principalmente para arquivar informações clínicas e administrativas, embora muitas novas aplicações usem esses dados médicos em uma variedade de tarefas, tais como a previsão de doenças crônicas (GOLDSTEIN et al., 2017), decomposição de risco clínico (HUANG; DONG; DUAN, 2015), tratamento da depressão (PERLIS et al., 2012) (ADEKKANATTU et al., 2023), entre outros.

A integração de fontes heterogêneas e análises de similaridade de pacientes auxiliam a compor um modelo de suporte para um diagnóstico clínico mais amplo e para apoiar a vigilância automatizada (PERER; WANG; HU, 2015) (BEG et al., 2022). Uma vez que a informação de saúde é estruturada, ela pode ser usada em conjunto com bancos de dados abertos e conectados para alimentar sistemas de recomendação de diagnóstico clínico (SCHWERTNER, 2020), (BRAGA et al., 2019).

Existe um grande aumento do número de prontuários em formato digital que acomodam dados não estruturados, com textos e observações em Linguagem Natural (LN) (TOPAZ et al., 2016) (SOLARES et al., 2020). A crescente adoção do *Electronic Health Records* (EHR) e o consequente aumento no volume de dados não estruturados disponíveis, bem como o interesse em utilizar tais dados para promover melhorias na saúde e o desenvolvimento de pesquisas, consistem em evidências de que o tratamento dos dados em LN contidos em tais registros consiste em um desafio e ao mesmo tempo em uma oportunidade (DHOLE; UKE, 2014) (SOLARES et al., 2020).

As informações médicas têm sido historicamente extraídas dos registros do paciente através de especialistas clínicos. Essa abordagem tem limitações de escalabilidade, além de ser demorada, trabalhosa e cara (KOLECK et al., 2019). A disponibilidade de EHRs para reutilização de dados secundários criou uma oportunidade para o Processamento de Linguagem Natural (PLN) ser usado para aproveitar o potencial das narrativas de texto livre para aplicar soluções que possam apoiar os profissionais em saúde (KREIMEYER et al., 2017) (JUHN; LIU, 2020).

Segundo Névéol, Zweigenbaum et al. (2017), a aplicação do PLN é considerada como um aspecto fundamental para fomentar a pesquisa em apoio à decisão clínica, bem como apoiar a análise automatizada de dados médicos (BECKER; BÖCKMANN, 2017) (SOUZA et al., 2020). O tratamento das informações não estruturadas contidas no EHR está associado à possibilidade de avanços significativos na pesquisa nessa área (ZHAO et al., 2018). Os dados

contidos no EHR, como relatórios clínicos e observações médicas, podem ser usados para registros de doenças, estudos epidemiológicos, vigilância de segurança de medicamentos, ensaios clínicos, auditorias de saúde e muito mais (FORD et al., 2016) (SOLARES et al., 2020).

O desenvolvimento de modelos que possibilitem uma melhor interpretação de textos em LN pelos sistemas computacionais, através da identificação e representação de seu significado, possibilita o processamento automatizado, com maior precisão, da semântica de bases textuais. Este tipo de tratamento de dados é cada vez mais uma necessidade, pois existe um consenso sobre como a aplicação das tecnologias de informação e comunicação nos sistemas de saúde pode aumentar sua eficiência, reduzir custos e auxiliar na melhoria da vida das pessoas, fornecendo subsídios ao procedimento diagnóstico e possibilitando a prevenção de doenças (AMATO et al., 2016) (MUSEN, 2015) (BADAR; HARIS; FATIMA, 2020). Como uma das consequências dessas capacidades, os sistemas integrados de informação médica estão se tornando uma parte essencial dos sistemas de saúde (DUAN; STREET; XU, 2011). O conceito de EHR é um exemplo nesse contexto, fornecendo dados relevantes sobre os pacientes em formato digital e de forma integrada (HEART; BEN-ASSULI; SHABTAI, 2017) (SOLARES et al., 2020).

A análise manual de dados não é possível de ser feita devido ao grande volume de dados existente, cuja tendência é o aumento continuado (NGIAM; KHOR, 2019). Sendo assim, há necessidade de uma abordagem que possibilite que essas informações sejam obtidas, comparadas e relacionadas automaticamente através de diferentes processos em execução rotineira em instituições de saúde. Isso é relevante para setores internos às instituições de saúde, bem como para aqueles setores externos situados junto às entidades governamentais, pois permite a estes setores que suas decisões sobre orçamento, análise de qualidade, planejamento e análise de tendências ou de situações críticas, entre outras, sejam tomadas com base em um cenário mais consistente (CAMMAROTA et al., 2020).

A área de Inteligência Artificial (IA) vem sendo utilizada de forma crescente na área médica, sendo integrada em diversos trabalhos junto com outras tecnologias, buscando atender necessidades da área da saúde. Em particular, abordagens de *Machine Learning* (ML) vem sendo integradas à prática clínica, com aplicações que variam de processamento de dados pré-clínicos, assistência a diagnósticos, tomada de decisão de tratamento e alerta precoce como parte da prevenção primária e secundária (ADLUNG et al., 2021).

Esta abordagem permite o reconhecimento de padrões nos dados e, com a abundância de conjuntos de dados disponível, vem sendo aplicada para extrair dados relevantes para aplicações em saúde (MAHESH, 2020), (HUTTER; KOTTHOFF; VANSCHOREN, 2019). De forma mais recente, observa-se a escalada do uso de abordagens de *Deep Learning* (DL), possibilitando que sejam obtidos resultados relevantes para tratamento de diferentes tarefas, desde a análise de imagens até extração de informações de texto (DAWUD; YURTKAN; OZTOPRAK, 2019),(CUOCOLO et al., 2020).

Observam-se trabalhos integrando recursos como extração de informação, uso de ontologias de saúde, aplicação de modelos vetoriais de linguagem para tratamento de textos médicos, uso

de Redes Neurais (RN) e sistemas de recomendação de saúde (ZHU; SOBHANI; GUO, 2016) (LEE et al., 2016) (FENG; XIANG; ZHOU, 2016) (ALEMI et al., 2012) (YANG et al., 2019a) (CHRISTOPOULOU et al., 2020).

No contexto de extração de informações e de relações, existe um cenário de rápido avanço nas técnicas empregadas. O trabalho de Ford et al. (2016) destacou que entre projetos que envolvem extração de informações dos registros eletrônicos do paciente e classificação automática de doenças, que um percentual de 67% das obras utiliza abordagens baseadas em regras. Como os sistemas baseados em regras são limitados (DHOLE; UKE, 2014), surge a necessidade do uso do PLN devido à sua capacidade de identificar um conceito expresso de muitas maneiras diferente, incluindo sinônimos e generalizações ou especificações de contexto (GONZALEZ-HERNANDEZ et al., 2017) (LAI et al., 2022).

A utilização de modelos *bidirectional encoder representations from Transformers* (BERT) tem apresentado bons resultados em muitas tarefas de PLN, tais como o Reconhecimento de Entidades Nomeadas (REN) e Extração de Relações (ER). Desde então, *Transformers* têm sido explorados em diferentes domínios de conhecimento, avaliando quanto o modelo genérico pode trazer de resultados significativos em relação às técnicas usualmente aplicadas, e quanto o *fine-tuning* em *datasets* específicos pode ser vantajoso (LEE et al., 2020) (LI et al., 2020a) (MORGAN; RANASINGHE; ZAMPIERI, 2021). Em particular, Li et al. (2020b) exploraram a eficácia dos modelos baseados em BERT para a tarefa de normalização de entidades com foco nos domínios biomédico e clínico. A proposta utiliza uma abordagem conjunta de *fine-tuning* sequenciais, onde o modelo BERT, treinado com dados Wikipédia, foi inicializado com um *fine-tuning* utilizando publicações da base PubMed, resultando no modelo BioBERT (domínios biomédicos e clínicos). Os autores Xue et al. (2019) também exploraram a utilização de *Transformers*, mas com foco na mineração de dados em domínio biomédico.

Apesar de existirem trabalhos explorando diferentes formas de integração de dados não estruturados, são escassos na literatura os trabalhos voltados para a área da saúde que explorem técnicas de aprendizagem profunda de forma a contemplar dados não estruturados e estruturados, sendo estes os aspectos de originalidade que serão explorados nesta pesquisa.

Este trabalho tem como objetivo o desenvolvimento de um modelo para o processamento de dados não estruturados em saúde e sua utilização para população de uma ontologia. O resultado pode ser aplicado no acompanhamento automatizado de subsídios ao diagnóstico clínico, análise de qualidade, projeção de cenários para orçamento, identificação de tendências ou epidemias, predição de cenários, recomendação de diagnósticos ou protocolos, entre outras atividades possíveis.

Atualmente existem iniciativas nesta área (YANG et al., 2019a) (LIN et al., 2020) mas ainda observa-se a necessidade de avanços na capacidade dos sistemas atuais de extrair, representar e integrar dados, especialmente quando se trata de dados não estruturados e da integração de dados multimodais. Para isso, serão realizadas integrações de técnicas de DL e PLN. Desta forma, espera-se obter avanços na utilização do conjunto de dados médicos existente atualmente

em registros médicos e em outras fontes.

Este trabalho traz contribuições em dois contextos integrados. Um deles situa-se no apoio ao desenvolvimento de aplicações com sistemas EHR para oncologia, ampliando a capacidade desses sistemas para o uso dos dados não estruturados. O segundo aspecto de contribuição está relacionado com a experimentação e proposta de avanços na computação em algoritmos para REN e ER. Além desses, faz parte dessas contribuições a integração de dados estruturados em uma ontologia.

O estado da arte mostra trabalhos relevantes e com bons resultados quanto ao uso do modelo BERT, bem como de arquiteturas recorrentes. Sendo assim, foram utilizado modelos de BERT e BiLSTM (*Bidirectional Long Short-Term Memory*) para REN e ER. Os modelos conhecidos exploram pouco os aspectos linguísticos e semânticos (MORGAN; RANASINGHE; ZAMPIERI, 2021) (LI et al., 2020a) (SCHNEIDER et al., 2020). As situações de uso que estão previstas são a aplicação na saúde na análise de dados oncológicos. Além deste fator, são escassos os recursos para tratamento deste problema em língua portuguesa, o que representará um ganho adicional promovido por esta pesquisa.

1.1 Questão de Pesquisa

Neste contexto, a questão de pesquisa adotada neste trabalho foi definida da seguinte forma: Quais elementos devem compor uma abordagem que permita extrair relações de dados não estruturados, estruturá-los e gerar uma base de conhecimento para uso na área da saúde?

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um modelo que possibilite o REN e ER de dados não estruturados gerados em sistemas EHR e estruturá-los em uma base de conhecimento através de uma ontologia.

1.2.2 Objetivos Específicos

1. Realizar estudos sobre sistemas médicos, PLN, DL e ontologias.
2. Avaliar trabalhos correlatos disponíveis na literatura quanto aos detalhes da sua abordagem de ER.
3. Realizar experimentos com dados na área da saúde.
4. Anonimizar e anotar um *dataset* em português na área da oncologia.

5. Descrever um modelo para realização da ER.

6. Desenvolver um protótipo para avaliação do modelo quanto à integração dos dados em ontologias.

7. Avaliar os resultados de experimentos com especialistas de um domínio.

1.3 Metodologia

A metodologia para o desenvolvimento da pesquisa tem como procedimento inicial a revisão do estado da arte e trabalhos relacionados para uma atualização sobre técnicas, projetos e métodos. Segue-se então o levantamento de requisitos, etapa na qual são identificados e validados os principais componentes da infraestrutura implementada na próxima etapa, qual seja, a definição detalhada de um novo modelo e a sua prototipação. Após esta etapa são realizados procedimentos de avaliação e de análise dos resultados. A fase de levantamento do estado da arte foi realizada, bem como o levantamento de requisitos e dos principais problemas e necessidades associados com a integração de dados. A partir destas duas atividades consideram-se consolidados os subsídios para o detalhamento do modelo construído, em contato com as parceiras acadêmicas e da área da saúde.

Essa pesquisa conta com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) através do Programa Doutorado Acadêmico para Inovação (DAI), reiterando a relevância da parceria entre universidades e empresas e trabalhando com o interesse de ampliar a capacidade de gerar inovação no setor empresarial. Com isso, este trabalho conta com a parceria com a empresa Interprocess[1], uma empresa especializada em soluções tecnológicas para o gerenciamento de dados e processos no segmento da saúde, do Portal de Inovação da UNISINOS (Universidade do Vale do Rio dos Sinos) e do CNPQ.

Atualmente, o grupo de pesquisa deste projeto conta com a colaboração da UFCSPA (Universidade Federal de Ciências da Saúde de Porto Alegre), a UEVORA (Universidade de Évora) e a FAU (Friedrich-Alexander-Universität Erlangen-Nürnberg). O projeto também conta com parceiras institucionais na área da Saúde, atualmente vinculadas formalmente ao projeto em desenvolvimento (CNPQ/DT). Estas instituições são a Secretaria Municipal de Saúde do município de Porto Alegre; o Grupo Hospitalar Conceição e a Irmandade da Santa Casa de Misericórdia de Porto Alegre. Estas instituições participam do projeto fornecendo a demanda inicial, o caso de estudo e proporcionando a avaliação dos resultados. Considera-se ainda a perspectiva de utilização futura dos resultados nestas instituições, como contrapartida do projeto.

Quanto aos procedimentos técnicos, considera-se o trabalho como bibliográfico e experimental (WAZLAWICK, 2009). O procedimento bibliográfico foi realizado através de pesquisas em trabalhos acadêmicos, livros, artigos e publicações referente a sistemas médicos, Ontologias, DL e técnicas de IA. O levantamento bibliográfico serviu de base para a definição da proposta

[1] https://www.interprocess.com.br

de modelo para auxiliar especialistas em saúde. Foram analisados os modos de extração, dados, técnicas e demais conceitos a serem inseridos na base de conhecimento.

Os objetivos deste trabalho são caracterizados como uma pesquisa do tipo exploratória, que visa o estudo de determinada área de ação, suas premissas e necessidades para que, a partir disso, possa ser desenvolvido o modelo que extraiu relações de dados não estruturados para população de base de conhecimento. Caracteriza-se o trabalho como de natureza exploratória e aplicada, pois atua no desenvolvimento de um modelo que busca solucionar problemas concretos encontrados na literatura e com especialistas quando se trata de estruturar dados. Neste conjunto de ações, houve a colaboração com profissionais das áreas de saúde e linguística, para o detalhamento de aspectos das aplicações práticas delimitadas pela pesquisa proposta.

A forma de abordagem é qualitativa, pois através da análise de especialistas e da literatura, buscou-se uma ampla compreensão de como são extraídas relações de dados não estruturados. Além disso, a medida em que a análise ocorreu, a mesma foi sendo validada junto a especialistas da área. Caracteriza-se o trabalho desenvolvido como experimental (KÖCHE, 2016), sendo realizada a análise e proposto o desenvolvimento de um modelo que visa solucionar o problema prático específico, observando como a tecnologia pode auxiliar na estruturação do conhecimento e o suporte que prestará para os especialistas quanto ao conhecimento de domínio.

Com o protótipo construído, realizou-se estudos de caso junto com as instituições ligadas ao contexto da Saúde. Os estudos de caso concretizaram-se através de aplicação dos métodos desenvolvidos junto com volumes de dados consistentes para atividades de avaliação quanto às métricas relevantes, além de possibilitarem a verificação da pertinência das soluções para as necessidades das instituições de saúde. Os resultados dos estudos de caso possibilitam avançar com melhorias no modelo e no protótipo, a escrita de relatórios técnicos e artigos científicos, bem como a disponibilização de produtos de softwares ou serviços de software.

Foram realizados dois tipos de avaliações do modelo. A primeira, denominada computacional, ocorreu a partir do treinamento do modelo. Métricas como Precisão, *Recall*, *F1 score*, Suporte, Acurácia, *Macro AVG* e *Weighted AVG* foram apuradas. A segunda, baseada em especialistas na área da saúde, ocorreu com base em cenários. Tendo em vista a área da saúde, o modelo e seus resultados foram validados com especialistas na área da oncologia, em estudos de casos e cenários aplicados ao Sistema Gemed Onco[1] (SGO). Foram realizados experimentos com coleta de dados em questionários e acompanhamentos com os profissionais para validar sua percepção quanto ao uso do modelo.

1.3.1 Escopo deste trabalho

Este trabalho se concentra na área de oncologia e tem como objetivo principal a aplicação de técnicas de REN e ER em dados não estruturados provenientes do SGO, sendo disponibilizado

[1] https://www.interprocess.com.br/gemed/

pela empresa Interprocess. O SGO é amplamente utilizado em instituições de saúde especializadas em tratamento oncológico. Sua utilização promove o armazenamento de várias informações relevantes, tais como diagnósticos, tratamentos, exames e respostas clínicas registrados por profissionais de saúde em formato de texto livre. Isso resulta em dados não estruturados que dificultam a análise e o aproveitamento pleno das informações disponíveis. A parceria com a Interprocess permitiu a utilização de dados reais, garantindo a relevância e a aplicabilidade dos resultados obtidos.

Foram aplicadas técnicas de PLN para identificar as entidades de domínios presentes nos textos, como tipos de exames, características de diagnóstico e outros conceitos relevantes. Além disso, foram extraídas as relações semânticas entre essas entidades, a fim de estruturar os dados em uma ontologia desenvolvida no contexto oncológico. Outro aspecto importante desta pesquisa é a colaboração com estudantes da área de saúde da UNISINOS para realizar a anotação de dados necessária para treinar e validar o modelo desenvolvido. Essa parceria proporcionou conhecimentos especializados para a identificação correta de conceitos da saúde presentes nos textos. Adicionalmente, a empresa privada Comsentimento[2] também se uniu ao processo de anotação. Essa parceria permitiu agilizar o processo de anotação, garantindo maior quantidade e qualidade dos dados anotados, essenciais para o treinamento adequado. O capítulo 4 apresenta os detalhes completos sobre o processo e metodologia desta pesquisa.

Em relação ao compartilhamento dos dados e recursos gerados nesta pesquisa, é importante destacar que, devido à parceria com a empresa Interprocess e ao DAI, os *datasets*, dados e outros recursos não serão disponibilizados publicamente. Conforme as regras do edital do DAI, esses materiais são considerados privados e de propriedade da empresa.

1.4 Contribuições

Neste contexto, este trabalho tem como objetivo o desenvolvimento de um modelo para o processamento de dados não estruturados de EHRs para apoiar a atividade de atualização de uma ontologia.

As contribuições dessa pesquisa estão presentes em dois aspectos relacionados. Um deles situa-se no nível da aplicação em sistemas EHR para oncologia, através da ampliação da capacidade desses sistemas para o uso dos dados não estruturados. O segundo aspecto de contribuição está relacionado com a experimentação e proposta de avanços na computação em abordagens para REN e ER, bem como sua integração com uma ontologia.

Este trabalho foi realizado no contexto de um estudo de caso em uma empresa que atua na área de Oncologia. Foram efetuadas análises detalhadas de um sistema amplamente utilizado em EHRs de clínicas de oncologia. A partir desta análise foi gerado um dos diferenciais do trabalho que é a indicação de componentes que possibilitam a utilização dos dados internos do sistema, em formato não estruturado, para geração de uma base de conhecimento.

[2]https://comsentimento.com.br/

Um segundo diferencial do trabalho é a composição de *datasets* inéditos e com dados reais de entidades e relações de evoluções médicas, contendo 1.622 documentos anotados, sendo 146.769 entidades e 111.716 relações.

Outro diferencial do trabalho está relacionado com o estudo e adaptação de uma ontologia de domínio para representar os dados estruturados deste estudo de caso.

Por fim, foram realizados experimentos com abordagens para REN e ER em texto, com resultados promissores para o modelos utilizados, cujo diferencial está localizado no treinamento de dados reais de oncologia e a construção de uma base de conhecimento através da ontologia.

1.5 Organização do Texto

No Capítulo 2 são apresentados os conceitos fundamentais sobre registros médicos e prontuários, PLN, DL e ontologias. Uma revisão sistemática sobre ER e DL é descrita no Capítulo 3. A abordagem proposta neste trabalho é apresentada no Capítulo 4. No Capítulo 5, são descritos resultados de experimentos. Por fim, as conclusões e os trabalhos futuros são descritos no Capítulo 6.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os temas utilizados como embasamento para esta tese, organizados em EHR, PLN, DL e Ontologias. Cada seção apresenta os conceitos gerais e também discute elementos da literatura em uma contextualização da sua aplicação em torno do tema central desta tese.

2.1 Registros médicos e prontuários eletrônicos (*Electronic Health Records* - EHR)

Os EHRs são uma rica fonte de dados para pesquisa, melhoria da qualidade e gerenciamento da população. A quantidade de dados disponíveis em EHRs vem crescendo exponencialmente e, portanto, métodos analíticos computacionais eficientes e eficazes são necessários para realizar plenamente seus benefícios potenciais (RUDRAPATNA; BUTTE, 2020) (WIKSTRÖM et al., 2019) (TURCHIN; MASHARSKY; ZITNIK, 2023). Um componente importante dos EHRs são os documentos narrativos. Eles contêm uma grande quantidade de dados não encontrados em tabelas de bancos de dados estruturados, tais como avaliações diferenciadas da condição do paciente, raciocínio por trás da escolha do tratamento, documentação de discussões paciente-profissional, etc (TURCHIN; MASHARSKY; ZITNIK, 2023) (SCHWERTNER et al., 2019).

Anotações clínicas ou registros médicos são essenciais para a continuidade do atendimento aos pacientes. Registros médicos adequados permitem reconstruir as partes essenciais de cada contato com o paciente. Portanto, eles devem ser abrangentes o suficiente para permitir a continuidade de trabalho entre os profissionais de saúde (SOCIETY, 2018). Escrever notas clínicas claras e descritivas é muito diferente da maioria dos outros tipos de escrita. O objetivo em escrever notas clínicas serve para criar um registro preciso e informativo do tratamento e do progresso do paciente (HODGES, 2011).

Os registros médicos devem resumir os principais detalhes de cada contato com o paciente, devendo incluir: achados clínicos relevantes, decisões tomadas, ações acordadas, quem está tomando as decisões, quem está concordando com as ações, as informações que são fornecidas aos pacientes, medicamentos prescritos ou outra investigação ou tratamento, quando o registro foi feito e, por fim, quem está fazendo o registro e quando. Em ocasiões subsequentes, os registros médicos também devem observar o progresso do paciente, achados no exame, medidas de monitoramento e acompanhamento, detalhes de consultas por telefone, detalhes sobre acompanhantes presentes e qualquer instância em que o paciente recusou-se a ser examinado ou obedecer ao tratamento (SCHWERTNER, 2020) (ALFATTNI; PEEK; NENADIC, 2020).

Tais registros podem abranger uma ampla variedade de materiais, incluindo: notas manuscritas, registros computadorizados, correspondência entre profissionais de saúde, relatórios laboratoriais, registros de imagens (incluindo raios-x), fotografias, vídeo e outras gravações, impressos de equipamentos de monitoramento e comunicação de texto ou e-mail com os pacientes. Quando um profissional de saúde precisa apresentar um paciente para discutir nas rondas,

a equipe (incluindo enfermeiras, residentes e médicos assistentes) ouvirá a apresentação para ter uma ideia do que está acontecendo com o paciente. As informações precisam ser corretas, bem organizadas e concisas (SCHWERTNER, 2020) (KOLECK et al., 2019).

Um recente aumento da disponibilidade de dados de saúde coletados rotineiramente facilitou o fornecimento de suporte baseado em evidências orientado por dados para atividades clínicas e de pesquisa. Esses dados são principalmente coletados ao longo do tempo e inseridos em EHR (RAZZAQUE; HAMDAN, 2021). O EHR de uma grande organização médica pode capturar as transações médicas de mais de 10 milhões de pacientes ao longo de uma década. Uma única hospitalização sozinha normalmente gera em torno de 150.000 dados. Os benefícios potenciais derivados desses dados são significativos. No total, um EHR desta escala representa 200.000 anos de sabedoria médica e 100 milhões de anos de dados de resultados de pacientes, cobrindo uma grande quantidade de doenças e condições raras (ESTEVA et al., 2019).

Os dados são registrados em EHRs em diferentes formas de dados estruturados, semiestruturados e não estruturados. As partes estruturadas geralmente consistem em informações codificadas e numéricas com registro de data e hora, incluindo dados demográficos dos pacientes, visitas clínicas, diagnósticos, biomarcadores, resultados laboratoriais, tratamentos e imunizações (RAZZAQUE; HAMDAN, 2021), enquanto as partes não estruturadas consistem em texto livre ou dados narrativos semiestruturados, imagens clínicas, etc.

O texto livre em LN é amplamente utilizado para registrar detalhes sobre o histórico médico do paciente (SCHWERTNER, 2020) (DEMNER-FUSHMAN; CHAPMAN; MCDONALD, 2009), pois captura a explicação do profissional da saúde em relação a condição do paciente, duração dos sintomas, hipóteses diagnósticas rejeitadas, preferências do paciente e experiência de tratamento. Esforços significativos foram feitos nos últimos anos para identificar e extrair automaticamente as principais informações clínicas (por exemplo, sintomas, diagnósticos e tratamentos) da narrativa de texto livre (ALFATTNI; PEEK; NENADIC, 2020). No entanto, tais eventos clinicamente relevantes precisam ser colocados no contexto temporal para ajudar a entender a ordem cronológica dos procedimentos clínicos (MATE et al., 2019), facilitar e melhorar o diagnóstico, por exemplo, observando a ordem em que os sintomas se desenvolvem (CHOI et al., 2017) e também melhorar o tratamento, por exemplo, anotando o tempo de tomar a medicação (CHOI et al., 2016a).

Os grandes bases de dados EHR, no entanto, muitas vezes não são acessíveis por vários motivos, incluindo o número limitado de casos para doenças novas ou raras; dificuldade na limpeza de dados e anotação, especialmente se coletados de fontes diferentes; e questões de governança que dificultam a aquisição de dados (RASMY et al., 2021). Outro ponto que dificulta o acesso é em relação a Lei Geral de Proteção de Dados (LGPD), criada em 2018 no Brasil, que garante a proteção quanto ao uso dos dados sem justificativa, autorização e consenso ou anonimização dos dados do paciente.

Tendo em vista que o câncer é uma das comorbidades crescentes nas estatísticas do Ministério da Saúde, a seguir serão detalhados aspectos de prontuários médicos e registros de saúde

ligados a oncologia. O processo de formação do câncer é chamado de carcinogênese ou oncogênese e, em geral, acontece lentamente, podendo levar vários anos para que uma célula cancerosa prolifere-se e dê origem a um tumor visível (BRASIL, 2021). O câncer é a causa mais comum de morte em países desenvolvidos e estima-se que o número de casos aumentará ainda mais com o envelhecimento da população. No Japão, quase 1 milhão de pessoas são diagnosticadas com câncer e quase 400.000 morrem da doença todos os anos. A pesquisa do câncer continuará, portanto, a ser uma das principais prioridades para salvar vidas na próxima década. (SHIMIZU; NAKAYAMA, 2020).

O Instituto Nacional do Câncer (INCA) traz a incidência, morbidade e a mortalidade das ocorrências desta doença no âmbito nacional. No ano de 2018 houve um aumento significativo em novos casos em homens. Destes casos, 31,7% a localização primária do câncer, manifestou-se na próstata. Nas mulheres, o aumento da doença em 2018 foi de 282.450, destes 29,5% dos casos originou-se nas mamas. Conforme os Registros Hospitalares de Câncer (RHC), referente ao intervalo de tempos em dias, entre a primeira consulta e o diagnóstico, dos 28.066 casos registrados, 50,1% ficaram sem diagnósticos e sem tratamento nos primeiros 15 dias de investigação. Neste mesmo período na faixa de consulta/tratamento, 41,7% receberam o diagnóstico, mas não haviam iniciado o tratamento. No período de 31 a 60 dias, 15,1% ficaram sem diagnóstico e tratamento e 18,9% com diagnóstico e sem tratamento (BRASIL, 2021).

Enquanto isso, registros médicos podem fornecer informações úteis sobre os casos de câncer já tratados. Depois de um longo período, as informações acumuladas em um banco de dados das instituições de saúde podem ser valiosas para auxiliar os médicos na tomada de decisão e ações apropriadas em casos específicos. As informações clínicas normalmente estão relacionadas com o acesso aos dados de saúde do paciente, geralmente fornecidas em uma variedade de formatos: dados estruturados (como registos de saúde eletrônicos), dados semiestruturados (como documentos XML (*Extensible Markup Language*) ou resultados laboratoriais tabulares) e dados não estruturados (como desenvolvimentos em texto livre e relatórios médicos), principalmente por causa da grande variedade de formatos de entrada e a diversidade de aspectos envolvidos na descrição de informações não-estruturadas em texto livre (SCHWERTNER; RIGO, 2018). O desenvolvimento de estratégias de avaliação e tratamento de sintomas mais eficazes é essencial para melhorar a qualidade de vida dos pacientes relacionada à saúde.

2.2 Processamento de Linguagem Natural

O PLN pode ser visto como uma forma de comunicação entre o homem e a máquina. Para que um ser humano possa interagir com uma máquina é necessário uma tradução entre as linguagens interpretadas por cada um (SILVA et al., 2019). Nesse contexto, são aplicadas técnicas e algoritmos para suportar a compreensão da semântica da linguagem (QI et al., 2020). Com isso, considera-se o seu uso significativo e útil, implicando em um grande valor em várias áreas, embora comumente seja utilizado para fins de pesquisa científica no meio acadêmico. (SILVA

et al., 2019)

Atualmente, o PLN é utilizado como uma técnica analítica amplamente usada na área de saúde (MEHTA; PANDIT, 2018) (KOLECK et al., 2019), sendo um campo interdisciplinar que combina linguística computacional, ciência da computação, ciência cognitiva e IA. As aplicações típicas de PLN incluem reconhecimento de fala, compreensão da linguagem falada, sistemas de diálogo, análise lexical, tradução automática, gráfico de conhecimento, recuperação de informações, resposta a perguntas, análise de sentimento, computação social, Geração de Linguagem Natural (GLN), resumo de LN (DENG; LIU, 2018), geração de legendas de imagens e geração de texto. Em grande parte, o PLN se concentra na análise de texto e fala para inferir o significado das palavras (ESTEVA et al., 2019). Fang et al. (2016) e Jensen, Jensen and Brunak (2012) destacam a pertinência no uso do PLN em conjunto com recursos linguísticos, estatísticos e heurísticos para extrair e processar a informação não estruturada contida no texto.

Utilizando as regras de PLN e semânticas, Sabra et al. (2018) extraiu informações sobre fatores de risco de tromboembolismo venoso, a partir de relatos clínicos. Eles reforçam que a extração de informações com maior precisão é fundamental para uma maior certeza no prognóstico de doenças. Entre vários exemplos de aplicações, a maioria dos artigos enfatizam a importância da obtenção dos dados de forma estruturada, completa e válida (POLPINIJ, 2011) (SHICKEL et al., 2017) (FERNANDES et al., 2009). Outros artigos criaram interfaces de comunicação para obter dados a partir de fontes estruturadas e ampliar as possibilidades da proposta (BUCUR et al., 2013) (SARAIVA et al., 2016). Vários trabalhos usam SVM (*Support Vector Machines*) em conjunto com outras técnicas. Polpinij (2011) e Zhang et al. (2017) usam SVM e Naive Bayes como um método de classificação de texto para extrair resumos clínicos. No caso do Islam et al. (2010), SVM é usado para extrair informações da literatura biomédica, biomarcador. Zhang et al. (2017) testaram vários métodos de PLN e IA para extrair a classificação da *New York Heart Association* de registros clínicos de texto livre, incluindo métodos baseados em regras e ML. Poucos estudos consideram a relação dos problemas dos pacientes entre eles, e com conceitos identificados como testes e tratamentos. São ainda raros os trabalhos que levam em conta as necessidades dos profissionais de saúde.

Embora a ambiguidade e a complexidade da linguagem médica tornem a aplicação do PLN um desafio, ele tem sido usado para uma variedade de propósitos relacionados à saúde, incluindo a identificação de fatores de risco de doenças, avaliação da eficiência do atendimento e custos e extração de informações clínicas de texto livre em EHRs. Muitos dos dados clínicos ricos e expressivos capturados em EHRs são documentados e armazenados dentro dessas narrativas de texto livre não estruturadas. Consequentemente, tais narrativas de texto livre têm sido a fonte de dados para a comunidade de saúde de PLN (KOLECK et al., 2019).

A medicina está rapidamente se tornando intensiva em dados. Stephens et al. (2015) afirmam que a aplicação de IA será fundamental para o uso intensivo de dados no que diz respeito à geração e análise de dados nos próximos 10 anos. O uso de PLN na área da medicina cresceu consideravelmente, sendo o câncer o principal assunto estudado, correspondendo a aproximada-

mente 25% de todas essas pesquisas na área (SOUZA et al., 2021a). Algoritmos automatizados que extraem padrões significativos podem fornecer conhecimento prático e mudar a maneira como os tratamentos são desenvolvidos, os pacientes são classificados e as doenças são estudadas (SCHWALBE; WAHL, 2020). Na área da saúde, tecnologias de linguagem podem potencializar as aplicações em domínios como EHRs (ESTEVA et al., 2019). Com o desenvolvimento do aprendizado profundo, várias RNs têm sido amplamente utilizadas para resolver tarefas de PLN aplicadas em diferentes demandas da área da saúde (QIU et al., 2020).

O PNL tem sido particularmente bem-sucedida na área de REN, permitindo a identificação de diagnósticos, medicamentos e outros conceitos descritos por uma única palavra ou várias palavras espaçadas (por exemplo, pneumonia ou infarto do miocárdio) (TURCHIN; MASHARSKY; ZITNIK, 2023). Por outro lado, construções linguísticas compostas por termos espaçados entre si na frase ou mesmo em frases diferentes, podem ser mais desafiadoras. Conceitos complexos dessa natureza, portanto, representam uma próxima fronteira crítica na análise de dados de EHR (GROVES et al., 2016).

2.3 Aprendizado de máquina e modelos de linguagem

Historicamente, a construção de um sistema de ML exigia experiência de domínio e engenharia humana para projetar extrações de recursos que transformavam dados brutos em representações adequadas a partir das quais um algoritmo de aprendizagem poderia detectar padrões (ESTEVA et al., 2019). Os métodos convencionais de ML eram limitados em sua capacidade de processar dados naturais em sua forma bruta. Durante décadas, a construção de um sistema de reconhecimento de padrões ou ML exigiu uma engenharia cuidadosa e considerável experiência de domínio para projetar uma representação interna adequada a partir do qual o subsistema de aprendizado, geralmente um classificador, pode detectar ou classificar padrões (LONDHE; BHASIN, 2019).

Os métodos de PLN sem uso de ML geralmente dependem fortemente dos recursos artesanais discretos, enquanto os métodos empregando RNs geralmente usam vetores densos e de baixa dimensão (também conhecidos como representação distribuída) para representar implicitamente os recursos sintáticos ou semânticos da linguagem. Essas representações são aprendidas em tarefas específicas da PLN, de modo que os métodos baseados em RNs tornam mais fácil o desenvolvimento de sistemas de PLN (QIU et al., 2020).

A extração de informação (MANNING; SCHUTZE, 1999) tem como um dos seus desafios o domínio da saúde, devido aos nomes complexos de sintomas e doenças e a diversidade de documentos encontrados (DHOLE; UKE, 2014) (RAMESH et al., 2021). Além destes, Sezgin and Özkan (2013) descrevem que as aplicações nesta área apresentam dificuldades para integrar múltiplas fontes de dados. Este argumento é retomado por Balzano and Del Sorbo (2014), que enfatiza como o acesso efetivo a informações relevantes sobre o domínio da saúde requer a capacidade de processar e organizar informações automaticamente, especialmente quando elas

estão contidas em grandes repositórios de dados.

A aplicação de IA na medicina compreende principalmente o ML, que envolve algoritmos matemáticos que ajudam a melhorar o aprendizado por meio da experiência. Existem três tipos principais de algoritmo de ML (LONDHE; BHASIN, 2019):

- Não supervisionado (encontrar padrões);
- Supervisionado (algoritmos de classificação e previsão com base no aprendizado anterior);
- Aprendizagem por Reforço (usando recompensas e punições para formar uma estratégia para operar em um espaço de problema projetado).

O aprendizado não supervisionado envolve um sistema que procura padrões nos dados de entrada não rotulados e classifica posteriormente os dados de entrada, dependendo desses padrões identificados. Na aprendizagem supervisionada, é fornecido ao sistema dados rotulados, o que ajuda o sistema a categorizar as várias entradas, dependendo do que foi aprendido com os dados rotulados (PEHLEVAN; CHKLOVSKII, 2019).

O aprendizado profundo (*Deep Learning*) pode ser adotado no aprendizado supervisionado. Os métodos de classificação supervisionados incorporam grande quantidade de conhecimento prévio na forma de um conjunto de dados de treinamento com rótulos conhecidos (RUSSELL; NORVIG, 2021). A partir desse conjunto de dados de treinamento, o algoritmo aprende os limites de decisão entre as classes em um espaço de recursos de alta dimensão, que pode então ser aplicado a dados de teste não rotulados. A saída na aprendizagem supervisionada é definida. Em contraste, a saída não é definida no aprendizado não supervisionado. Em vez disso, essa abordagem funciona na suposição de que pode haver algum padrão observado nos dados de entrada que serão refletidos na saída (KELLEHER, 2019).

A Aprendizagem por Reforço utiliza algoritmos baseados em um sistema de recompensa e punição. Um algoritmo de aprendizado por reforço, ou agente, aprende interagindo com seu ambiente. O agente recebe recompensas pelo desempenho correto e penalidades pelo desempenho incorreto. O agente aprende sem intervenção de um humano, maximizando sua recompensa e minimizando sua penalidade (KAISER et al., 2019).

A aplicação dos conceitos acima aos cuidados de saúde levou a uma era da computação cognitiva, na qual tais modelos podem ingerir uma grande quantidade de dados para detectar padrões aos quais foram expostos anteriormente. Por exemplo, terapias direcionadas têm sido usadas em pacientes com câncer com sucesso limitado, apesar dos oncologistas tentarem definir subconjuntos de pacientes que podem se beneficiar de um tratamento específico. No entanto, essa definição é baseada em grandes quantidades de dados associados à genômica do paciente, imagem, comorbidades e tratamentos anteriores (BERTSIMAS; WIBERG, 2020).

As abordagens de ML podem ajudar os oncologistas a analisar esses dados durante os diferentes estágios da terapia, incluindo diagnóstico, tratamento, acompanhamento e prognóstico.

A IA pode manipular e classificar enormes quantidades de dados, usando-os para prever padrões que, quando aplicados à terapêutica, pode tornar o processo de identificação de grupos de pacientes mais rápido, fácil e preciso. Assim, quando combinada com experiência e conhecimento humanos, a IA pode melhorar significativamente o diagnóstico de câncer e o atendimento ao paciente (BERTSIMAS; WIBERG, 2020) (LAMY et al., 2019). A saúde e a medicina podem ser beneficiadas com o aprendizado profundo devido ao grande volume de dados sendo gerados e seu potencial de contribuição em problemas diversos (ESTEVA et al., 2019).

O DL é uma abordagem de aprendizagem na qual uma máquina é alimentada com dados brutos e desenvolve suas próprias representações necessárias para o reconhecimento de padrões, sendo composta de várias camadas de representações. Essas camadas são normalmente organizadas sequencialmente e compostas por um grande número de operações não lineares primitivas. Conforme os dados fluem através das camadas do sistema, o espaço de entrada torna-se iterativamente distorcido até que os pontos de dados se tornem distinguíveis. Desta forma, funções altamente complexas podem ser aprendidas (ESTEVA et al., 2019).

Modelos de DL podem ser escalados para grandes conjuntos de dados - em parte devido à sua capacidade de execução em hardware de computação especializada - e continuam a melhorar com mais dados, permitindo-lhes superar muitas abordagens clássicas de ML. Os sistemas de DL podem aceitar vários tipos de dados como entrada, tendo uma relevância particular para dados heterogêneos de saúde. Os modelos mais comuns são treinados usando aprendizado supervisionado, no qual os conjuntos de dados são compostos de pontos de dados de entrada (por exemplo, imagens de lesões cutâneas) e rótulos de dados de saída correspondentes (por exemplo, 'benigno' ou 'maligno') (PENG; YE; CHEN, 2019). Aprendizagem por Reforço, em que agentes computacionais aprendem por tentativa e erro ou por demonstração de especialista, progrediu com a adoção de aprendizagem profunda, alcançando feitos notáveis em áreas como jogos, podendo ser útil na área de saúde sempre que o aprendizado requer demonstração do médico (ESTEVA et al., 2019).

O DL permite que modelos computacionais compostos por várias camadas de processamento aprendam representações de dados com vários níveis de abstração. Esses métodos melhoraram o reconhecimento de fala de última geração, o reconhecimento visual de objetos, a detecção de objetos e muitos outros domínios, como descoberta de medicamentos e genômica. Redes Neurais Convolucionais (RNC) são exemplos que trouxeram avanços no processamento de imagens, vídeo, fala e áudio, enquanto as Redes Neurais Recorrentes (RNR) tratam eficientemente de dados sequenciais, como texto e fala (LECUN; BENGIO; HINTON, 2015).

Modelos preditivos baseados em DL a partir de EHRs oferecem desempenho positivo em muitas tarefas clínicas (RASMY et al., 2021). A IA como previsão de doenças teve um desenvolvimento considerável nos últimos anos. Atualmente, ela pode melhorar a precisão do diagnóstico, permitir a prevenção de doenças por meio de alerta precoce, agilizar a tomada de decisões clínicas e reduzir os custos de saúde. Outras abordagens também foram amplamente aplicadas na modelagem preditiva clínica e obtiveram inúmeros sucessos. Com amostras de

treinamento suficientes, os modelos de DL podem alcançar um desempenho comparável ou até melhor do que os especialistas de domínio no diagnóstico de certas doenças (MUNIR et al., 2019).

Uma tecnologia recente que vem ganhando destaque na literatura são os *Transformers* (VASWANI et al., 2017), que tornaram-se rapidamente a arquitetura dominante para PLN, superando modelos neurais alternativos, RNC e RNR no desempenho de tarefas de compreensão de LN e GLN. A arquitetura é dimensionada com dados de treinamento e tamanho do modelo, facilita o treinamento paralelo eficiente e captura recursos de sequência de longo alcance (WOLF et al., 2020). O pré-treinamento de modelo permite que os modelos sejam treinados em *datasets* genéricos e, subsequentemente, sejam facilmente adaptados a tarefas específicas com forte desempenho (MCCANN et al., 2017) (DEVLIN et al., 2018a) (WOLF et al., 2020).

A arquitetura *Transformers* é particularmente propícia ao pré-treinamento em extensas quantidades de texto, levando a grandes ganhos de precisão nas tarefas, incluindo a classificação de texto (YANG et al., 2019b), compreensão da linguagem (LIU et al., 2019a), tradução (LAMPLE; CONNEAU, 2019), resolução de correferência (JOSHI et al., 2020), inferência de senso comum (BOSSELUT et al., 2019) e resumo (LEWIS et al., 2019), entre outros. Esse avanço leva a uma ampla gama de desafios práticos que devem ser enfrentados para que esses modelos sejam extensamente utilizados.

Representações de linguagem com o modelo BERT apresentam bons resultados para lidar com entradas sequenciais, como textos com numerosas variações. O BERT também vem sendo utilizado amplamente no domínio clínico. No entanto, esses modelos foram pré-treinados em texto clínico e são apenas para tarefas clínicas de PLN (XUE et al., 2019). Um EHR estruturado, como uma fonte primária de entrada para a previsão de doenças, oferece informações ricas e bem estruturadas que refletem na progressão da doença de cada paciente e é um dos recursos valiosos disponíveis para análise de dados de saúde (LEHMAN et al., 2021).

2.3.1 BERT

O *Transformer* é uma RN com uma arquitetura encoder-decoder baseada em mecanismos de atenção. Ele recebe uma sequência de palavras como entrada, codifica-as em representações utilizando camadas de atenção e, em seguida, as decodifica novamente em palavras. À medida que o modelo processa cada palavra na sequência de entrada, um mecanismo de atenção calcula a importância das palavras em outras posições para codificar a palavra atual (VASWANI et al., 2017).

A rede *Transformer* é dividida em dois módulos: o codificador (*encoder*) e o decodificador (*decoder*). Cada um desses módulos consiste em uma pilha de seis camadas idênticas. No módulo codificador, cada camada é composta por duas subcamadas: uma chamada de "*Multi-Head Self-Attention*" e outra de uma RN *feedforward*. Nas camadas do decodificador, além das duas subcamadas mencionadas, há uma terceira subcamada de atenção que processa os dados

provenientes do módulo codificador (RAFFEL et al., 2020).

Inicialmente, o *Transformer* foi proposto para a tarefa de tradução automática de idiomas devido à sua arquitetura *encoder-decoder*. O módulo codificador recebe uma sequência de palavras em um determinado idioma, como o português, codifica as palavras em representações contextuais e as transfere para o módulo decodificador. No decodificador, as representações são processadas passo a passo e as saídas são alimentadas em uma camada de saída, onde as previsões do modelo são geradas, resultando na tradução da sentença para um idioma alvo (VASWANI et al., 2017) (RAFFEL et al., 2020).

Atualmente, os modelos de texto mais avançados usam *Transformers* para ensinar como representar o texto (DEVLIN et al., 2018b), e estão cada vez mais encontrando seu uso em vários ramos do ML, mais frequentemente quando as informações de entrada e saída são uma sequência. Esses modelos usam uma combinação de RNR e RNC. As RNC geralmente têm dificuldade em capturar informações de contexto em longas sequências (DEVLIN et al., 2018b). No entanto, em texto natural, a representação do *token* pode ser influenciado pelo contexto através de várias palavras e até mesmo frases do próprio *token*. Para ter em conta a influência de longo alcance, LSTMs (Long Short-Term Memory) - um tipo de arquitetura de RNR - são usados em conjunto com o mecanismo de atenção para melhorar a eficiência do aprendizado, levando em consideração a influência de *tokens* distantes (KOROTEEV, 2021). Diversos modelos surgiram a partir do *Transformer*, como o BERT (DEVLIN et al., 2018b).

O BERT é uma arquitetura de RN projetada para PLN, e tem se destacado nessa área. Ele é baseado na arquitetura *Transformer*, que permite a aprendizagem de representações de palavras e frases em um contexto amplo e sem supervisão (LIU et al., 2019b). O modelo é treinado em grandes quantidades de texto rotulados, capturando a relação entre as palavras de forma bidirecional e alcançando um melhor entendimento do contexto semântico das sentenças. Essa abordagem tem proporcionado avanços significativos em tarefas como classificação de texto, resposta a perguntas e tradução automática (SANH et al., 2019).

Este modelo destina-se ao aprendizado preliminar profundo de representação de texto bidirecional para uso subsequente em modelos de ML. Existem duas categorias de tarefas de processamento de texto natural: holística, operando com texto no nível da frase, e os *token*izados, como responder a uma pergunta e atribuição de entidades, que produzem uma saída mais detalhada em o nível de elementos de texto individuais (DEVLIN et al., 2019). Ambas as categorias de problemas têm usado recentemente modelos pré-treinados, o que pode reduzir significativamente o tempo para projetar e treinar modelos privados, mantendo um alto nível de eficiência (DEVLIN et al., 2019) (KOROTEEV, 2021).

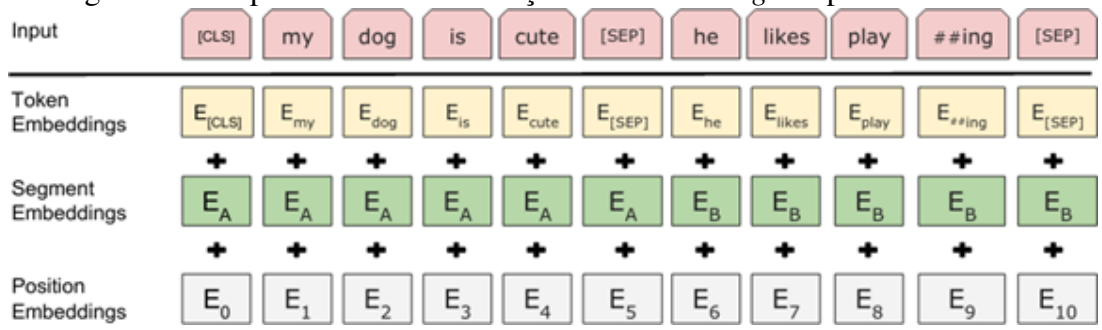
Tecnicamente, o BERT consiste em 12 camadas do codificador do *Transformer* empilhadas, excluindo a parte decodificadora geralmente encontrada em modelos geradores de texto, como o GPT-2. O BERT utiliza a técnica de *Transfer Learning* (MOZAFARI; FARAHBAKHSH; CRESPI, 2020), na qual seus parâmetros são pré-treinados em grandes conjuntos de textos de forma não supervisionada. Para adaptar-se a tarefas específicas em um domínio particular,

requer apenas pequenas modificações, como a inclusão de uma camada de classificação seguida por um treinamento com dados rotulados (RADFORD et al., 2019).

Em vez de realizar o pré-processamento dos textos em nível de caracteres ou palavras, o BERT divide a entrada em subunidades de palavras, seguindo a abordagem *Wordpiece* (BOUKKOURI et al., 2020). Essa abordagem lida adequadamente com palavras que estão fora do vocabulário conhecido. Por exemplo, a frase "como você está bonito" seria convertida para "como voc ##ê est ##á bo ##nito", dividindo a palavra "bonito" em duas subunidades. Portanto, uma sentença com quatro palavras se torna uma entrada para o modelo composta por seis subunidades de palavras. Além disso, são adicionados dois *tokens* especiais: "[CLS]" e "[SEP]", que dividem a entrada em subunidades de palavras, seguindo a abordagem *Wordpiece* (BOUKKOURI et al., 2020) (DEVLIN et al., 2019) (JOHNSON et al., 2017).

O *token* "[CLS]" é adicionado no início de todas as entradas e é usado como uma representação agregada de todo o texto para classificações de nível de sentença. Por outro lado, o *token* "[SEP]" é usado para separar diferentes sentenças dentro da mesma entrada, conferindo versatilidade ao BERT. Com esse *token*, o modelo pode realizar tarefas que consideram apenas uma sentença, como análise de sentimentos de um texto, assim como tarefas que envolvem pares de textos, como encontrar a resposta para uma pergunta. Nesses casos, a pergunta e o trecho que contém a resposta são separados pelo *token* "[SEP]", que também é repetido no final da entrada. A Figura 1 ilustra um exemplo de texto pré-processado (DEVLIN et al., 2019) (BOUKKOURI et al., 2020) (JOHNSON et al., 2017).

Figura 1 – Pré-processamento formação dos *embeddings* na primeira camada.



Fonte: Devlin et al. (2019)

Enquanto o BERT original (base) foi desenvolvido para análise de texto de domínio geral, versões subsequentes especializadas foram propostas para análise de texto biomédico narrativo, incluindo BioBERT (LEE et al., 2020) e o ClinicalBERT (HUANG; ALTOSAAR; RANGANATH, 2019). A Google disponibiliza modelos pré-treinados em diversas configurações, como modelos próprios para inglês e chinês, assim como multilinguais, treinados em quantidades enormes de textos. O pré-treinamento em inglês, por exemplo, é feito sobre um conjunto de livros, o BookCorpus (BANDY; VINCENT, 2021), e a Wikipedia inteira, totalizando 3 bilhões e 300 milhões de palavras, levando quatro dias para ser completado em uma configuração potente de GPUs (JOHNSON et al., 2017).

2.3.2 Bidirectional Long Short-Term Memory (BILSTM)

Uma entidade nomeada refere-se a uma sequência de palavras que representa uma entidade do mundo real. A tarefa de REN envolve processar automaticamente um *dataset* de textos em LN, identificando e anotando as entidades nomeadas. As anotações fornecem uma classificação das entidades nomeadas com base em um conjunto pré-definido de categorias (WANG et al., 2022). Em essência, o REN produz textos estruturados. Extrair entidades nomeadas é um desafio significativo, que pode se tornar ainda mais complexo dependendo do domínio dos textos analisados. Isso é especialmente evidente em textos médicos, como os EHRs, nos quais não há um padrão estrito para se referir às entidades nomeadas (DONG et al., 2018).

Já a ER é uma tarefa de PLN que envolve identificar e classificar as relações entre entidades nomeadas em um texto. É um desafio importante no campo da IA e do DL, pois requer compreensão semântica e contextual das sentenças. Eles são treinados em grandes conjuntos de dados rotulados, onde as relações entre entidades são explicitamente anotadas, permitindo que o modelo aprenda a mapear os contextos em que as relações ocorrem (GUO; ZHANG; LU, 2019).

REN e a ER são tarefas essenciais no PLN. Nos últimos anos, modelos baseados em DL têm se mostrado eficazes no domínio médico. As RNRs são um tipo de modelo de PLN capazes de lidar com sequências de dados, como texto, e apresentam uma memória interna que permite que informações contextuais sejam mantidas ao longo do tempo. No entanto, as RNRs tradicionais sofrem do problema de desvanecimento do gradiente, o que dificulta a aprendizagem de dependências de longo prazo (HOCHREITER; SCHMIDHUBER, 1997).

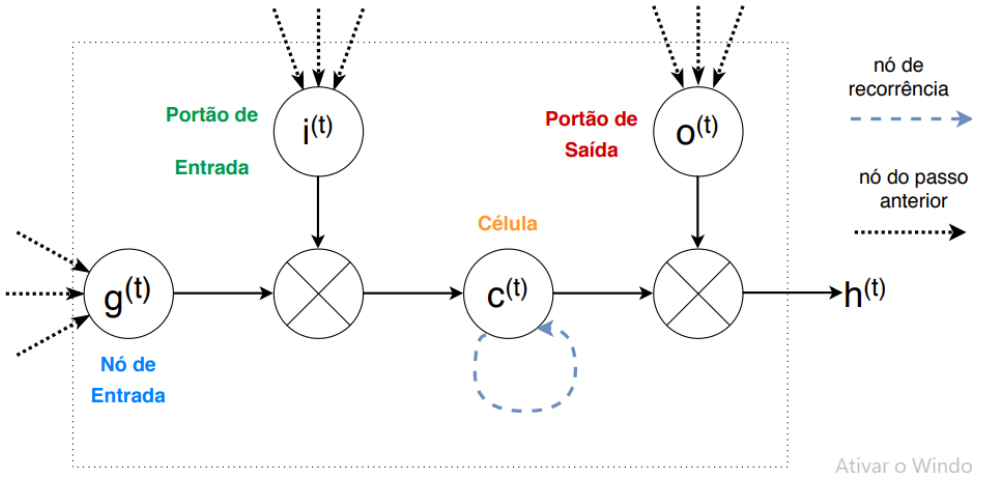
Para resolver o problema de desvanecimento do gradiente em RNRs, foram propostas as redes LSTM. As redes LSTM utilizam unidades de memória especializadas chamadas células LSTM, que possuem um mecanismo de controle de fluxo de informação através de portões. Esses portões permitem que a rede aprenda a lembrar ou esquecer informações relevantes em diferentes etapas da sequência (HOCHREITER; SCHMIDHUBER, 1997).

A LSTM é uma RN com uma estrutura em cadeia, semelhante à RNR. Ela possui blocos chamados de células de memória, que desempenham um papel central. Cada célula de memória contém um nó conhecido como estado da célula, responsável por armazenar as informações relevantes (WU; HE, 2019). Essas informações são controladas por unidades neurais denominadas portões. Os portões são compostos por uma camada *feedforward* seguida por uma função de ativação sigmoide. Sua função é regular o fluxo de informações na camada em que estão localizados. Quando o valor de um nó é próximo de 1, a informação é adicionada à célula, enquanto valores próximos de zero indicam que a informação deve ser removida (CESAR; MANSO-CALLEJO; CIRA, 2023).

Na arquitetura desenvolvida por Hochreiter and Schmidhuber (1997), encontra-se dois tipos de portões: o portão de entrada e o portão de saída. O primeiro é responsável por adicionar informações relevantes ao estado da célula, enquanto o segundo seleciona quais informações

da célula serão utilizadas como saída. A Figura 2 apresenta uma representação visual de uma célula em uma LSTM.

Figura 2 – Representação visual de uma célula em uma LSTM



Fonte: Hochreiter and Schmidhuber (1997)

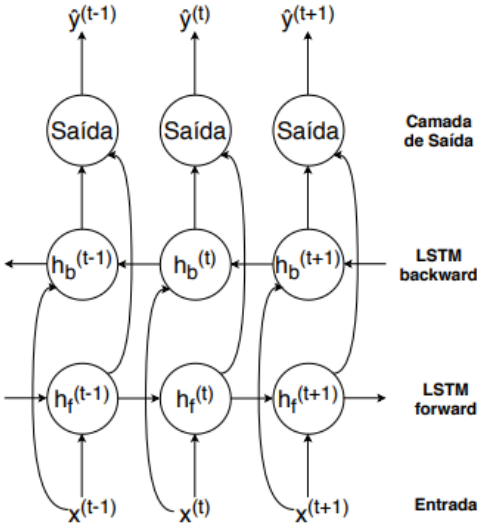
A arquitetura BILSTM foi proposta por Schuster and Paliwal (1997), sendo desenvolvida para aproveitar o contexto tanto passado quanto futuro em uma sequência de dados. Ela consiste em duas camadas de LSTM, uma processando a sequência no sentido direto e a outra no sentido inverso. Quando os modelos *forward* e *backward* são combinados, é possível criar uma Rede Neural Bidirecional, que tem a capacidade de unir as saídas das duas redes em uma representação única, capturando informações de ambos os sentidos. Isso permite que a rede tenha acesso a informações contextuais anteriores e posteriores a cada posição da sequência, melhorando o desempenho em tarefas de PLN, como REN e ER (GRAVES; SCHMIDHUBER, 2005). A Figura 3 ilustra uma BILSTM.

A cada etapa t, os estados ocultos da LSTM *forward* e *backward* são mesclados por meio de uma operação, como concatenação, multiplicação ou soma. Essas saídas são então utilizadas como entrada para uma camada de pós-processamento, como uma RN *feedforward* com uma função de ativação *softmax*, para gerar as previsões do modelo.

No REN, o objetivo é identificar e classificar palavras em categorias pré-definidas, como nomes de pacientes, medicamentos e procedimentos médicos, entre outras. A BILSTM tem se mostrado uma abordagem promissora para essa tarefa, devido à sua capacidade de capturar informações contextuais. Estudos, como o de Ma and Hovy (2016), demonstraram que a BILSTM supera abordagens tradicionais, como Conditional Random Fields (CRF), alcançando melhores resultados em benchmarks de REN. A abordagem BILSTM tem sido aplicada com sucesso no REN em vários domínios, incluindo o médico. Estudos mostram que o uso de BILSTMs em combinação com outros recursos, como word *embeddings*, melhora significativamente o desempenho do REN em textos clínicos (CHOI et al., 2016b).

Além do REN, a ER entre entidades é importante para compreender as interações e associa-

Figura 3 – Representação visual de uma BILSTM



Fonte: Silva (2020)

ções em textos médicos. A ER busca identificar e classificar as relações entre pares de entidades em um texto. A BILSTM tem sido aplicada com sucesso nessa tarefa, permitindo a identificação de relações entre entidades nomeadas em um texto. A estratégia de codificar as entidades e suas relações em vetores de características tem sido adotada em várias pesquisas nessa área (ZHOU et al., 2016). Zeng et al. (2014) propuseram um modelo baseado em BILSTM para ER e demonstraram sua eficácia em diferentes domínios, como a ER médicas.

Para melhorar ainda mais o desempenho das BILSTMs, podem ser utilizados recursos linguísticos, como word *embeddings* pré-treinados. Essas representações distribuídas de palavras capturam informações semânticas e contextuais que são úteis para as tarefas de REN e ER. O uso de *embeddings* pré-treinados tem se mostrado eficaz na melhoria do desempenho das BILSTMs em vários estudos (PETERS et al., 2018). Um aspecto importante para o desempenho da BILSTM é a representação das palavras. A utilização de *embeddings* pré-treinados, como o Word2Vec (Mikolov et al., 2013), tem sido comumente adotada para capturar informações semânticas das palavras. Além disso, técnicas como o uso de caracteres e subpalavras também têm sido aplicadas para melhorar a representação de palavras desconhecidas (LAMPLE et al., 2016).

A aplicação de mecanismos de atenção em modelos baseados em BILSTM tem se mostrado promissora para melhorar o desempenho do REN e ER. Yang et al. (2016) propuseram um modelo de atenção baseado em BILSTM para REN e mostraram que ele supera abordagens anteriores em benchmarks como o CoNLL-2003. O treinamento de uma BILSTM envolve a definição de uma função de perda, como a entropia cruzada, e a otimização dos pesos da rede por meio de algoritmos como o gradiente descendente estocástico (GOODFELLOW; BENGIO; COURVILLE, 2016).

A transferência de aprendizado tem sido amplamente explorada na aplicação de BILSTM

para REN e ER. Utilizando técnicas como *fine-tuning* de modelos pré-treinados, é possível aproveitar o conhecimento adquirido em tarefas semelhantes para melhorar o desempenho em um novo domínio ou conjunto de dados (PETERS et al., 2018). Ele consiste em inicializar os pesos da rede com parâmetros pré-treinados em tarefas relacionadas, como o treinamento em grandes conjuntos de dados genéricos ou tarefas semelhantes de PLN. Essa abordagem permite que a rede comece com um conhecimento prévio, acelerando o processo de convergência (YOSINSKI et al., 2014).

A avaliação do desempenho de uma BILSTM para REN e ER pode ser realizada utilizando métricas como precisão, *recall* e *F1-score*. Essas métricas permitem quantificar a capacidade do modelo de identificar corretamente as entidades e relacionamentos em um texto. Além disso, técnicas como validação cruzada podem ser aplicadas para avaliar a generalização do modelo em diferentes conjuntos de dados (SOKOLOVA; LAPALME, 2009).

Abordagens de aprendizado auto-supervisionado têm ganhado destaque na área de PLN. A utilização de representações latentes pré-treinadas em grandes conjuntos de dados não rotulados, como o BERT (DEVLIN et al., 2019), tem se mostrado eficaz na melhoria do desempenho de modelos baseados em BILSTM para REN e ER (PETERS et al., 2018).

A disponibilidade de conjuntos de dados anotados adequadamente é crucial para o desenvolvimento e avaliação de modelos de REN e ER. Diferentes conjuntos de dados têm sido utilizados, como o CoNLL-2003 para REN (TJONG KIM SANG; DE MEULDER, 2003) e o SemEval-2010 Task 8 para ER (HENDRICKX et al., 2010). Esses conjuntos de dados fornecem referências sólidas para comparação de desempenho entre diferentes abordagens. Ao mesmo tempo, a redução da necessidade de dados rotulados é um desafio importante no desenvolvimento de modelos de REN e ER. Técnicas como o aprendizado ativo e a aprendizagem semi-supervisionada têm sido exploradas para minimizar a dependência de grandes conjuntos de dados rotulados(SETTLES, 2009) (MINTZ et al., 2009).

Os EHRs contêm uma grande quantidade de informações médicas que podem ser exploradas para melhorar a qualidade do atendimento e a tomada de decisões clínicas. A aplicação de BILSTMs para REN e ER em EHR pode fornecer insights valiosos sobre a evolução dos pacientes, identificando condições médicas, tratamentos e resultados (JAGANNATHA; YU; LIU, 2016). Apesar dos avanços recentes, ainda existem desafios a serem enfrentados no campo do REN e da ER. A falta de conjuntos de dados anotados e a necessidade de anotações especializadas em domínios médicos são obstáculos para treinar modelos robustos. Além disso, o processamento de textos médicos não estruturados e a interpretação de termos ambíguos representam desafios adicionais (PRADHAN; ELHADAD; SOUTH, 2013). Além disso, a aplicação de BILSTM em línguas diferentes do inglês também merece atenção (LI et al., 2022).

2.4 Ontologias

As ontologias são modelos de representação que relatam um domínio formalmente, determinando seus conceitos, bem como relacionamentos, restrições e regras relativas. Utiliza-se de uma estrutura de dados abstrata que descreve o aspecto semântico de um conjunto de informações (GUARINO; OBERLE; STAAB, 2009), (GRUBER, 1995), (SALATINO et al., 2020). Esta estrutura é facilmente processada por computadores, possibilitando assim sua reutilização e compartilhamento (ZANATTA et al., 2012). Segundo URBANSKYY (2018), as ontologias permitem que o conhecimento tácito seja explicitado, formalizando uma visão relevante do mundo e tornando este modelo passível de processamento e interpretação por parte dos computadores.

Em geral, o objetivo de uma ontologia é representar o conhecimento através de variáveis e relações relevantes entre elas, fornecendo assim o entendimento comum para um domínio (KALET et al., 2017), (CONG, 2014). Tal entendimento facilita a comunicação precisa e eficaz das informações, o que, por sua vez, leva a outros benefícios, como a interoperabilidade, reutilização e compartilhamento (ARSENE; DUMITRACHE; MIHU, 2011) (COTA et al., 2020). A representação é feita através de uma base de conhecimento, que se torna disponível para aplicações que necessitam usar e/ou compartilhar o conhecimento de um domínio. Uma ontologia pode ser facilmente modificada, estabelecer hierarquias e relacionamentos entre entidades e, por fim, utilizada para realizar inferências lógicas. Estas características têm se provado úteis em vários campos (KALET et al., 2017).

As ontologias vêm sendo aplicadas em diferentes áreas da computação, dentre elas, a Integração de Dados, IA e a Web Semântica. Visando tais áreas, as ontologias têm sido usadas como uma maneira de uniformizar o compartilhamento de informações em diferentes tipos de aplicações, relatando um domínio de interesse, seja ele de forma a representar uma classificação de conceitos e/ou axiomas de um teorema ou um esquema de banco de dados (SOUZA; DURAN; VIEIRA, 2014). A principal funcionalidade de uma ontologia é elevar a compreensão do conhecimento de um determinado domínio, usando um método de conceituação formal, facilmente compartilhável. Tal representação de conhecimento é formalmente especificado de maneira explícita para facilitar a interpretação e compreensão tanto de computadores como de pessoas (YAN et al., 2013). Através desta abordagem existe a possibilidade de utilização desta representação do conhecimento do mundo real por parte dos computadores, a fim de lidar com fatos de interesse (ZANATTA et al., 2012) (QI; DING; LIM, 2020).

Na ciência da computação, as ontologias são consideradas um complemento para artefatos de software e também como linguagem para modelagem de recursos específicos. As mesmas tornaram-se a ferramenta de escolha das comunidades de IA para representar o conhecimento de uma área específica, de modo a facilitar o compartilhamento e reutilização das informações (SOUZA; DURAN; VIEIRA, 2014), (URBANSKYY, 2018), (SILVA et al., 2018). Existem hoje inúmeras ferramentas disponíveis para ajudar no desenvolvimento de uma ontologia, tais como Protégé, Ontop, OntoDebug, entre outras. Como o uso de ontologias tem crescido, torna-

se mais fácil reutilizar outras ontologias para construção ou ampliação em algum caso específico de necessidade de um domínio do conhecimento. Isso ajuda a evitar o problema de silos de conhecimento (KALET et al., 2017).

Dentre as vantagens de seu uso na computação, destaca-se a possibilidade dada aos programadores de reaproveitar o conhecimento, podendo realizar alterações e implementar extensões para sua utilização. A criação de estruturas de conhecimento é uma das atividades mais longas, caras e complexas de um sistema especialista (GEORGE; LAL, 2019). Há um alto ganho de tempo, trabalho e investimento quando reutilizada uma ontologia, visto que existe uma ampla quantidade de ontologias disponíveis para uso, reutilização e integração, podendo ser complementadas de acordo com os conceitos dos domínios (ZANETTI; BONACIN, 2014).

A construção de ontologias envolve peculiaridades arquiteturais e metodológicas que a diferenciam da análise e desenvolvimento de sistemas de informação tradicionais. Dessa forma, no intuito de sistematizar essa atividade, foram propostas diversas metodologias específicas para a elaboração de ontologias. Entretanto, nenhuma das metodologias existentes é totalmente madura, principalmente se comparadas com metodologias de engenharia de software, não havendo, assim, um padrão para sua construção (URBANSKYY, 2018).

O desenvolvimento em linguagens de ontologias vem sendo realizado com o uso do OWL (*Web Ontology Language*), aprovado pelo *World Wide Web Consortium* (W3C) para promover a visão da Web Semântica, tornando-se a linguagem de uso padrão (ARSENE; DUMITRACHE; MIHU, 2011) (RASMUSSEN et al., 2021). O W3C inaugurou um ambiente para pesquisa e troca de informações, principalmente considerando-se a quantidade de dados que armazena. No entanto, estas informações são, em sua maioria, armazenadas sem qualquer critério de organização, padronização ou classificação quanto ao significado que carregam. Isso dificulta o processamento dos dados por meio de mecanismos inteligentes e automatizados (ARSENE; DUMITRACHE; MIHU, 2011).

Por esta razão, tem-se hoje a necessidade de transformar a Web atual em uma Web Semântica na qual a informação é provida tanto para o consumo humano quanto para o processamento de máquina. No entanto, para que a Web Semântica funcione, componentes de software devem ter acesso às informações e também as regras de inferência e significado destes dados. Nesse sentido, a forma de se representar a informação no contexto da Web Semântica é de vital importância nos dias atuais (SHARMA; KUMAR; RANA, 2017).

Devido a existência de diversas aplicações de ontologias, é importante enfatizar as vantagens dessa utilização (URBANSKYY, 2018), tais como pode ser observado nos pontos a seguir:

- Comunicação e colaboração entre pessoas: ontologias reduzem os conflitos conceituais e terminológicos na organização, uma vez que definem uma conceituação unificada da mesma.

- Formalização: devido a natureza formal de uma ontologia, há uma eliminação de contradições e inconsistências

- Interoperabilidade entre sistemas: ontologias podem ser utilizadas na tradução de diferentes representações de bases de dados, fornecendo uma conceituação única a partir da qual outras conceituações se normalizam.

- Representação de conhecimento e reuso: ontologias formam um vocabulário de consenso e representam o conhecimento de forma explícita no seu mais alto nível de abstração, possuindo um elevado potencial de reuso.

Uma ontologia representa, explicitamente, as classes de entidades de um domínio de aplicação, suas propriedades, suas relações, os papéis que podem desempenhar, como eles são decompostos em partes, os eventos e os processos que estão envolvidos. O conhecimento pode ser extraído de uma ontologia usando o raciocínio lógico, que explora tanto as relações entre classes (conceitos) e os fatos armazenados (as instâncias das classes) (ANDREA; FRANCO, 2012) (ABEYSINGHE et al., 2020). A utilização das ontologias permite o acesso automático a bases de conhecimento de modo que, em conjunto com motores de inferência, podem ser usados para inferir novos conhecimentos a partir de fatos já conhecidos (FENZ, 2012) (ABEYSINGHE et al., 2020).

Nos últimos anos, diversas aplicações foram desenvolvidas com a utilização de ontologias, inclusive em áreas relacionadas à saúde, tais como, informações armazenadas em EHRs e conhecimento medicinal (ZANATTA et al., 2012), (URBANSKYY, 2018), (CHEN et al., 2019). Atualmente, há mais de 500 ontologias biomédicas registradas no BioPortal[1]. Em padrões de sintaxe e semânticas internacionais, uma ontologia tem ampla aplicabilidade em uma variedade de contextos (KALET et al., 2017). O conhecimento medicinal é estruturado nas ontologias mediante a sua complexidade e o formalismo usado para a representação de conceitos, propriedades e relações. Esse formalismo possui um amplo poder expressivo, sendo altamente adequado para organizar grandes conjuntos de conhecimento (ZANATTA et al., 2012) (CHEN et al., 2019).

Quanto ao uso de ontologias na saúde, existe um paralelo com os avanços na área que podem ser feito. Existe uma transição do formato de armazenamento do histórico de saúde dos pacientes, normalmente registrado manualmente em prontuários armazenados em grandes arquivos físicos localizados geralmente nos subsolos de hospitais e clínicas (ADEL et al., 2019a). Atualmente a maioria das instituições de saúde mantêm alguma forma de digitalização de dados e observa-se o desenvolvimento de prontuários eletrônicos (SCHWERTNER; RIGO, 2018).

Com isso, dois protocolos envolvendo informações médicas foram criados como forma de padronizar os dados e garantir a melhor interoperabilidade entre os estabelecimentos de saúde: *Digital Imaging and Communications in Medicine* (DICOM) e o *Health Level Seven International* (HL7). Tais padrões, embora importantes, são extremamente limitados e não resolvem o problema da semântica da informação (KALET et al., 2017). Por outro lado, esta semântica está sendo representada em formatos de ontologias, gerando um cenário favorável à novas

[1] bioportal.bioontology.org

aplicações na área da saúde (GRUBER, 1995) (ADEL et al., 2019b).

Após o aporte teórico sobre registros médicos, PLN, DL e Ontologias, o próximo capítulo apresenta os trabalhos correlatos.

3 TRABALHOS RELACIONADOS

Para o estudo de trabalhos correlatos, foi realizada uma revisão sistemática de literatura na qual, após um processo de filtragem de 964 publicações, foram analisados 23 trabalhos com foco em ER. Um artigo desta revisão sistemática foi enviado para o *Journal* JCBM[1] (*Computers in Biology and Medicine*) sob o título "*Deep Learning And Natural Language Processing Applied To Health Data, Information Extraction, And Ontologies: A Systematic Literature Review*".

A revisão sistemática foi desenvolvida seguindo uma mescla do protocolo da pesquisadora (KITCHENHAM; CHARTERS, 2007), referente à área da computação, e o protocolo de Recomendação PRISMA (PRISMA, 2021), direcionado para à área da saúde. O Protocolo completo desta revisão sistemática está disponível no Apêndice A deste trabalho. Esta abordagem permite que sejam levados em consideração aspectos do processo de escolha e análise dos artigos, ampliando a assertividade dos resultados, quando comparados com os resultados obtidos a partir de revisões diretas de literatura (COOPER, 2016).

O tema principal desta revisão sistemática consiste na ER, DL e PLN, buscando conexões com a saúde e técnicas de extração de dados e uso de ontologias. O protocolo foi desenvolvido com a ajuda de especialistas da área de computação ligados a área da saúde e a área de IA. Os itens a seguir descrevem o trabalho desenvolvido em cada etapa do protocolo adotado.

3.1 Elaboração das questões de pesquisa da revisão sistemática

As questões de pesquisa da revisão sistemática levaram este estudo a identificar trabalhos que poderiam estar ligados à ER, em contextos de uso destes recursos em atividades ligadas à saúde. Para este trabalho, foram definidas duas Perguntas Gerais (PG) e oito Perguntas Focalizadas (PF). O propósito das PGs é entender de que forma os trabalhos foram aplicados em relação ao assunto. O objetivo das PFs é identificar quais as tecnologias usadas para construir seus modelos. Por último, o propósito associado com as PEs é encontrar dados estatísticos sobre o estudo realizado. Estas questões são apresentadas na Tabela 1.

Tabela 1: Perguntas de pesquisa adotadas

Referência	**Questão**
Perguntas Gerais	
PG1	Os artigos foram aplicados à área da saúde?
PG2	Qual o objetivo da pesquisa?
Perguntas Focalizadas	
	Continua na próxima página

[1] https://www.sciencedirect.com/journal/computers-in-biology-and-medicine

Tabela 1 – continua da página anterior

Referência	Questão
PF1	Apresentou uso de ontologias?
PF2	O artigo foi aplicado em algum protótipo?
PF3	Qual o idioma alvo?
PF4	Quais relações foram extraídas?
PF5	Qual a área de atuação?
PF6	Qual a origem do *dataset*?
PF7	Quais são as abordagens, técnicas e métodos usados?

Elaborado pelo autor.

3.2 Construção do processo de pesquisa

Através do processo definido pelo protocolo, foram escolhidas as bases de dados para aplicar a sequência de pesquisa, a obtenção dos resultados e a identificação dos principais trabalhos. Foram utilizadas as seguintes bases de dados: A Web Of Science, que proporciona o acesso a periódicos em diversas áreas do conhecimento (SCIENCE, 2021). A IEEExplore, que fornece acesso publicações importantes em Ciência da Computação, (XPLORE, 2021). A ScienceDirect, que é uma das principais fontes mundiais de pesquisa científica na área médica (SCIENCEDIRECT, 2021).

Esta escolha de bases de dados a consultar buscou a identificação, de forma ampla e consistente, de material bibliográfico sobre o tema principal de pesquisa, com a utilização de fontes complementares integrando tanto a área de computação como a área de saúde. A *string* de busca foi criada conforme as palavras-chave definidas, tendo como elementos obrigatórios da pesquisa o uso de "Processamento de Linguagem Natural" (*natural language processing*), "Extração de Relações" (*relation extraction*) e "Deep Learning" ("*deep learning*"). Os termos "Saúde" (*health*) e "Ontologia" (*ontology*) serão definidos como elementos auxiliares. A *string* de busca é apresentada no Quadro 1.

Quadro 1 – Query inserida nas bases de dados

(("*deep learning*") AND ("*relation extraction*") OR ("*natural language processing*" OR "NLP") OR ("*health*") OR ("*ontology*")).

Na primeira etapa, foi executada a *string* de busca nas bases de dados selecionadas. Todas as publicações encontradas foram exportadas no formato BibTex para serem cadastradas na ferramenta StArt (HERNANDES et al., 2010). Após o registro, foram validados os critérios de inclusão e exclusão. Avante, foram realizadas as leituras de todos os títulos, palavras-chave e

resumos, passando, em seguida, para uma validação de introdução e conclusão e, na fase final, a leitura integral dos artigos selecionados. As fases do processo de seleção de artigos são exibidas a seguir:

- Fase 1: Validação dos critérios de inclusão e exclusão;
- Fase 2: Leitura do título, palavras-chave e resumo;
- Fase 3: Leitura da introdução e conclusão;
- Fase 4: Leitura integral dos artigos e validação das respostas para as perguntas.

3.3 Definição dos critérios de inclusão/exclusão

Os estudos foram filtrados para selecionar os artigos relevantes. Para isso, alguns critérios de inclusão/exclusão (CIE) foram definidos, conforme indicado a seguir:

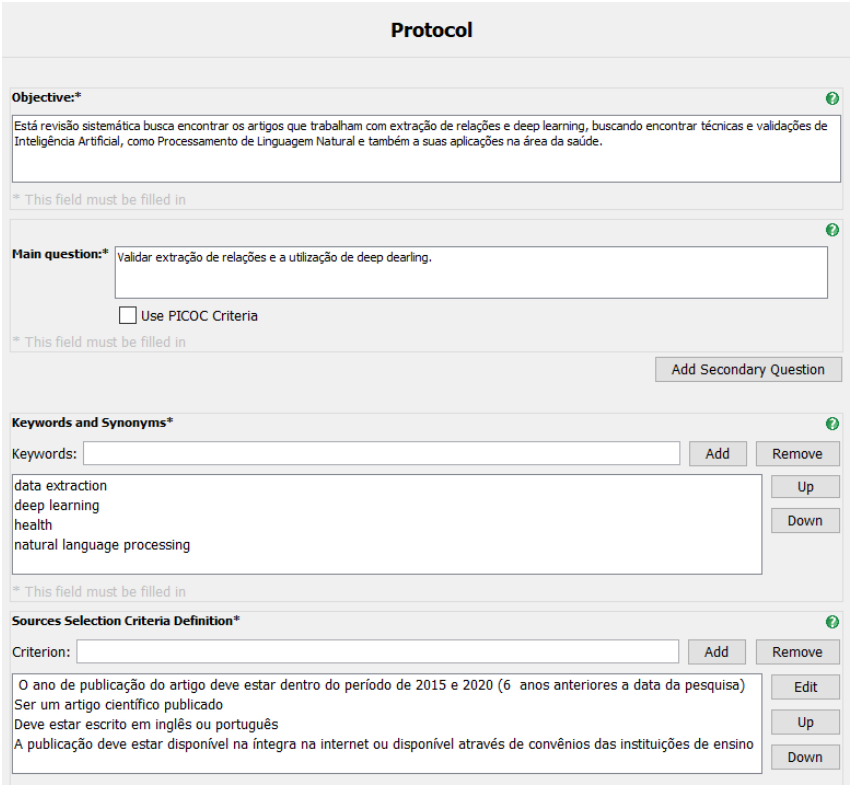
- CIE1: O ano de publicação do artigo deve estar dentro do período de 2016 e 2022 (6 anos anteriores a data da pesquisa);
- CIE2: Ser um artigo científico publicado;
- CIE3: Deve estar escrito em inglês ou português;
- CIE4: A publicação deve estar disponível na íntegra na internet ou disponível através de convênios das instituições de ensino.
- CIE5: O artigo não deve ter menos de 6 páginas.
- CIE6: O *dataset* utilizado na pesquisa deve estar em inglês ou português.
- CIE7: Deve ser aplicado em alguma área da saúde.

3.4 Ferramentas de apoio

Para a classificação dos artigos, foi utilizado o software StArt[2], desenvolvido para apoiar todo o processo de uma revisão sistemática. Por meio de uma árvore hierárquica, ele disponibiliza funcionalidades que apoiam a execução de cada fase. Inicialmente, é cadastrado o protocolo criado (conforme o apêndice A). Após, o software disponibiliza funcionalidades para apoiar as etapas de condução, seleção e extração das informações. Para a importação dos dados na ferramenta, as publicações devem ser exportadas no formato BibTex (HERNANDES et al., 2010). A Figura 4 ilustra o cadastro do protocolo na ferramenta StArt.

[2]http://lapes.dc.ufscar.br/tools/start_tool

Figura 4 – Cadastro do protocolo na ferramenta StArt



Fonte: Elaborado pelo autor

3.5 Fases De Seleção

A inserção da *string* de busca nas bases de dados foi realizada no dia 20 de novembro de 2022. Somando os resultados das três plataformas propostas nesta pesquisa, foram encontradas 964 publicações relacionadas a ER e DL. Após os filtros de período e idioma, restaram 680 artigos a serem analisados, que são de interesse desta pesquisa. A Tabela 2 ilustra os resultados encontrados de acordo com a base pesquisada. A seguir, serão apresentados os resultados de acordo com cada fase de seleção.

Tabela 2 – Comparação dos resultados nas bases de dados

Base	**Quantidade**	**Porcentagem**
Web of Science	96 Artigos	14%
IEEE Xplore	31 Artigos	4%
ScienceDirect	553 Artigos	82%

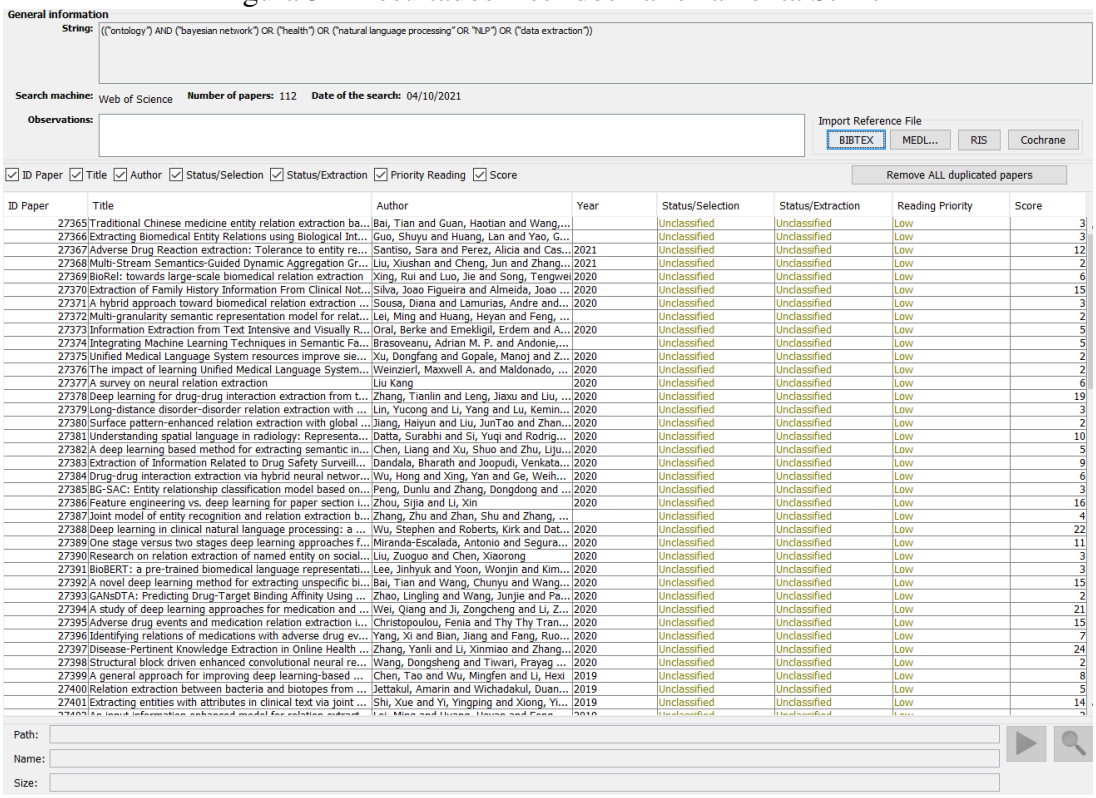
Fonte: Elaborado pelo autor.

3.5.1 Fase 1 - Critérios de inclusão/exclusão;

Como os motores de busca utilizados disponibilizam opções de filtros, antes da exportação dos resultados, foram aplicados os critérios de exclusão referentes ao ano de publicação e o

idioma, excluindo do retorno da busca todos os artigos publicados antes do ano de 2016 e que não estejam escritos em inglês ou português. Os resultados foram exportados no formato BibTex e inseridos na ferramenta StArt, como mostra a Figura 5.

Figura 5 – Resultados inseridos na ferramenta StArt



ID Paper	Title	Author	Year	Status/Selection	Status/Extraction	Reading Priority	Score
27365	Traditional Chinese medicine entity relation extraction ba...	Bai, Tian and Guan, Haotian and Wang,...		Unclassified	Unclassified	Low	3
27366	Extracting Biomedical Entity Relations using Biological Int...	Guo, Shuyu and Huang, Lan and Yao, G...		Unclassified	Unclassified	Low	3
27367	Adverse Drug Reaction extraction: Tolerance to entity re...	Santiso, Sara and Perez, Alicia and Cas...	2021	Unclassified	Unclassified	Low	12
27368	Multi-Stream Semantics-Guided Dynamic Aggregation Gr...	Liu, Xiushan and Cheng, Jun and Zhang...	2021	Unclassified	Unclassified	Low	2
27369	BioRel: towards large-scale biomedical relation extraction	Xing, Rui and Luo, Jie and Song, Tengwei	2020	Unclassified	Unclassified	Low	6
27370	Extraction of Family History Information From Clinical Not...	Silva, Joao Figueira and Almeida, Joao ...	2020	Unclassified	Unclassified	Low	15
27371	A hybrid approach toward biomedical relation extraction ...	Sousa, Diana and Lamurias, Andre and...	2020	Unclassified	Unclassified	Low	3
27372	Multi-granularity semantic representation model for relat...	Lei, Ming and Huang, Heyan and Feng, ...		Unclassified	Unclassified	Low	2
27373	Information Extraction from Text Intensive and Visually R...	Oral, Berke and Emekligil, Erdem and A...	2020	Unclassified	Unclassified	Low	5
27374	Integrating Machine Learning Techniques in Semantic Fa...	Brasoveanu, Adrian M. P. and Andonie,...		Unclassified	Unclassified	Low	5
27375	Unified Medical Language System resources improve sie...	Xu, Dongfang and Gopale, Manoj and Z...	2020	Unclassified	Unclassified	Low	2
27376	The impact of learning Unified Medical Language System...	Weinzierl, Maxwell A. and Maldonado, ...	2020	Unclassified	Unclassified	Low	2
27377	A survey on neural relation extraction	Liu Kang	2020	Unclassified	Unclassified	Low	6
27378	Deep learning for drug-drug interaction extraction from t...	Zhang, Tianlin and Leng, Jiaxu and Liu, ...	2020	Unclassified	Unclassified	Low	19
27379	Long-distance disorder-disorder relation extraction with ...	Lin, Yucong and Li, Yang and Lu, Kemin...	2020	Unclassified	Unclassified	Low	3
27380	Surface pattern-enhanced relation extraction with global ...	Jiang, Haiyun and Liu, JunTao and Zhan...	2020	Unclassified	Unclassified	Low	2
27381	Understanding spatial language in radiology: Representa...	Datta, Surabhi and Si, Yuqi and Rodrig...	2020	Unclassified	Unclassified	Low	10
27382	A deep learning based method for extracting semantic in...	Chen, Liang and Xu, Shuo and Zhu, Liju...	2020	Unclassified	Unclassified	Low	5
27383	Extraction of Information Related to Drug Safety Surveill...	Dandala, Bharath and Joopudi, Venkata...	2020	Unclassified	Unclassified	Low	9
27384	Drug-drug interaction extraction via hybrid neural networ...	Wu, Hong and Xing, Yan and Ge, Weih...	2020	Unclassified	Unclassified	Low	6
27385	BG-SAC: Entity relationship classification model based on...	Peng, Dunlu and Zhang, Dongdong and ...	2020	Unclassified	Unclassified	Low	3
27386	Feature engineering vs. deep learning for paper section i...	Zhou, Sijia and Li, Xin	2020	Unclassified	Unclassified	Low	16
27387	Joint model of entity recognition and relation extraction b...	Zhang, Zhu and Zhan, Shu and Zhang, ...		Unclassified	Unclassified	Low	4
27388	Deep learning in clinical natural language processing: a ...	Wu, Stephen and Roberts, Kirk and Dat...	2020	Unclassified	Unclassified	Low	22
27389	One stage versus two stages deep learning approaches f...	Miranda-Escalada, Antonio and Segura...	2020	Unclassified	Unclassified	Low	11
27390	Research on relation extraction of named entity on social...	Liu, Zuoguo and Chen, Xiaorong	2020	Unclassified	Unclassified	Low	3
27391	BioBERT: a pre-trained biomedical language representati...	Lee, Jinhyuk and Yoon, Wonjin and Kim...	2020	Unclassified	Unclassified	Low	3
27392	A novel deep learning method for extracting unspecific bi...	Bai, Tian and Wang, Chunyu and Wang...	2020	Unclassified	Unclassified	Low	15
27393	GANsDTA: Predicting Drug-Target Binding Affinity Using ...	Zhao, Lingling and Wang, Junjie and Pa...	2020	Unclassified	Unclassified	Low	2
27394	A study of deep learning approaches for medication and ...	Wei, Qiang and Ji, Zongcheng and Li, Z...	2020	Unclassified	Unclassified	Low	21
27395	Adverse drug events and medication relation extraction i...	Christopoulou, Fenia and Thy Thy Tran...	2020	Unclassified	Unclassified	Low	15
27396	Identifying relations of medications with adverse drug ev...	Yang, Xi and Bian, Jiang and Fang, Ruo...	2020	Unclassified	Unclassified	Low	7
27397	Disease-Pertinent Knowledge Extraction in Online Health ...	Zhang, Yanli and Li, Xinmiao and Zhang...	2020	Unclassified	Unclassified	Low	24
27398	Structural block driven enhanced convolutional neural re...	Wang, Dongsheng and Tiwari, Prayag ...	2020	Unclassified	Unclassified	Low	2
27399	A general approach for improving deep learning-based ...	Chen, Tao and Wu, Mingfen and Li, Hexi	2019	Unclassified	Unclassified	Low	8
27400	Relation extraction between bacteria and biotopes from ...	Jettakul, Amarin and Wichadakul, Duan...	2019	Unclassified	Unclassified	Low	5
27401	Extracting entities with attributes in clinical text via joint ...	Shi, Xue and Yi, Yingping and Xiong, Yi...	2019	Unclassified	Unclassified	Low	14

Fonte: Elaborado pelo autor

Após a inserção das publicações no software, automaticamente foram excluídos 10 artigos detectados como duplicados entre as bases selecionadas. Foi realizada uma revisão manual e detectou-se mais 19 artigos duplicados. Neste momento, os mesmos já estavam sendo classificados de acordo com os critérios de inclusão e exclusão previamente definidos no protocolo. Decidiu-se que alguns critérios seriam realizados nas próximas fases, a fim de otimizar o tempo da análise, tais como a verificação da disponibilidade dos artigos na íntegra e as formas de validações, que serão usadas na Fase 3. A ferramenta Start permite a classificação do artigo, se ele deve ser aceito ou rejeitado para a próxima fase, e ainda informa por qual critério de exclusão ou inclusão foi aceito ou rejeitado. Após a validação desta etapa, 651 artigos foram aceitos para a Fase 2.

3.5.2 Fase 2 – Título, palavras-chave e resumo

Na fase 2, realizou-se a leitura dos títulos, palavras-chave e resumo, analisando a presença dos termos “*deep learning*” e “*relation extraction*”. Após a leitura dos resumos e a validação dos novos critérios de exclusão, 96 publicações foram aceitas para a terceira fase.

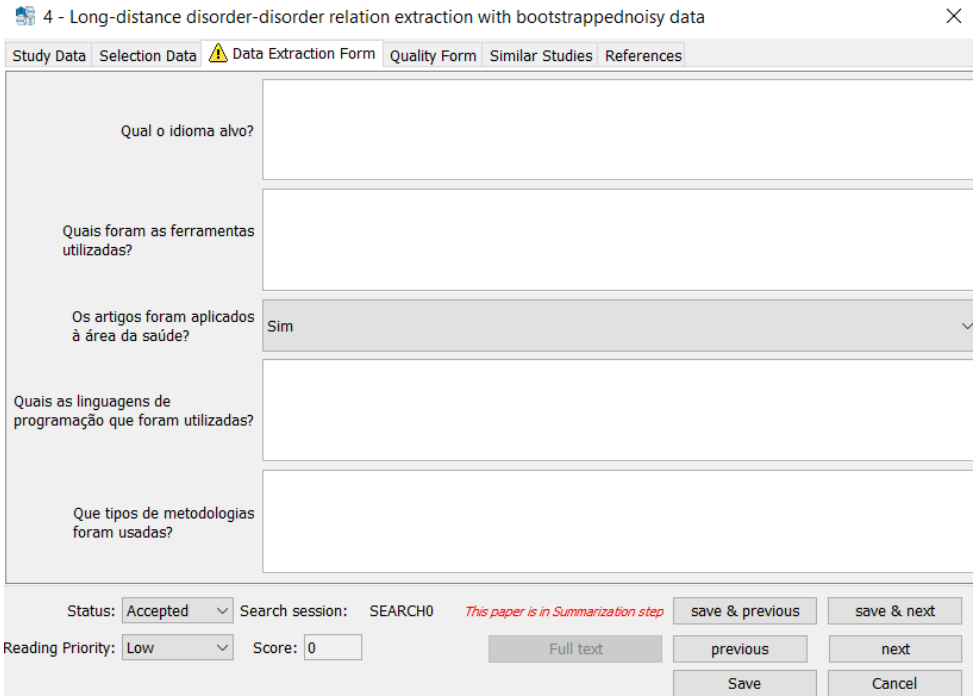
3.5.3 Fase 3 - Introdução e conclusão

Na terceira fase, foram lidas as introduções e conclusões dos artigos aprovados na fase anterior, buscando obter uma ideia mais clara sobre o foco dos artigos que possam ter sido aceitos sem tanta exatidão, bem como, analisar a qualidade dos resultados obtidos. Verificou-se que 71 artigos não estavam adequados à proposta deste trabalho e foram excluídos. Após a leitura e a análise, 25 publicações foram aprovadas para a última fase, na qual foi realizada a leitura completa dos mesmos.

3.5.4 Fase 4 - Leitura integral

A fase 4 consistiu na leitura completa dos 25 artigos que restaram, visando extrair todas as informações das publicações encontradas e organizá-las para que, assim, possam ser respondidas as perguntas pré-definidas no protocolo. Dois artigos selecionados não foram encontrados na íntegra, sendo ambos excluídos por essa validação. A ferramenta StArt disponibiliza uma aba de extração de dados dos artigos, onde foram cadastradas as perguntas e respostas propostas, classificando conforme cada publicação e as informações que se buscava extrair, como mostra a Figura 6.

Figura 6 – Cadastro de extração de dados na ferramenta



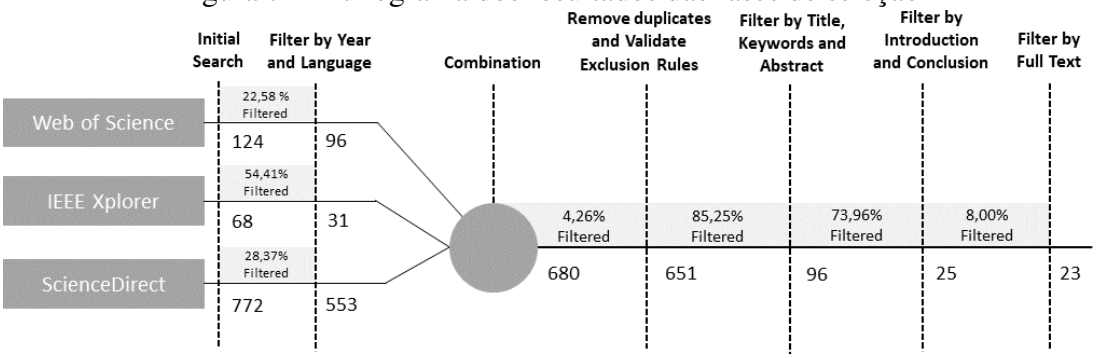
Fonte: Elaborado pelo autor

Após a leitura das 23 publicações e a extração dos dados concluída, iniciou-se a análise dos resultados obtidos. Alguns destes resultados foram simplificados no formato de gráficos e tabelas. A apresentação dos resultados será exibida no próximo item.

3.6 Resultados da Revisão Sistemática

Mediante a conclusão da leitura e a extração das informações de 23 publicações selecionadas para a etapa final, é possível responder as perguntas de pesquisa levantadas no protocolo desta revisão sistemática. Para exemplificar o processo ocorrido na Seção anterior, a Figura 7 ilustra um fluxograma geral dos resultados das classificações dos artigos em cada fase de seleção.

Figura 7 – Fluxograma dos resultados das fases de seleção



Fonte: Elaborado pelo autor

A Tabela 3 mostra os artigos considerados como relevantes para esta revisão sistemática, o seu ano de publicação e a base.

Tabela 3: Título e base dos artigos selecionados

Título	**Base**
Christopoulou et al. (2020)	Web of Science
Li and Yu (2019)	Web of Science
Wei et al. (2020)	Web of Science
Suárez-Paniagua et al. (2019)	Web of Science
Purushotham et al. (2018)	ScienceDirect
Lee et al. (2020)	Web of Science
Lamurias et al. (2019)	Web of Science
Yoon et al. (2019)	Web of Science
Fabregat, Araujo and Martinez-Romo (2018)	Web of Science
Manley et al. (2020)	ScienceDirect
Lamy et al. (2019)	ScienceDirect
Khan and Shamsi (2021)	ScienceDirect
Yang et al. (2020)	Web of Science
Li et al. (2019)	Web of Science
Lin et al. (2020)	ScienceDirect

Continua na próxima página

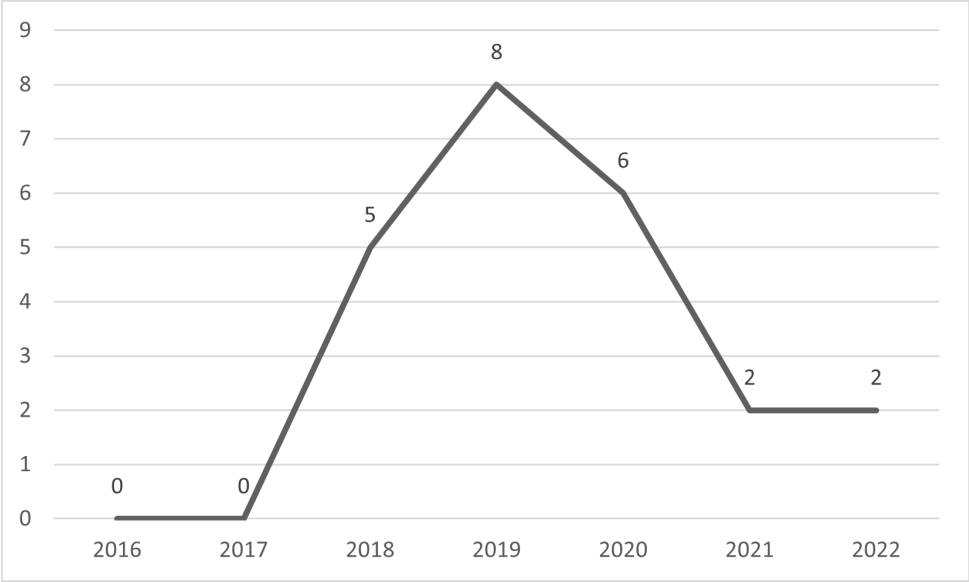
Tabela 3 – continua da página anterior

Título	Base
Meyer et al. (2018)	ScienceDirect
Yang et al. (2019a)	Web of Science
Mahmoud, Elbeh and Abdlkader (2018)	IEEE Xplore
Huang et al. (2019)	ScienceDirect
Ershoff et al. (2020)	ScienceDirect
Hooley et al. (2021)	ScienceDirect
Zhou (2022)	IEEE Xplore
Yu et al. (2022)	Web of Science

Elaborado pelo autor.

Para uma análise detalhada, gerou-se alguns gráficos comparativos com relação ao ano das publicações selecionadas e as bases encontradas. Um detalhe importante para ressaltar é que o ano de 2022 é referente somente as publicações realizadas até o mês de novembro, data em que foi realizada a busca desta revisão. A Figura 8 apresenta o gráfico comparativo do ano das publicações.

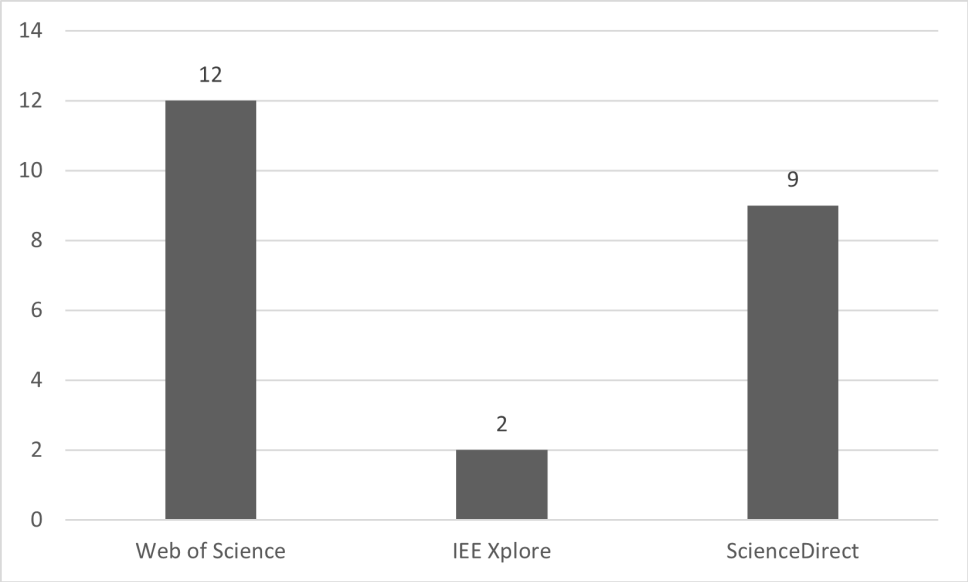
Figura 8 – Comparativo do ano de publicação dos artigos selecionados



Fonte: Elaborado pelo autor

Através do gráfico é possível verificar que o estudo de ER e DL obteve destaque no passar dos anos, tendo crescido muito o interesse nas técnicas. A Figura 9 ilustra o gráfico comparativo das bases de dados.

Figura 9 – Comparativo da base de dados dos artigos selecionados



Fonte: Elaborado pelo autor

A base de dados da Web Of Science obteve grande êxito quanto as publicações desta pesquisa, seguindo da base ScienceDirect. Vale destacar que a maioria dos artigos duplicados ocorreram entre estas duas bases, fazendo com que a ScienceDirect obtivesse mais relevância na seleção pela quantidade de informações que oferecia. Mas, de forma geral, pode-se dizer que ambas as bases tiveram uma proporção semelhante quanto ao êxito dos resultados. Já a bases IEEE Xplore não obteve muita relevância para este estudo, tendo, no final, uma seleção de dois artigos. Também foi possível listar as revistas onde os artigos selecionados foram publicados, visando pesquisar a relevância da revista para a possibilidade de uma publicação. Verificou-se que a revista "*Journal Of The American Medical Informatics Association*" possui a maior quantidade de publicações sobre o tema da pesquisa. A Tabela 4 ilustra todas as revistas onde foram publicados os artigos.

3.6.1 Análise dos artigos

Conforme os dados obtidos nos artigos, foi possível responder a todas as perguntas pré-estabelecidas nessa revisão sistemática. Essa Seção apresentará uma análise aprofundada de cada questão, buscando apresentar os resultados, causas e observações sobre os artigos selecionados para a fase final.

3.6.1.1 Os artigos foram aplicados à área da saúde?

Com base no objetivo da proposta, procurou-se analisar a aplicação dos trabalhos em alguma área da saúde. Apesar da palavra Saúde (*Health*) ter sido utilizada de forma opcional nas

Tabela 4 – Revistas das quais os artigos selecionados foram publicados

Revista	**Qtd.**
Artificial Intelligence In Medicine	3
Bmc Bioinformatics	3
Bmc Medical Informatics And Decision Making	1
Clinical Breast Cancer	1
Computer Methods And Programs In Biomedicine	2
Drug Safety	1
International Computer Engineering Conference (ICENCO)	1
Journal Of Biomedical Informatics	3
Journal Of King Saud University - Computer And Information Sciences	1
Journal Of The American Medical Informatics Association	4
The Lancet Digital Health	1
The Lancet Respiratory Medicine	1
Transplantation Proceedings	1

Fonte: Elaborado pelo autor.

pesquisas, em todos os trabalhos que passaram nos requisitos havia citação da palavra.

Sendo assim, os trabalhos recuperados apresentam aplicações na área da saúde, sendo detalhados com relação à área de aplicação mais específica no item que indica a área de atuação do trabalho.

3.6.1.2 Qual o objetivo da pesquisa?

Buscou-se analisar os objetivos das pesquisadas realizadas pelos autores, visando identificar a aplicação da qual foram executadas e planejadas. A tabela 5 demonstra os objetivos e os autores que as utilizaram.

Tabela 5: Objetivo dos artigos selecionados

Resposta	**Autor(es)**
Auxílio no Diagnóstico	Lin et al. (2020) e Khan and Shamsi (2021)
Avaliar a Eficácia do Modelo de Aprendizagem de Máquina	Li and Yu (2019)
Classificar Consulta	Lamy et al. (2019), Manley et al. (2020), Hooley et al. (2021) e Li et al. (2019)
Detecção de medicamentos	Yang et al. (2019a), Fabregat, Araujo and Martinez-Romo (2018), Yang et al. (2020) e Wei et al. (2020)
Continua na próxima página	

Tabela 5 – continua da página anterior

Resposta	**Autor(es)**
Experimentos e Análises	Mahmoud, Elbeh and Abdlkader (2018), Huang et al. (2019), Ershoff et al. (2020), Purushotham et al. (2018), Lamurias et al. (2019), Yoon et al. (2019), Suárez-Paniagua et al. (2019), Yu et al. (2022), Zhou (2022) e Christopoulou et al. (2020)
Prevenção de complicações	Meyer et al. (2018)
Mineração de Texto	Lee et al. (2020)

Elaborado pelo autor.

Quanto aos objetivos, 43,48% dos autores tinham como objetivo vencer um desafio lançado em programas de aprendizagem, buscando fazer experimentos e análises para comparar resultados com outros trabalhos dos eventos. A busca pela detecção de medicamentos, principalmente na área da Biomedicina, foi observada em 17,39% dos artigos. A classificação de consultas foi observada em 17,39% dos autores, sendo elas classificações de características em notas clínicas, riscos ou casos clínicos. Apenas 8,69% dos trabalhos buscaram aplicar sua pesquisa para auxiliar diagnósticos clínicos, extraindo relações consideradas importantes para os médicos. Por fim, 4,34% dos artigos tem como objetivo a mineração do texto, a avaliação da eficácia do modelo desenvolvido, comparando com outros modelos similares ou trabalharam com prevenção de complicações de saúde. A Figura 10 ilustra um gráfico com esses resultados.

3.6.1.3 Apresentou uso de ontologias?

Com base no objetivo desse trabalho, buscou-se identificar os trabalhos que utilizaram, de alguma forma, ontologias em suas pesquisas. A tabela 6 mostra os artigos que fizeram uso de ontologias.

Tabela 6: Uso de ontologias nos artigos selecionados

Resposta	**Autor(es)**
Sim	Lamurias et al. (2019), Lamy et al. (2019) e Mahmoud, Elbeh and Abdlkader (2018)

Continua na próxima página

Tabela 6 – continua da página anterior

Resposta	**Autor(es)**
Não	Christopoulou et al. (2020), Li and Yu (2019), Wei et al. (2020), Suárez-Paniagua et al. (2019), Purushotham et al. (2018), Lee et al. (2020), Yoon et al. (2019), Fabregat, Araujo and Martinez-Romo (2018), Manley et al. (2020), Khan and Shamsi (2021), Yang et al. (2020), Li et al. (2019), Lin et al. (2020), Meyer et al. (2018), Yang et al. (2019a), Huang et al. (2019), Hooley et al. (2021), Yu et al. (2022), Zhou (2022) e Ershoff et al. (2020)

Elaborado pelo autor.

Em relação ao uso de ontologias, verificou-se que 86,96% dos autores não fizeram uso do recurso. Já 13,04% dos autores usaram ontologias em sua pesquisa. Nesses trabalhos, foram construídas ontologias com as relações extraídas de texto biomédico. Estas tinham como objetivo a estruturação dos dados extraídos ou o aproveitamento de dados específicos de domínio para auxiliar na representação de cada entidade como a sequência das relações.

3.6.1.4 O artigo foi aplicado em algum protótipo?

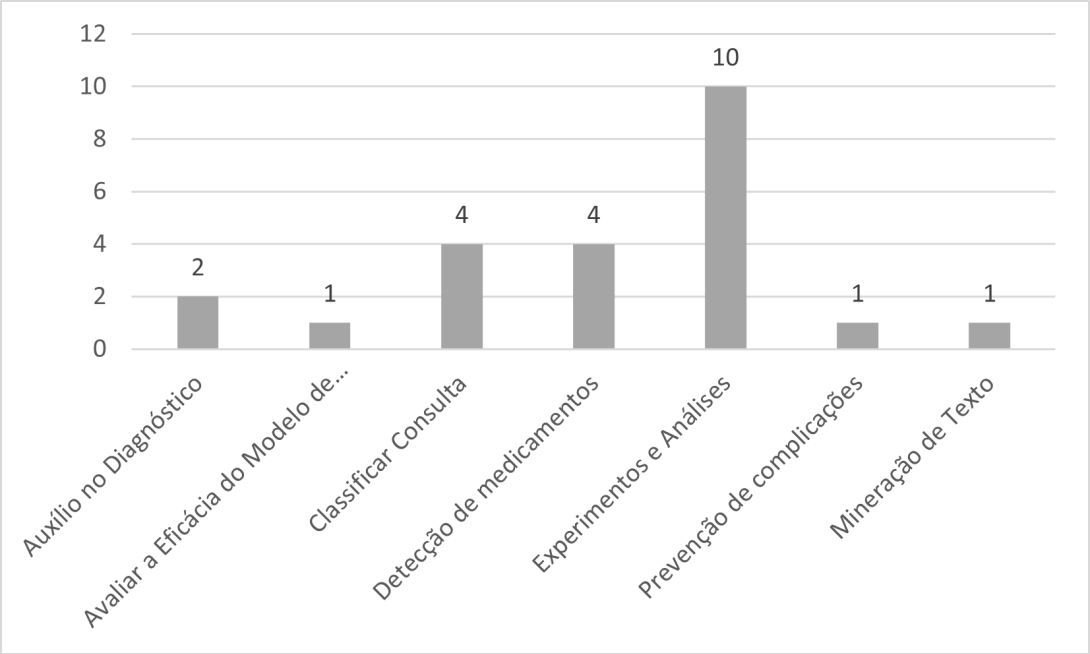
Identificou-se os autores que criaram e utilizaram algum tipo de protótipo para realizar a sua pesquisa. A Tabela 7 mostra se foram desenvolvidos protótipos e os autores que os apresentaram.

Tabela 7: Protótipos dos artigos selecionados

Resposta	**Autor(es)**
Sim	Huang et al. (2019), Yang et al. (2019a), Li and Yu (2019), Lamy et al. (2019), Khan and Shamsi (2021), Lee et al. (2020), Lamurias et al. (2019), Yoon et al. (2019) e Suárez-Paniagua et al. (2019)
Não	Christopoulou et al. (2020), Wei et al. (2020), Purushotham et al. (2018), Fabregat, Araujo and Martinez-Romo (2018), Manley et al. (2020), Yang et al. (2020), Li et al. (2019), Lin et al. (2020), Meyer et al. (2018), Mahmoud, Elbeh and Abdlkader (2018), Hooley et al. (2021), Yu et al. (2022), Zhou (2022) e Ershoff et al. (2020)

Elaborado pelo autor.

Figura 10 – Comparativo de objetivos dos artigos selecionados



Fonte: Elaborado pelo autor

Em relação ao desenvolvimento de protótipos, verificou-se que 39,13% dos autores criaram um protótipo para seu experimento ou tinham como objetivo/pretenção aplicar em um protótipo existente, seguindo de 47,82% dos trabalhos que não construíram ou não ilustraram seus resultados em um protótipo, focando principalmente no desempenho dos seus algoritmos.

3.6.1.5 Qual o idioma alvo?

Quanto ao idioma, buscou-se analisar qual era a linguagem, principalmente em relação ao *dataset*, que os trabalhos focaram. Os artigos descreveram trabalhos com idioma alvo com foco em inglês.

Portanto, quanto ao idioma, verificou-se que 100% dos autores trabalharam com a língua inglesa em seu *dataset*, visto a grande gama de materiais disponíveis para essa linguagem. Isso ressalva um grande diferencial dessa pesquisa e trabalhar com dados em português.

3.6.1.6 Quais relações foram extraídas?

Buscou-se identificar quais eram os tipos de relações que os autores extraíram de suas pesquisas, identificando assim, uma grande variedade de tipos de extração. A Tabela 8 mostra as relações extraídas e seus autores.

Tabela 8: Relações extraídas dos artigos selecionados

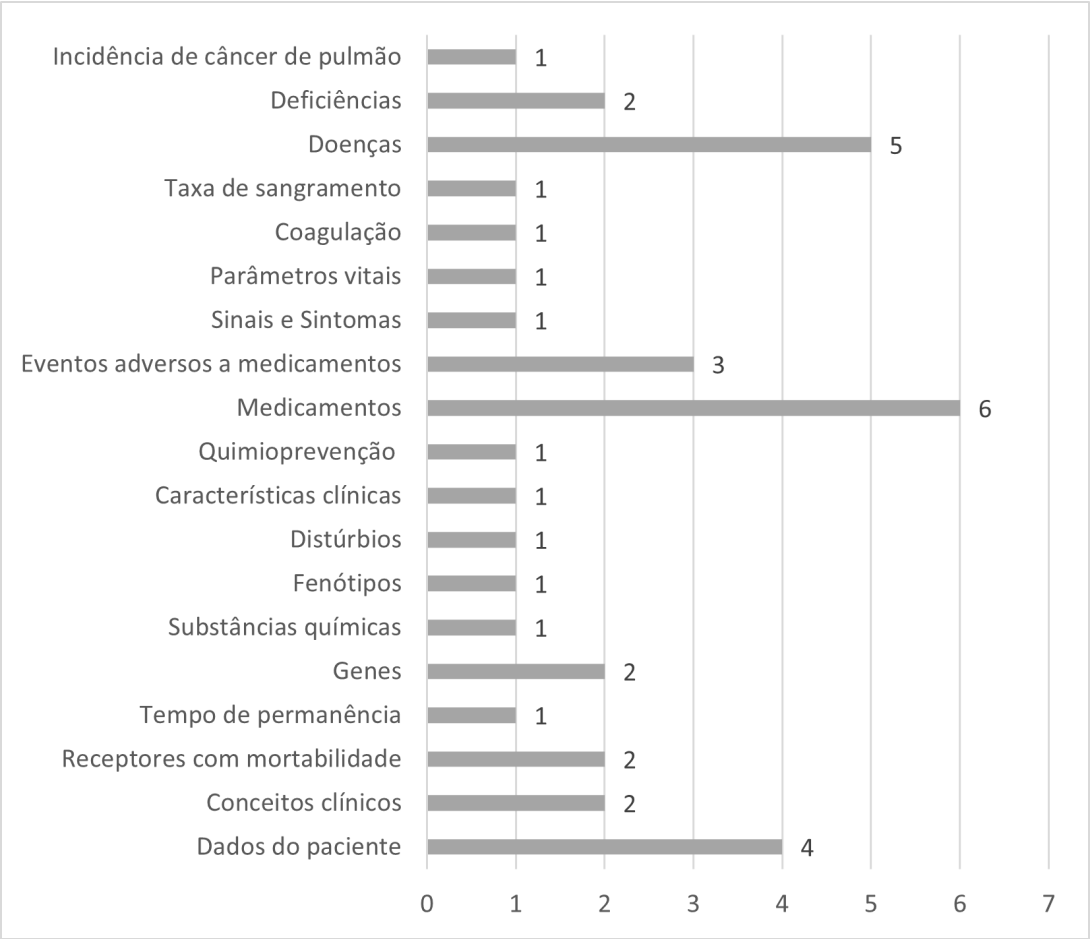
Resposta	**Autor(es)**
Dados do paciente	Lamy et al. (2019), Zhou (2022), Hooley et al. (2021) e Meyer et al. (2018)
Conceitos clínicos	Yang et al. (2019a) e Mahmoud, Elbeh and Abdlkader (2018)
Receptores com mortabilidade	Ershoff et al. (2020) e Purushotham et al. (2018)
Tempo de permanência	Purushotham et al. (2018)
Genes	Lamurias et al. (2019) e Yoon et al. (2019)
Substâncias químicas	Lamurias et al. (2019)
Fenótipos	Lamurias et al. (2019)
Distúrbios	Lamurias et al. (2019)
Características clínicas	Manley et al. (2020)
Quimioprevenção	Manley et al. (2020)
Medicamentos	Li and Yu (2019), Lee et al. (2020), Yang et al. (2020), Suárez-Paniagua et al. (2019), Wei et al. (2020) e Christopoulou et al. (2020)
Eventos adversos a medicamentos	Li and Yu (2019), Yang et al. (2020) e Wei et al. (2020)
Sinais e Sintomas	Lin et al. (2020)
Parâmetros vitais	Meyer et al. (2018)
Coagulação	Meyer et al. (2018)
Taxa de sangramento	Meyer et al. (2018)
Doenças	Lee et al. (2020) Fabregat, Araujo and Martinez-Romo (2018), Yu et al. (2022), Khan and Shamsi (2021) e Li et al. (2019)
Deficiências	Fabregat, Araujo and Martinez-Romo (2018) e Li et al. (2019)
Incidência de câncer de pulmão	Huang et al. (2019)

Elaborado pelo autor.

Procurou-se identificar os tipos de relações que os autores extraíram de suas pesquisas, identificando assim vários tipos de extração. Quanto à extração, 26,08% dos trabalhos selecionados extraíram dados de medicamentos, com uma ampla variação de trabalhos voltados para a biomedicina. Um total de 21,74% dos trabalhos extraíram dados de doenças em geral. 17,38% dos trabalhos extraíram os dados do paciente. Apenas 8,68% extraíram dados de deficiências, con-

ceitos clínicos, mortalidade ou informações genéticas. Por fim, 4,34% dos autores trabalharam com o tempo de internação dos pacientes, com a extração de substâncias químicas, extração de fenótipos, dados sobre distúrbios, dados sobre características clínicas, quimioprevenção, sinais e sintomas, parâmetros vitais, coagulação, taxa de sangramento ou especificamente incidências de câncer de pulmão. A Figura 11 ilustra um gráfico com esses resultados.

Figura 11 – Comparativo de extrações dos artigos selecionados



Fonte: Elaborado pelo autor

3.6.1.7 Qual a área de atuação?

Também, buscou-se identificar quais as áreas de atuação que os autores trabalharam em suas pesquisas, ou se trabalharam de forma geral. A tabela 9 mostra as áreas de atuação e seus autores.

Tabela 9: Área de atuação dos artigos selecionados

Resposta	**Autor(es)**
Câncer de Pulmão	Huang et al. (2019)
Medicamentos	Yang et al. (2019a), Yang et al. (2020), Christopoulou et al. (2020), Wei et al. (2020) e Suárez-Paniagua et al. (2019)
Geral	Lin et al. (2020), Khan and Shamsi (2021), Zhou (2022) e Li et al. (2019)
Terapia Intensiva	Meyer et al. (2018)
Câncer de Mama	Mahmoud, Elbeh and Abdlkader (2018), Lamy et al. (2019) e Manley et al. (2020)
Transplante de Figado	Ershoff et al. (2020)
Oncologia	Li and Yu (2019)
Cardiologia	Li and Yu (2019)
Transgênero e Inconformidade de Gênero	Hooley et al. (2021)
Mortabilidade	Purushotham et al. (2018)
Biomedicina	Lee et al. (2020), Lamurias et al. (2019) e Yoon et al. (2019)
Doenças Raras	Fabregat, Araujo and Martinez-Romo (2018)
Retinopatia diabética	Yu et al. (2022)

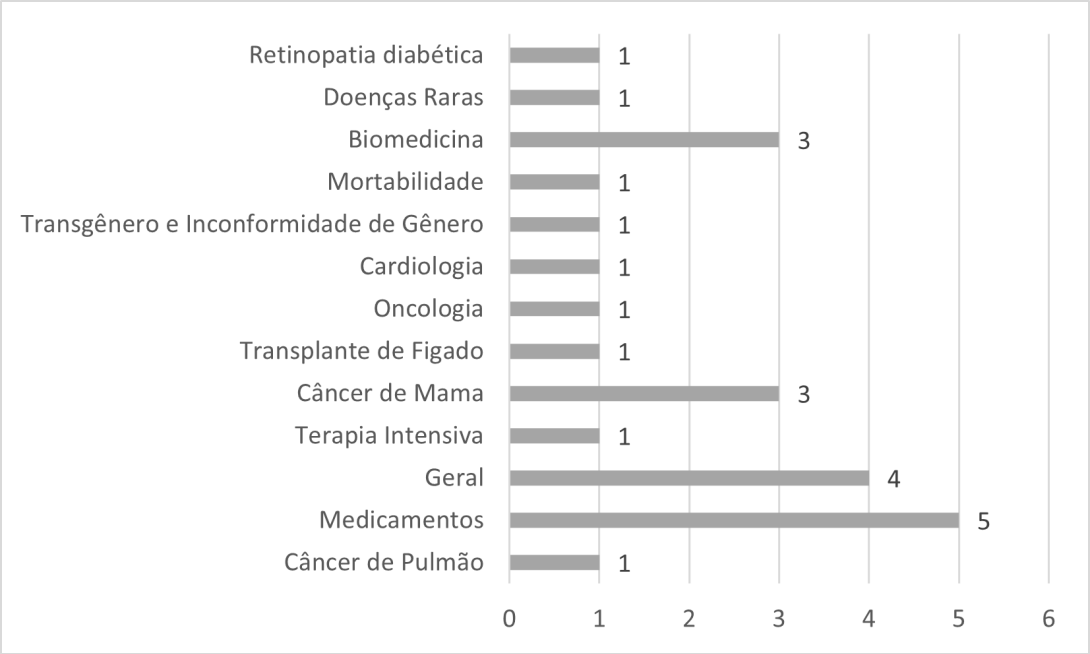
Elaborado pelo autor.

Quanto a área de atuação, verificou-se que 21,74% dos trabalhos focaram especificamente na área de medicamentos (biomedicina). Um percentual de 17,39% dos artigos não trabalharam com uma área específica, sendo classificados como geral. Apenas 13,04% dos trabalhos focaram em câncer de mama. Outros 13,04% também focaram na área da biomedicina, mas de forma geral. Por fim, 4,34% dos trabalhos focaram em terapia intensiva, em transplante de fígado, na oncologia, na relação com a mortabilidade, em doenças raras ou em câncer de pulmão. A Figura 12 ilustra um gráfico com esses resultados.

3.6.1.8 Qual a origem do *dataset*?

Buscou-se analisar a origem dos *dataset* utilizados nas pesquisas, a fim de identificar os tipos de dados e anotações que podem ser uteis nesse trabalho. A Tabela 10 ilustra os *datasets* utilizados e seus autores.

Figura 12 – Comparativo das áreas de atuação dos artigos selecionados



Fonte: Elaborado pelo autor

Tabela 10: Origem do *dataset* dos artigos selecionados

Resposta	**Autor(es)**
DESIREE (DESIMS)	Lamy et al. (2019)
GENIA 7	Mahmoud, Elbeh and Abdlkader (2018)
MADE 1.0	Li and Yu (2019)
Próprio	Khan and Shamsi (2021), Yang et al. (2019a) e Manley et al. (2020)
Dados do Medscape eMedicine	Lin et al. (2020)
Prontuários Eletrônicos	Hooley et al. (2021), Zhou (2022) e Meyer et al. (2018)
STAR	Ershoff et al. (2020)
Medical Information Mart for Intensive Care III	Purushotham et al. (2018)
Resumos do PubMed	Lee et al. (2020)
Desafios	Huang et al. (2019), citetextoyu2022identify, Lamurias et al. (2019), Li et al. (2019) e Suárez-Paniagua et al. (2019)
Resumos Científicos	Fabregat, Araujo and Martinez-Romo (2018)

Continua na próxima página

Tabela 10 – continua da página anterior

Resposta	**Autor(es)**
MIMIC-III	Yang et al. (2020), Christopoulou et al. (2020) e Wei et al. (2020)
BC2GM	Yoon et al. (2019)
BC4CHEMD	Yoon et al. (2019)
BC5CDR	Yoon et al. (2019)
JNLPBA	Yoon et al. (2019)
NCBI	Yoon et al. (2019)
I2b2 / VA	Li et al. (2019)

Elaborado pelo autor.

Conforme a análise, há uma grande variedade quanto ao uso dos *datasets*. Um percentual de 21,74% dos trabalhos utilizaram um *dataset* que foi disponibilizado em desafios de programação. Já 13,04% utilizaram *dataset* desenvolvidos por eles mesmos ou utilizaram o *dataset* disponibilizado pelo projeto MIMIC-III. 13,04% dos artigos usaram dados de sistemas EHR. Os restantes 4,34% dos artigos utilizaram Dados do Medscape eMedicine como *dataset*, utilizaram dados de Prontuários Eletrônicos, dados de resumos do PubMed, dados de resumos científicos ou dados disponibilizados em outros eventos, tais como: DESIREE (DESIMS), GENIA 7, MADE1.0, STAR, *Medical Information Mart for Intensive Care III*, BC2GM, BC4CHEMD, BC5CDR, JNLPBA, NCBI ou I2b2/VA. A Figura 13 ilustra um gráfico com esses resultados.

3.6.1.9 Quais são as abordagens, técnicas e métodos usados?

Para determinar a metodologia empregada na ER, uma análise minuciosa foi realizada considerando as técnicas, abordagens e métodos adotados em cada estudo. As tabelas 11, 12 e 13 demonstram quais são as técnicas, abordagens, métodos e os autores que as utilizaram.

Tabela 11: Modelos dos artigos selecionados

Resposta	**Autor(es)**
Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)	Mahmoud, Elbeh and Abdlkader (2018)
Bidirectional Long Short-Term Memory - Conditional Random Field (BiLSTM-CRF)	Yoon et al. (2019) e Wei et al. (2020)
	Continua na próxima página

Tabela 11 – continua da página anterior

Resposta	**Autor(es)**
Campo aleatório condicional (*Conditional Random Field* - CRF)	Suárez-Paniagua et al. (2019), Yang et al. (2019a) e Christopoulou et al. (2020)
Classificação multirrótulo (CMR)	Khan and Shamsi (2021)
Classificador Softmax	Li et al. (2019) e Suárez-Paniagua et al. (2019)
Long short-term memory (LSTM)	Yu et al. (2022) e Yang et al. (2019a)
Máquinas de vetores de suporte (*Support Vector Machines* - SVM)	Yang et al. (2019a)
Modelo perceptron multicamadas (*Multilayer Perceptron* - MLP)	Li and Yu (2019)
BiLSTM	Li et al. (2019), Suárez-Paniagua et al. (2019) e Christopoulou et al. (2020)
Rede Neural Convolucional	Yoon et al. (2019), Li et al. (2019) e Suárez-Paniagua et al. (2019)
Rede Neural Recorrente	Yang et al. (2019a) e Meyer et al. (2018)
Rede neural recorrente com unidades de memória de longo prazo (*Back-off - Long Short-Term Memory* - BO-LSTM)	Lamurias et al. (2019)
Rede *Transformer*	Christopoulou et al. (2020)
Redes Neurais Profundas (RNP)	Ershoff et al. (2020)
Outros	Hooley et al. (2021) e Zhou (2022)
Weighted K-nearest Neighbor (WKNN)	Lamy et al. (2019)

Elaborado pelo autor.

Tabela 12: Recursos dos artigos selecionados

Resposta	**Autor(es)**
BERT	Yu et al. (2022) e Lee et al. (2020)
Dicionário	Khan and Shamsi (2021)
Deepl	Huang et al. (2019)
Embeddings	Mahmoud, Elbeh and Abdlkader (2018), Lamurias et al. (2019) e Yoon et al. (2019)
Ontologias	Lamy et al. (2019), Lamurias et al. (2019) e Mahmoud, Elbeh and Abdlkader (2018)
	Continua na próxima página

Tabela 12 – continua da página anterior

Resposta	**Autor(es)**
Super Learner (SL)	Purushotham et al. (2018)
Word2Vec (W2V)	Mahmoud, Elbeh and Abdlkader (2018) e Lee et al. (2020)
WordNet	Lamurias et al. (2019)

Elaborado pelo autor.

Tabela 13: Outros tipos de abordagens dos artigos

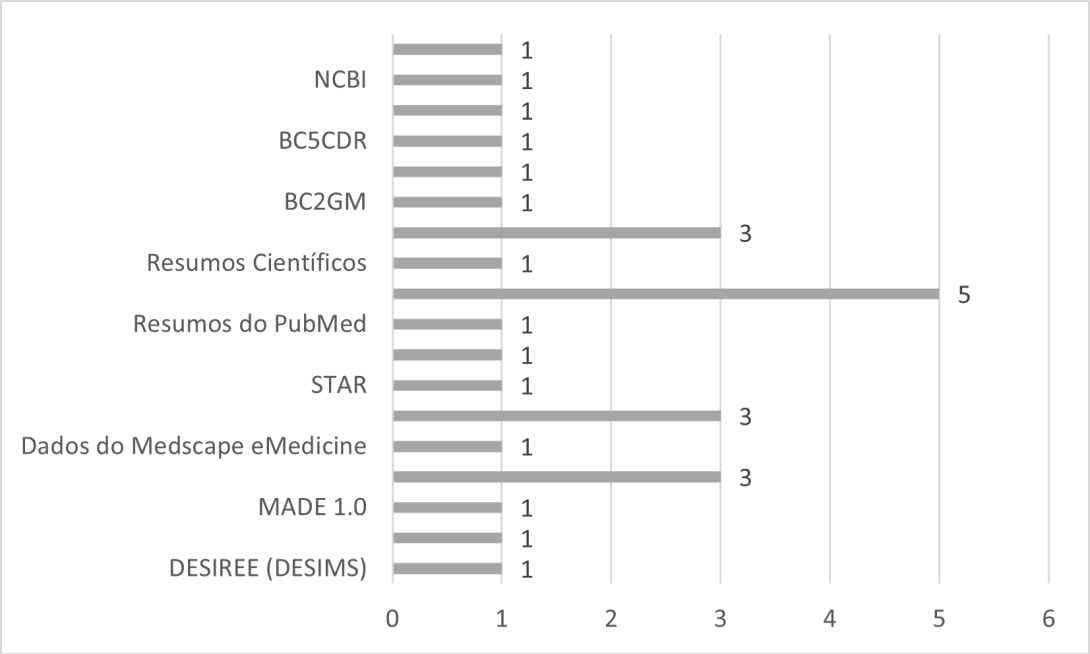
Resposta	**Autor(es)**
CapNet	Li and Yu (2019)
Freebase	Lin et al. (2020)
Multidimensional Scaling (MDS)	Lamy et al. (2019)
Reconhecimento de entidades nomeadas	Yang et al. (2019a), Lee et al. (2020), Suárez-Paniagua et al. (2019), Wei et al. (2020) e Christopoulou et al. (2020)

Elaborado pelo autor.

Um total de 26,08% dos artigos utilizaram Campo Aleatório Condicional (CRF) como recurso em sua pesquisa, seguido de 21,74% dos autores que fizeram uso do REN. Já 13,04% dos trabalhos utilizaram ou uma BiLSTM ou uma CNN para estruturação da RN. Também, 13,04% do autores fizeram uso de *embeddings* e Ontologias em seus experimentos. Já 8,69% dos trabalhos utilizaram uma BiLSTM-CRF & 8,69% ou uma RNR em suas pesquisas, seguido também do uso do Word2Vec (W2V) ou do Classificador Softmax. 8,69% usaram *Transformers* e 8,69% dos trabalhos usaram uma LSTM. Por fim, 4,34% dos autores utilizaram uma variedade grande de recursos em seus trabalhos, tais como: CapNet, Classificação Multirrótulo (CMR), Dicionário, *Freebase*, LSTM, Máquinas de vetores de suporte (SVM), Modelo perceptron multicamadas (MLP), *Multidimensional Scaling* (MDS), Rede neural recorrente com unidades de memória de longo prazo (BO-LSTM) e, Redes *Transformers*, Redes Neurais Profundas (RNP), *Super Learner* (SL), *Weighted K-nearest Neighbor* (WKNN), *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH), BERT e, para finalizar, WordNet.

Na análise, é possível identificar que há uma grande variação de técnicas exploradas nos trabalhos correlatos, que vão desde métodos como SVM, MLP, até Redes Recorrentes e *Trans-*

Figura 13 – Comparativo de *dataset* dos artigos selecionados



Fonte: Elaborado pelo autor

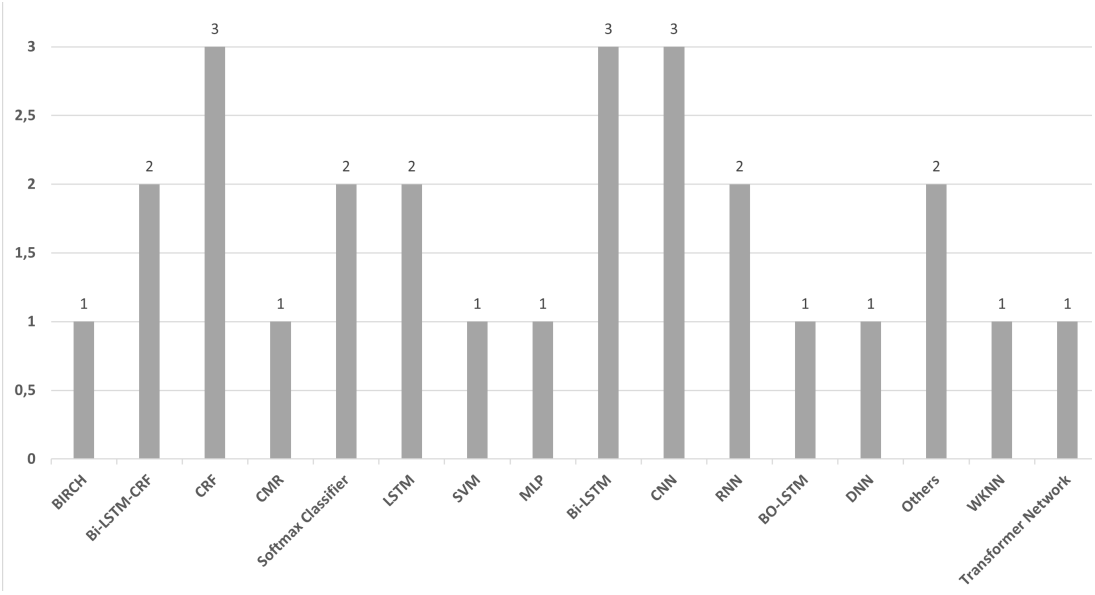
formers. Além dos métodos, o estudo destes artigos permitiu observar as principais bibliotecas utilizadas, tais como *keras*, *scikit-learn* e *spacy*. Junto a estas bibliotecas, destaca-se o uso de recursos adicionais, tais como dicionários, ontologias, léxicos (como, por exemplo, o *Wordnet*) e uma grande variação de *embeddings* e modelos de linguagem (tais como BERT, Word2vec, etc). Destaca-se também o uso de recursos não populares, como o MDS (Uma técnica de Escalonamento Multidimensional que é usada para representar espacialmente, em 2D ou 3D, uma matriz de proximidades), o Freebase (um banco de dados de gráficos criado de forma colaborativa para a estruturação do conhecimento humano), e o Capnet, uma rede cápsula de aprendizagem profunda para extração de relação de domínio único e modos totalmente compartilhados). As Figuras 14, 15 e 16 ilustram os gráficos com esses resultados.

3.7 Lacunas e Desafios

Acesso a dados clínicos em tempo real, segurança de dados, aprovação médica, os resultados gerados e a avaliação de desempenho são aspectos importantes da implementação de uma estratégia de dados baseada em DL. Uma abordagem que considera os dados clínicos pode melhorar a precisão das estratégias estabelecidas. Por exemplo, no contexto do câncer de mama, várias aplicações de DL foram propostas para prever o risco do paciente, mostrando bons desempenhos (MAHMOUD; ELBEH; ABDLKADER, 2018) (LAMY et al., 2019) (MANLEY et al., 2020). Algoritmos de ML processaram rapidamente grandes conjuntos de dados e forneceram *insights* úteis sobre o conhecimento em relação a serviços de saúde.

Com a análise da seção anterior, foi possível identificar questões importantes para esse tra-

Figura 14 – Comparativo de Métodos dos artigos selecionados



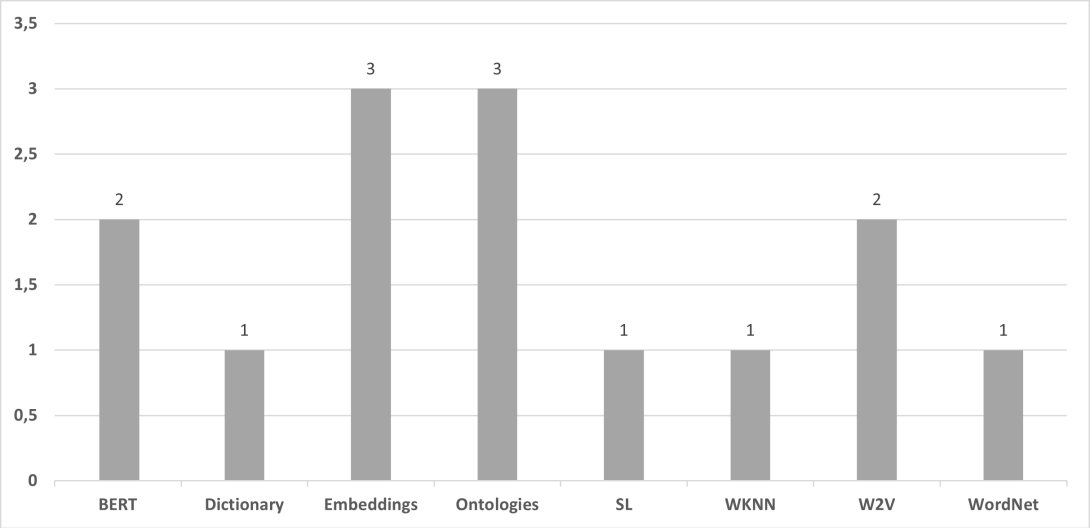
Fonte: Elaborado pelo autor

balho, sendo estas o uso de Ontologias e as possibilidades do modelo *Transformer*. No que se refere a ontologias, identificou-se o uso ainda incipiente nas pesquisas. Os principais benefícios proporcionados pelo uso de ontologias, conforme citados na literatura, estão relacionadas com oportunidades para explorar o uso de abordagens baseadas em grafos de conhecimento e sua integração com abordagens de DL. O modelo *Transformers* também apresentou poucos exemplos de uso, tendo em vista que é um método recente. Entretanto, apresenta resultados benéficos no que se refere a extração de informações e também pode ser visto como tendência quanto a apoio em tarefas do tipo identificação de entidades e extração de informação.

Outro detalhe que foi identificado está relacionado com os *datasets* trabalhados. Nenhum trabalho selecionado explorou a língua portuguesa, o que consiste em um diferencial para essa pesquisa. Modelos de ER em língua portuguesa ainda apresentam um grande desafio e limitações para utilização, uma vez que os modelos de linguagem usualmente são treinados e aplicado em experimentos em inglês. Este cenário é também uma constante em experimentos no Brasil, sendo prática comum a realização do processo de tradução de notas e narrativas clínicas do português para o inglês, em consequência do ferramental disponível para PLN e atualizações, quando há novas descobertas. Outro diferencial está em relação a exploração de um *dataset* com dados reais, visto que os trabalhos relacionados ou trabalharam com *datasets* prontos de desafios ou eventos (HUANG et al., 2019), (LAMURIAS et al., 2019), ou criaram seus próprios *datasets* para teste (KHAN; SHAMSI, 2021), (YANG et al., 2019a).

Analisou-se também aspectos da área de aplicação dos trabalhos, entre os quais há uma grande gama na área da Biomedicina, principalmente relacionado a medicamentos (YANG et al., 2019a), (YANG et al., 2020). Já na Oncologia, verificou-se que os trabalhos tratam especificamente de um tipo de câncer (como câncer de mama, pulmão, etc...) (HUANG et al.,

Figura 15 – Comparativo de recursos dos artigos selecionados



Fonte: Elaborado pelo autor

2019) (MANLEY et al., 2020), sendo que nessa pesquisa o objetivo é tratar de todas as possíveis ocorrências tratadas em texto médico na área de oncologia.

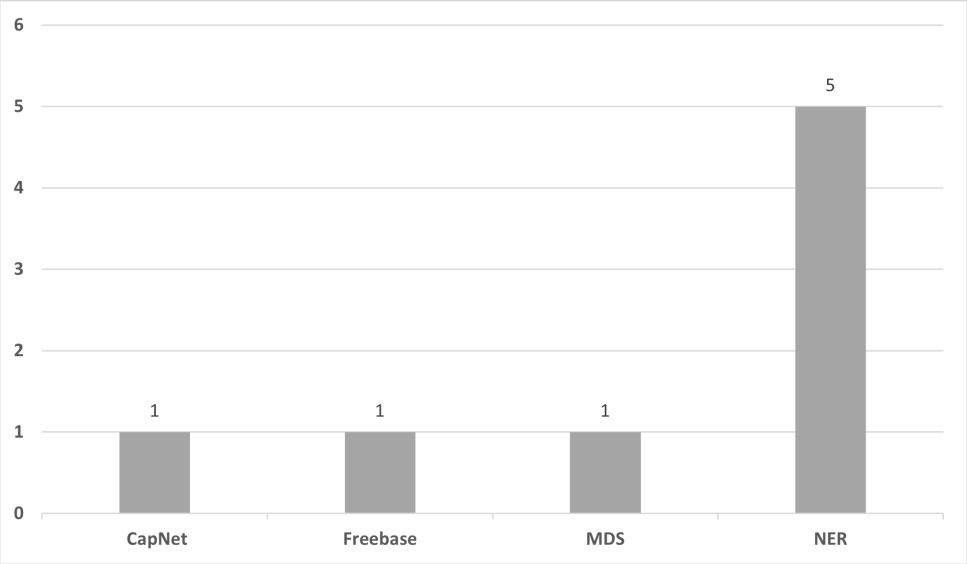
A adaptação da estrutura de aprendizagem para EHR estruturado é uma ideia clara em diversos trabalhos, baseando-se principalmente na analogia entre texto em LN e EHR, ou seja, onde ambos são modalidades sequenciais para *tokens* de um grande vocabulário. Alguns pesquisadores redirecionam diretamente as camadas internas de modelos profundos treinados (por exemplo, RNR) de uma tarefa existente para uma nova tarefa, mas esse aprendizado pode ser muito estreitamente associado a especificações e sua generalização não foi bem estabelecida.

O BERT, incluindo sua arquitetura e sua metodologia de treinamento, vem sendo usado com destaque para modelos de treinamento em grandes base de dados EHR. Notavelmente, outras estruturas de incorporação pré-treinadas contextualizadas no domínio de PLN, como ULMFiT46 e ELMo26, também podem ser observadas no domínio EHR. Existem outros dois estudos relevantes na literatura do domínio clínico: BEHRT (LI et al., 2020a) e G-BERT (MORGAN; RANASINGHE; ZAMPIERI, 2021). Esses modelos, no entanto, possuem algumas limitações, descritas a seguir. BEHRT visa desenvolver modelos pré-treinados para prever a existência de quaisquer códigos médicos em certas evoluções. Usa *embeddings* para distinguir evoluções diferentes e adiciona uma camada de idade para implicar ordens temporais.

O G-BERT aplicou um modelo de rede neural de grafos (*Graph Neural Network* - GNN) para expandir o contexto de cada código clínico por meio de ontologias e treinou em conjunto os *embeddings* GNN e BERT. No entanto, no trabalho as entradas são todas amostras de uma única evolução, que são insuficientes para capturar informações contextuais de longo prazo em EHR. Além disso, o tamanho do conjunto de dados de pré-treinamento não é grande, o que o torna difícil para avaliar todo o seu potencial.

Além disso, nem BEHRT nem G-BERT usam tarefas de previsão de doenças como a ava-

Figura 16 – Comparativo de outras abordagens dos artigos selecionados



Fonte: Elaborado pelo autor

liação de seu modelo pré-treinado por *fine-tuning*. Para a aprendizagem de pré-treinamento, estudos anteriores sobre EHR mostraram alguns sucessos, mas eles se concentraram principalmente em *embeddings* estáticos, como word2vec e GloVe, que falhou em capturar informações de contexto profundo.

Enquanto diversos trabalhos são realizados acerca dos modelos BERT e derivações, existe também a limitação quanto a aplicação destes em língua portuguesa. Algumas práticas para contornar essa deficiência são a tradução dos *tokens* para o inglês, e então a utilização de um modelo treinado para língua inglesa, e neste formato pode-se perder parte das entidades, dependendo do tipo de *dataset* utilizado. Schneider et al. (2020) propõe transferir as informações aprendidas do modelo BERT-multilíngue para um *dataset* de artigos científicos biomédicos e narrativas clínicas em Português Brasileiro. O modelo BioBERTpt foi avaliado com experimentos REN, em dois *dataset* anotados contendo narrativas clínicas e os resultados foram comparados com modelos BERT existentes.

Neste trabalho, serão abordadas as seguintes lacunas que precisam ser tratadas e aprimoradas. Em primeiro lugar, destaca-se a falta de recursos disponíveis para a língua portuguesa atualmente. Em segundo lugar, há uma escassez de trabalhos que utilizam dados da área da saúde, que sejam representativos e provenientes de contextos reais. A maioria dos estudos existentes utiliza dados de outras áreas. O terceiro ponto refere-se às técnicas utilizadas, com poucos trabalhos explorando o uso do BERT e dos *Transformers*, que estão se tornando cada vez mais relevantes no estado da arte atual. Por fim, é importante abordar a estruturação de dados com base na percepção de recursos limitados para a criação de estruturas de dados, como ontologias.

4 ABORDAGEM PROPOSTA

Este capítulo descreve a abordagem proposta. Inicialmente, no item 4.1, uma contextualização geral das motivações do trabalho e de sua conexão com uma aplicação específica em um estudo de caso é apresentada. Em seguida, no item 4.2, são descritos os elementos principais que integram a abordagem e descrita a dinâmica geral de seu funcionamento integrado. Os demais itens descrevem em maiores detalhes os principais elementos envolvidos neste trabalho, tais como o sistema adotado no estudo de caso (4.3), abordagens para tratamento de dados e geração de *datasets* (4.4), os modelos adotados para os experimentos de REN e ER (4.5), a estrutura da ontologia utilizada (4.6) e aspectos de avaliação adotados, levando em conta tanto as questões ligadas à aplicação e estudo de caso como as questões da extração e representação de informação (4.7).

4.1 Contextualização

Por meio do estudo dos trabalhos correlatos, foram identificadas informações importantes para o desenvolvimento desta pesquisa. Conforme observado na literatura, os sistemas EHR são geralmente projetados para manipular informações estruturadas. No entanto, o campo da saúde gera uma grande quantidade de informações em formatos de dados não-estruturados, tais como os registros textuais descrevendo a situação observada e intervenções realizadas em pacientes.

Sendo assim, esse trabalho parte do objetivo de experimentar técnicas de DL, em especial o uso de *Transformers*, BERT e BiLSTM, para REN e ER. Após este reconhecimento as informações são estruturadas uma ontologia. Com isso, estima-se possibilitar o apoio aos profissionais de saúde no que se refere à facilidade de acesso a dados para apoio nas decisões diagnósticas, tratamentos, protocolos e afins.

Conforme visto no item 3.7, há uma lacuna quanto ao uso de dados em português, principalmente se tratando de *Transformers* e BERT. A utilização de dados reais em oncologia agrega valor a essa pesquisa. A estruturação desses dados também constitui-se de uma lacuna de pesquisa, principalmente tratando-se de ontologia como base de conhecimento para apoio de atividades profissionais.

Esta pesquisa conta com a parceria da empresa Interprocess, a qual possui um sistema EHR de oncologia denominado SGO. Dezenas de clínicas e centros de oncologia utilizam esse sistema para controle de sua operação, especialmente na área assistencial. A empresa trabalha com uma grande quantidade de clínicas e hospitais, com dados de atendimentos médicos, de enfermagem e farmacêuticos, entre outras especialidades da saúde. Este trabalho desenvolveu um estudo de caso na área da oncologia, utilizando dados e outros recursos do sistema de modo a avaliar a possibilidade de implementação de componentes de software que permitam a extração automática de entidades e relações de textos médicos e a sua inserção em um formato estruturado em uma ontologia de domínio.

Os *Datasets* Gemed Onco (DGO) são um elemento essencial nesta pesquisa e sua construção constitui uma das contribuições deste trabalho. Os *datasets* foram elaborados a partir de dados reais, anonimizados por um algoritmo desenvolvido especificamente para o sistema EHR e anotados por especialistas em saúde. A construção do modelo de ER e a estruturação dos dados em uma ontologia também constituem contribuições deste trabalho.

4.2 Visão geral da abordagem proposta

Nesta seção, uma visão geral do contexto da abordagem proposta é descrita. A Figura 17 apresenta essa visão contendo os elementos presentes nesse contexto, que considera clínicas médicas e expressa as necessidades dos profissionais de saúde.

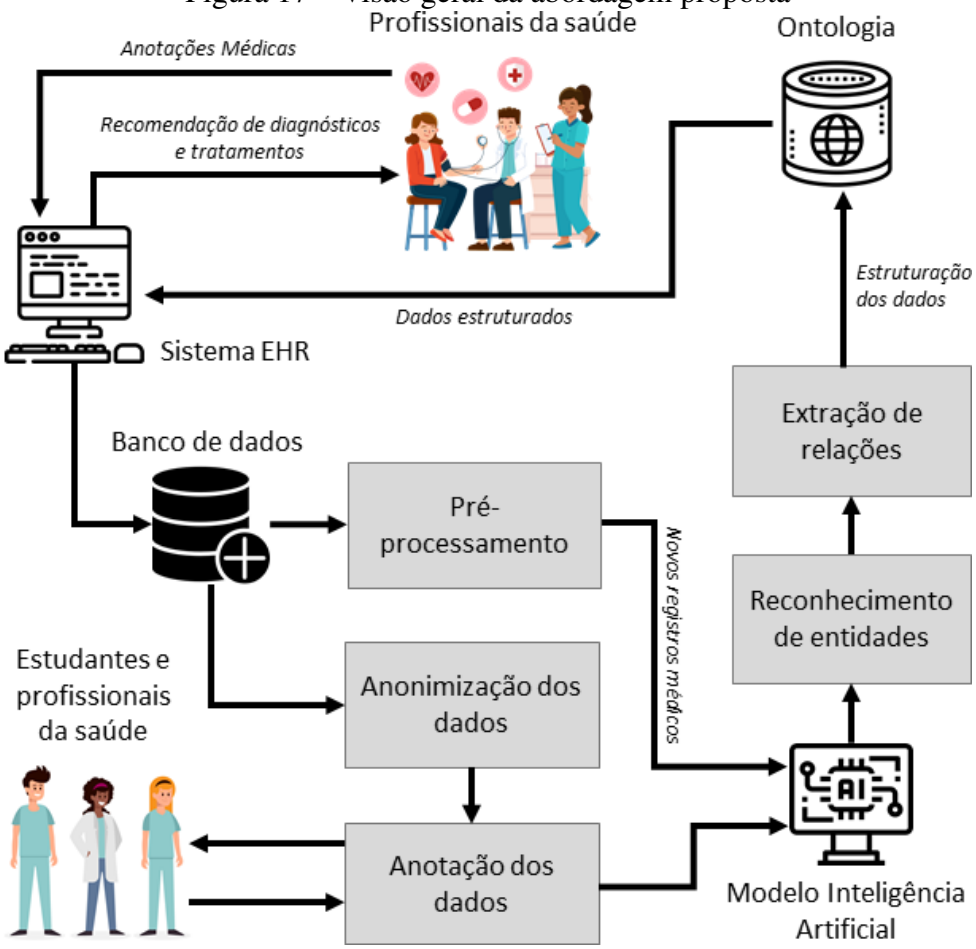
Este trabalho traz contribuições em dois contextos integrados. Um deles situa-se no nível da aplicação em sistemas EHR para oncologia, ampliando a capacidade desses sistemas para o uso dos dados não estruturados, tais como na integração com sistemas de recomendação, sistemas de auxílio no diagnóstico e tratamento, predição de dados, entre outros. Também apresenta diferenciais para o uso na saúde em relação ao tratamento dos dados, como o uso de técnicas de anonimização de dados, geração de *dataset* com dados oncológicos reais em português e a anotação desses dados. O segundo aspecto de contribuição está relacionado a experimentação e proposta de avanços na computação em algoritmos para REN e ER. Os experimentos com modelos baseados em DL geraram uma precisão superior ou semelhante aos trabalhos relacionados. Desta forma, oportunizaram o uso de dados oncológicos reais em português para integração de dados estruturados em uma ontologia.

O processo geral considera o ponto de início na criação de um registro médico ou anotação de um profissional de saúde em um Sistema EHR. Nos casos observados, essa situação gera registros médicos textuais não estruturados e informações clínicas estruturadas. Estes aspectos estão ilustrados no lado esquerdo superior da Figura 17.

A aplicação da proposta possibilitará atender e dar assistência para profissionais da saúde, sendo eles médicos, enfermeiros e farmacêuticos, que atendem pacientes no campo da oncologia em clínicas e hospitais de diversos pontos do Brasil. O processo geral de uso destes recursos está resumido a seguir. A cada consulta, um especialista registra uma evolução do paciente, na qual são preenchidas informações relevantes para o tratamento do paciente, acompanhado também de outros dados, como diagnóstico, conduta, remédios e outros. Essas observações podem ser compostas de texto livre e dados estruturados. Essas anotações médicas são interligadas em outras áreas, que complementam as informações de acordo com seu domínio de especialidade. Esse itens são ilustrados no lado esquerdo superior da Figura 17.

Após a conclusão desta etapa, as anotações médicas são inseridas em um sistema EHR, descrito a seguir, no item 4.3. Como este trabalho está sendo realizado em parceria com a empresa Interprocess, uma primeira abordagem foi um estudo de caso específico, portanto, considera-se que os profissionais de saúde caracterizados usam o SGO para inserir suas observações sobre o

Figura 17 – Visão geral da abordagem proposta



Fonte: Elaborado pelo autor

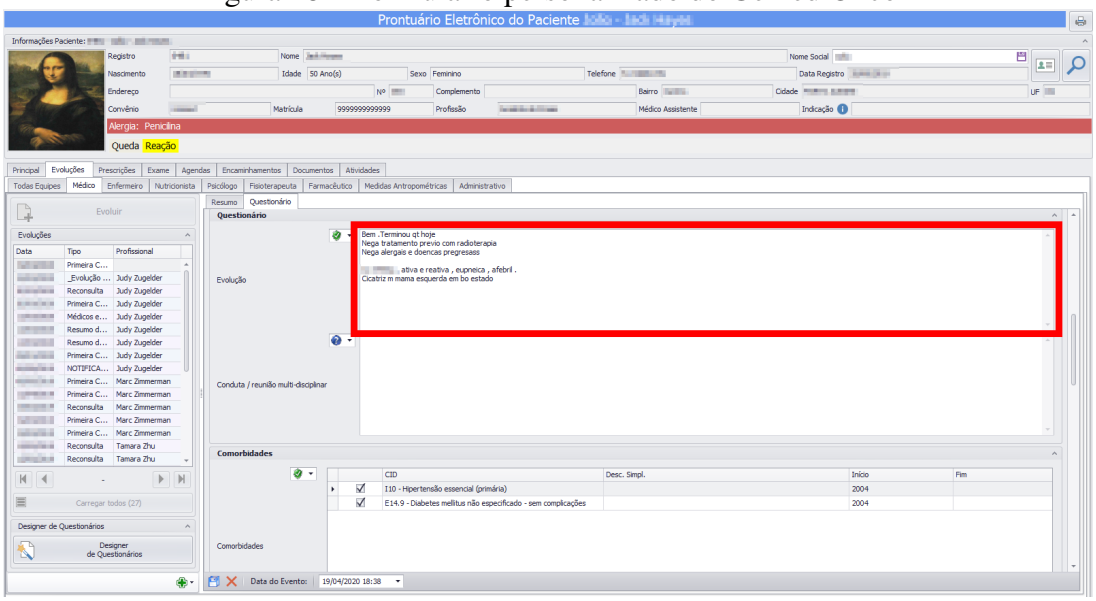
paciente.

Os dados gerados neste processo foram utilizados para a realização das atividades de pré-processamento e anotação de dados, gerando o subsídio necessário para a etapa seguinte, na qual foram estruturados e avaliados modelos para REN e também de ER. Por fim, os resultados das atividades de REN e ER foram empregados na geração de instâncias de uma ontologia de domínio para o estudo de caso.

4.3 Sistema EHR Gemed Onco

Como estudo de caso, foi utilizado o SGO como Sistema EHR. No sistema, as notas médicas contêm eventos clínicos registrados por médicos, enfermeiras e outros profissionais de saúde, com informações clínicas sobre o paciente. Para registrar essas informações, o profissional de saúde usa um formulário personalizável do sistema EHR para cada evento clínico. Esses formulários contêm uma ou mais perguntas, com vários tipos de campos, como números, datas e horas, listas e entradas de texto livre. As entradas de textos livres permitem que os profissionais de saúde insiram textos em LN. A figura 18 apresenta um exemplo de formulário deste sistema.

Figura 18 – Formulário personalizado do Gemed Onco



Fonte: Elaborado pelo autor

A Figura 18 ilustra um formulário personalizável no modo de inserção, com um campo de texto livre destacado dentro da caixa vermelha. Abaixo deste campo, há outro campo de texto livre e um campo estruturado.

Os formulários personalizáveis permitem que os profissionais de saúde criem opções personalizadas de inserção de dados, que são utilizadas para fazer anotações médicas. Durante o processo de criação do formulário, os profissionais de saúde (neste caso, ”usuários do sistema EHR” ou apenas ”usuários”) podem registrar suas dúvidas para aplicá-las e respondê-las durante um evento clínico, como uma consulta. Cada pergunta possui um rótulo e algumas propriedades, que descrevem o tipo de resposta esperada. Estes são os tipos de resposta disponíveis:

- Texto: este tipo possui três subtipos, sendo eles texto, número e data-hora. O tipo de texto é o mais comumente usado e permite que os usuários insiram dados de texto livre com anotações médicas;
- Lista: define que a resposta é um item (pode ser mais de um) de uma lista designada de valores;
- Componente: este tipo representa o uso de componentes predefinidos construídos para casos específicos. Quando usados, é necessário definir qual componente deve ser aplicado. Exemplos de componentes são a caixa de pesquisa ICD, as medidas do paciente (ou seja, altura, peso, área da superfície corporal), o diagnóstico principal do paciente e assim por diante.

O tipo ”componente” permite que o usuário insira as informações por meio de componentes de propósito específicos, como o diagnóstico principal com base no código ICD, medições

do paciente, alergias, medicamentos e assim por diante. Além disso, o usuário pode usar o recurso de prescrição eletrônica para definir o tratamento do paciente com base em protocolos de quimioterapia. Esses recursos permitem que o sistema EHR obtenha os dados e os armazene em um formato estruturado.

O banco de dados do sistema EHR armazena todas as informações sobre a saúde dos pacientes, como anotações médicas, prescrições eletrônicas e histórico de consultas dos pacientes. Para cada clínica de oncologia que utiliza o sistema, é criada uma instância do banco de dados para armazenar suas informações separadamente. A base de dados armazena não só informações clínicas, mas também financeiras, faturamentos de seguro saúde, controles de farmácia e relatórios gerenciais. As informações clínicas são segmentadas em cinco seções:

- A identificação pessoal contém a identificação do paciente e do profissional;
- Estrutura de formulários personalizáveis contém as informações sobre o formulário personalizado, seus grupos de perguntas e suas perguntas individuais;
- As informações clínicas estruturadas de EHR são um grupo de tabelas que contêm as informações clínicas do paciente em um formato estruturado. Foram consideradas apenas as tabelas utilizadas neste trabalho;
- A prescrição de EHR é um grupo de tabelas com informações sobre a prescrição de quimioterapia do paciente. Cada prescrição é gerada por um ou mais protocolos, que contêm os esquemas de medicamentos para gerar o tratamento de quimioterapia do paciente;
- A estrutura de eventos clínicos contém as informações sobre as consultas ao paciente ou qualquer informação sobre um evento relacionado ao paciente.

Depois de explicado como é feito o armazenamento dos dados descritos nessa sessão, parte-se para a análise de dados, descritos na próxima seção.

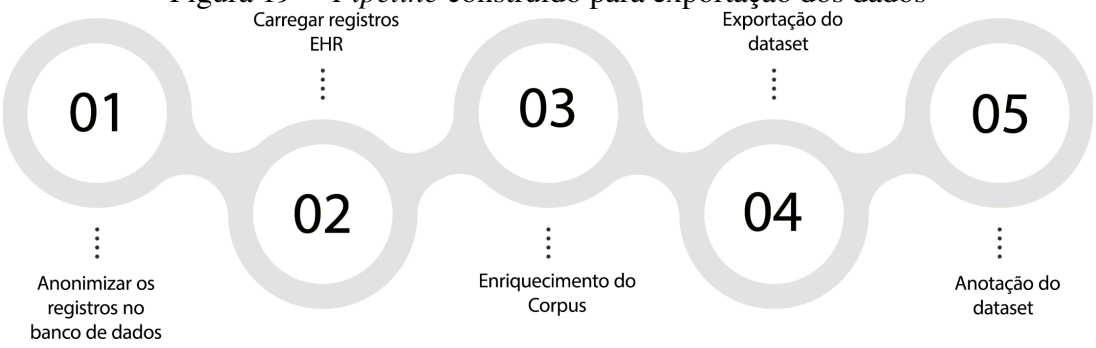
4.4 Exportação, Anonimização e Anotação dos dados

A seguir serão descritos os procedimentos gerais para geração dos *datasets* que serão utilizados para treinar o modelo e para a realização de experimentos preliminares. Esse processo foi exclusivamente realizado para possibilitar a realiação de experimentos. Um *pipeline* foi construído para realizar o processamento dos dados de interesse, conforme ilustrado na Figura 19.

As seguintes etapas compõe o *pipeline*: Anonimização dos dados do sistema EHR; Carregamento dos dados do banco de dados; Enriquecimento do *dataset* com informações clínicas estruturadas; Exportação do *dataset*; Anotação dos dados com alunos e especialistas na saúde.

O *pipeline* foi implementado usando *scripts* SQL e linguagem de programação Python. O primeiro foi executado no Microsoft SQL Server Management Studio para executar a etapa

Figura 19 – *Pipeline* construído para exportação dos dados



Fonte: Elaborado pelo autor

de pseudonimização, e o último foi um aplicativo de console Python que executa as etapas de carregamento, anotação e enriquecimento.

Na primeira etapa, um *script* SQL definido (consultas e procedimentos) foi executado no banco de dados para identificar suas informações. Esse processo foi aplicado para evitar a identificação exclusiva de pacientes e profissionais antes que essas informações fossem processadas e exportadas. Foi implementada uma aplicação que se conecta ao banco de dados do sistema EHR para consultar os registros com os campos relevantes. Enquanto os registros são extraídos, as etapas de anotação do *dataset* e enriquecimento com informações clínicas estruturadas são aplicadas.

4.4.1 Anonimização

A primeira etapa, necessária para gerar o *dataset* e utilizar a abordagem proposta no futuro, consiste em anonimizar e exportar os registros do banco de dados. Frequentemente, os pesquisadores precisam trabalhar com dados extraídos de bancos de dados para usar o histórico clínico do paciente. No entanto, é necessário aplicar técnicas de agregação sem a divulgação de dados pessoais do paciente para manter sua privacidade. Isso pode ser conseguido com o uso da pseudo-anonimização (PEREIRA et al., 2021).

A anonimização é o processo que remove a associação entre o conjunto de dados de identificação e o titular dos dados. A pseudo-anonimização é um tipo específico de anonimato que remove a associação com um titular dos dados e adiciona uma associação entre um determinado conjunto de características relacionadas ao titular dos dados e um ou mais pseudônimos. A pseudo-anonimização pode ser recuperável ou irrecuperável, ou seja, com ou sem a possibilidade de identificar o titular dos dados. Um esquema recuperável pode ser uma tabela de consulta secreta que, quando autorizada, pode ser usada para identificar o titular dos dados. Da mesma forma, um esquema irrecuperável pode ser uma tabela temporária que é destruída no final do processo.

Nessa pesquisa, foi aplicado um processo de anonimização para evitar a divulgação e identificação única de pacientes e profissionais de saúde. Nesta etapa, um processo de pseudo-

anonimização irrecuperável foi aplicado aos identificadores diretos, ou seja, sem a possibilidade de reverter esse processo e revelar a original identidade do paciente e do profissional. Além disso, os dados demográficos dos pacientes ou profissionais não foram exportados, tornando mais difícil reverter o processo de pseudo-anonimização.

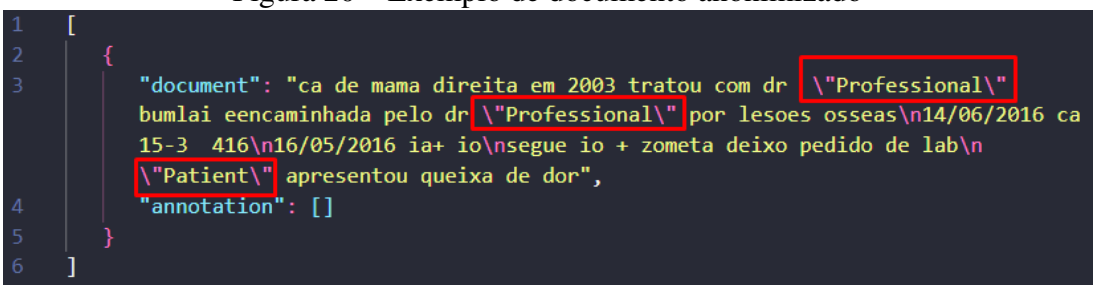
Este processo consiste na troca dos identificadores diretos por pseudônimos e sua aplicação aos dados criados aleatoriamente. Os seguintes campos foram alterados nas tabelas do banco de dados de pacientes e profissionais: nome, número da carteira de identidade do brasileiro (RG, CPF), telefones e endereço. Os dados pseudo-anonimizados consistem em:

- Uma lista de nomes fictícios (nomes e sobrenomes) combinados aleatoriamente;
- Números fictícios de carteira de identidade;
- Números de telefone fictícios;
- Endereços fictícios.

Os campos listados acima foram atualizados com esses novos dados. foi aplicada a pseudo-anonimização dos campos de texto livre, em um processo para remover nomes próprios. Os campos de nome próprio são utilizados a partir das seguintes tabelas de identificação pessoal: Paciente, Profissional, Fornecedores. Uma lista com nomes e sobrenomes foi criada com os nomes dessas tabelas. Essa lista foi usada para encontrar os nomes e sobrenomes nos textos e quando encontrados substituí-los por um nome pseudo-anonimizado.

Na etapa final do *pipeline* ilustrado na Figura 19, após todos os registros serem anonimizados, carregados e enriquecidos, os mesmos são exportados para um formato adequado, juntamente com algumas informações clínicas estruturadas, fazendo com que na próxima etapa possam ser aproveitados os rótulos de dados de texto livre gerados pela equipe médica ao usar o sistema EHR. A Figura 20 ilustra um exemplo de documento anonimizado. Após concluir esse processo, o *dataset* está pronto para ser anotado, conforme descrito na próxima seção.

Figura 20 – Exemplo de documento anonimizado



Fonte: Elaborado pelo autor

4.4.2 Anotação

A Anotação de dados é o processo de rotulagem de dados. Esses dados podem estar em vários formatos, tais como: imagens, vídeo, áudio ou texto, sendo muito utilizada para treinar diferentes modelos de ML. Para o aprendizado de máquina supervisionado, por exemplo, conjuntos de dados rotulados são necessários, de modo que o algoritmo possa compreender facilmente e claramente os padrões de entrada, buscando precisão para uma previsão correta de saída (MICELI; SCHUESSLER; YANG, 2020).

Essa pesquisa contou com o apoio de pesquisadores especialistas na área de saúde da UNISINOS, profissionais da saúde da Inteprocess e de uma empresa especialista em anotação de dados. Em todos estes casos, as equipes utilizaram os dados extraídos do banco de dados e, com ajuda de um software específico, anotaram os dados clínicos. Um fragmento de documento anotado no formato IOB (*Beginning*, *Inside*, *Outside*) é ilustrado na Figura 21.

Figura 21 – Fragmento de documento anotado

```
neo B-Diagnostico
de O
mama B-Orgao
esquerda B-Local_Orgao
# O
cirurgia O
conservadora O
-ap: O
carcinoma O
lobular O
invasivo O
multifocal O
```

Fonte: Elaborado pelo autor

Nesse trecho específico, pode-se identificar diferentes entidades nomeadas relacionadas ao diagnóstico de câncer de mama. A primeira entidade é ”neo”, marcada como B-Diagnóstico, indicando o início da anotação da entidade ”Diagnóstico”. Em seguida, tem a palavra ”de” marcada como O (*Outside*), o que significa que não faz parte de uma entidade nomeada específica.

A próxima entidade é ”mama”, marcada como B-Orgão, indicando que se refere ao órgão ”Mama”. Em seguida, temos ”esquerda” marcado como B-Local Orgão, sugerindo que é uma especificação do local do órgão, indicando que se refere à mama esquerda. Após a palavra ”esquerda”, encontra-se caracteres especiais, representados por ”# o”, que não faz parte de nenhuma entidade nomeada específica.

A seguir, temos as palavras ”cirurgia” e ”conservadora”, ”capro”, ”carcinoma”, ”lobular” e ”invasivo”, todas marcadas como O, indicando que não estão relacionadas a entidades nomeadas específicas.

Após essa etapa, foi possível construir os *datasets* utilizados para o treinamento do modelo proposto, o qual denominaremos nesse trabalho de DGO. Foram gerados três *datasets*, sendo uma versão anotada automaticamente para a aplicada de atividades preliminares, de classificação, denominado de *Dataset* Gemed Onco - PROTOCOLOS (DGO-P). Os outros dois *datasets* foram anotados manualmente, dedicados para as atividades de extração de entidades e de relações, sendo denominados de *Dataset* Gemed Onco - EXAMES (DGO-E) e *Dataset* Gemed Onco - Características do Diagnóstico (DGO-CD). Estes dois *datasets* são descritos em detalhes nas seções 5.1.3 e 5.1.4.

Com os *datasets* anonimizados e anotados, os modelos de DL para REN e de ER puderam ser treinados, conforme descrito na próxima seção.

4.5 Identificação de entidades e extração de relações

Nessa etapa, foi desenvolvida uma abordagem para a extração de entidades e de relações, no domínio de textos oncológicos. A abordagem foi composta por dois modelos: BERT e BiLSTM. O BERT foi treinado para efetuar o reconhecimento das entidades, assim como a extração de características que representarão o texto. Já a BiLSTM foi responsável por, dado um conjunto de entidades identificadas e seus tipos, classificar a relação existente entre elas.

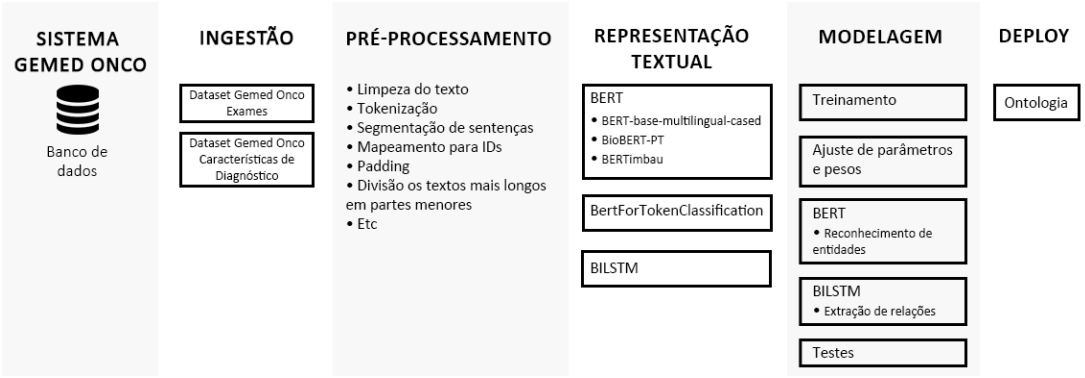
Para avaliar a performance do modelo BERT, diferentes modelos iniciais foram incluídos nos experimentos e o modelo de melhor resultado na tarefa de reconhecimento das entidades foi utilizado para a classificação das relações com a BiLSTM. Os modelos selecionados inicialmente foram o "BERT-base-multilingual-cased" (SOUZA et al., 2021b) (SOUZA et al., 2021b), "BioBERT-PT" (SOUZA et al., 2021b) (JI; WEI; XU, 2020) e "BERTimbau" (LOPES; CORREA; FREITAS, 2021) (SOUZA; NOGUEIRA; LOTUFO, 2020).

Quanto ao modelo BiLSTM, além de parametrizações do tamanho do modelo e quantidade de camadas, também foi realizado um teste com BiGRUs, para avaliar se esta abordagem poderia apresentar resultados melhores em um cenário com menos exemplos de treinamento.

Os modelos foram aplicados nos dados reais dos DGO. Primeiramente foi obtido o melhor resultado possível com o BERT para o reconhecimento das entidades, e depois foi realizado o treinamento da BiLSTM. A Figura 22 ilustra esse processo.

Em resumo, conforme o processo ilustrado na Figura 22, O modelo utiliza do Banco de Dados do SGO (que contém os dados EHR) como ponto de partida. O processo começa com a ingestão de dois conjuntos de dados que foram anotados manualmente: o DGO-E e o DGO-CD. Esses conjuntos de dados contêm informações relevantes para o REN e ER. Em seguida, o pré-processamento dos dados é realizado. Esse processo inclui a limpeza do texto, *tokenização* (divisão em unidades significativas), segmentação de sentenças, mapeamento para IDs (Identificador) únicos, padding (preenchimento para garantir que todas as sequências tenham o mesmo tamanho) e a divisão de textos mais longos em partes menores, se necessário. Essas etapas visam preparar os dados para a representação textual adequada.

Figura 22 – Exemplo de estruturação do modelo



Fonte: Elaborado pelo autor

A representação textual é realizada por meio de diferentes modelos, como os BERTs (BERT base-multilingual-cased, Biobert-pt e BERTimbau), Bertfortokenclassification e a BILSTM. Esses modelos são utilizados para REN e ER. O BERT é aplicado para identificar e classificar as entidades nomeadas nos textos, enquanto a BILSTM é empregada para extrair as relações entre essas entidades. Após o treinamento do modelo e o ajuste de parâmetros e pesos, o processo de deploy é realizado. Nesse estágio, uma ontologia é utilizada para estruturar as informações extraídas dos EHRs. A ontologia fornece uma representação estruturada e organizada das entidades e relações extraídas, facilitando a interpretação e o uso dos dados, da qual é descrita a seguir.

O modelo traz contribuições quanto à avaliação do desempenho do BERT e BiLSTM para REN e ER em textos oncológicos em português no contexto de EHRs. Durante o treinamento, foram explorados diferentes modelos iniciais de BERT, incluindo o ”BERT-base-multilingual-cased”, ”BioBERT-PT” e ”BERTimbau”, com o objetivo de encontrar o modelo com melhor desempenho no domínio específico de EHRs oncológicos em português. Essa seleção foi fundamental para obter resultados mais precisos na tarefa de ER em combinação com a BiLSTM.

É importante destacar o processo de *fine-tuning* realizado com os *datasets* DGO-E e DGO-CD, que desempenhou um papel fundamental na adaptação e melhoria do desempenho dos modelos BERT e BiLSTM para as tarefas específicas de REN e ER, permitindo que os modelos fossem ajustados e calibrados para lidar com as características e nuances presentes nos textos reais relacionados à área oncológica em português. Ao utilizar dados reais dos EHRs para o *fine-tuning*, o modelo foi capaz de aprender com exemplos específicos da área, tornando-se mais adequado para a tarefa proposta. Esse processo contribuiu para melhorar a precisão, resultando em um desempenho mais eficiente e acurado nos experimentos realizados. Além disso, também foram conduzidos testes com a abordagem BiGRUs, buscando avaliar se essa alternativa poderia alcançar resultados melhores em cenários com menos exemplos de treinamento. A investigação de diferentes abordagens e arquiteturas demonstra o esforço em explorar diferentes possibilidades para melhorar o desempenho dos modelos e obter resultados mais robustos e confiáveis.

4.6 Estruturação da Ontologia

Na etapa final os resultados gerados pela etapa de extração de entidades e de relações foram estruturados em uma base de conhecimento. Para essa pesquisa, optou-se pela adaptação e uso de uma ontologia de domínio projetada para representar as entidades, conceitos e relações específicas identificadas na extração dos dados. A construção da ontologia foi iniciada junto a Interprocess e publicada em (SCHWERTNER et al., 2019), sendo ampliada nessa pesquisa.

Com isso, são identificados conjuntos de informações que podem ser obtidos com recursos de PLN, com base nas necessidades mais frequentes de informações complementares que os profissionais de saúde podem usar em seu trabalho diário.

Após a estruturação na ontologia, será possível utilizar os dados integrados ao sistema EHR. Estes poderão ser usados como auxílio na tomada de decisão, recomendação de diagnósticos e tratamentos, dentre outros. Os detalhes da ontologia são descritos na seção 5.5.

A próxima seção descreve a metodologia de avaliação desta pesquisa.

4.7 Metodologia de avaliação

A avaliação do modelo partiu de dois contextos, sendo um contextualizado na área da saúde e outro destinado aos aspectos computacionais. Tendo em vista a área da saúde, o modelo e seus resultados foram validados com especialistas na área da oncologia, em estudos de casos e cenários aplicados ao Gemed Onco. Foram realizados experimentos com coleta de dados em questionários e acompanhamentos com os profissionais para validar sua percepção quanto ao uso do modelo.

Segundo Dey (2001), a avaliação baseada em cenários tem sido utilizada pela comunidade acadêmico-científica para avaliar projetos sensíveis ao contexto e em ambientes ubíquos e baseados na web (SATYANARAYANAN, 2001). Empregar o uso de avaliação baseada em cenários é uma prática adotada em situações onde o sistema em desenvolvimento é resultado de interações entre atores externos, usuários ou outros sistemas (BERGMANN, 2003). Nessa avaliação de cenário, as situações práticas previstas pelo modelo proposto são previamente planejadas e descritas, antes de sua execução no sistema. Em seguida, comenta-se as etapas de sua execução e destaca-se os pontos relevantes para a avaliação do protótipo quanto ao atendimento dos objetivos propostos. O cenário principal adotado visa demonstrar a correta identificação das informações necessárias.

No contexto computacional, foram avaliadas as respostas do modelo utilizando métricas padronizadas, tais como Precisão (*Acurracy*), *Recall*, F1 Score, *Macro AVG* e *Weighted AVG*. Além dessas métricas, os resultados foram comparados com outras pesquisas semelhantes encontradas no estado da arte. Também, o modelo desenvolvido com dados oncológicos busca principalmente melhorar o modelo atual e gerar um melhor resultado. Estes resultados foram avaliados comparando-os com os experimentos e modelos BERT anteriores.

5 EXPERIMENTOS

Ao longo dessa pesquisa, foram exploradas diversas áreas dentro de ML e DL, a fim de experimentar e medir possíveis resultados que possam agregar valor a este trabalho. Para isso, foram realizados experimentos, tais como geração de *datasets*, predição, REN, ER e inserção de dados em ontologia, com base em registros médicos. Neste capítulo são descritos e analisados experimentos e seus resultados.

A estrutura deste capítulo compreende inicialmente a descrição dos *datasets* utilizado, no item 5.1. Em seguida, um experimento inicial de predição é apresentado no item 5.2. O primeiro experimento com extração de relações utilizou um *dataset* aberto e ele é apresentado no item 5.3, tendo sido importante para preparar os experimentos seguintes, que utilizaram os *datasets* anotados neste trabalho. Os experimentos de extração de entidades e de relações com os dois *datasets* anotados neste projeto são descritos no item 5.4. Em seguida são apresentados os experimentos de utilização destes resultados para a inserção de dados em ontologia, no item 5.5.

5.1 Descrição dos *datasets*

Durante esta pesquisa, foram desenvolvidos três *datasets* que contêm informações reais sobre oncologia. Esses dados foram extraídos do SGO e serão descritos a seguir. Um quarto *dataset* é chamado *Drug-Drug Interaction* (DDI) e foi extraído de um evento específico. A descrição do DDI será fornecida para facilitar a compreensão do experimento no qual o mesmo foi utilizado.

5.1.1 *Dataset* Gemed Onco - Protocolos

Para proporcionar os dados necessários para experimentos de classificação, foram desenvolvidas três versões de testes do *dataset* Gemed Onco - Protocolos (DGO-P), disponibilizados pela empresa Interprocess. Os dados são reais e não-estruturados. As informações foram extraídas de textos livres e foram usadas para treinar os modelos de classificação de texto.

As seguintes informações dos pacientes foram selecionadas: o diagnóstico de ICD e o protocolo de quimioterapia utilizado no tratamento do paciente. Foram elaboradas diversas versões ao longo do experimento, diversificando entre clínicas pequenas, médias e grandes, para medir a precisão dos dados. Os dados do *dataset* são classificados como evento clínico, os quais foram carregados e informados ao sistema EHR.

Um exemplo do *dataset* é ilustrado na Figura 23, com alguns textos (coluna "*SENTENCE*") e seu rótulo de anotação (coluna "*SENTENCE_CATEGORY*").

Um evento clínico contém uma ou mais questões, de acordo com sua forma personalizável. O carregamento por evento clínico gerou um único arquivo CSV (*Comma-Separated Values*)

Figura 23 – Exemplo de *dataset* por evento clínico

	A	B	C	D	E	F
1		SENTENCE_ID	SENTENCE	SETENCE_CATEGORY	ARGUMENT_2	ARGUMENT_2_CATEGORY
20	18	18	Exame físico normal	Ectoscopia / Pele e anexos	C50.9 - Mama	Diagnostico
21	19	19	Continua Anastrozol, Eligard e Zometa Solicito Ex lab e TCs de reestadiamento.	Conduta	C50.9 - Mama	Diagnostico
22	20	20	# Carcinoma ductal invasivo de mama D / Triplo negativo / Pré menopausa # 4x AC + 12 T até 23/09/2015 # RXT 25/11/2015 # Seguimento 5 meses > Recidiva óssea (CO : lesão em oitavo arco costal D) # 17 x Xeloda e Zometa # SVCS pelo PC - TEP - Stent com Dr. Luis Otávio em GYN	Histórico Oncológico	C50.9 - Mama	Diagnostico
23	21	21	Melhora dos sintomas de SVCS após retirada de PC e stent HB 13,8 L 3600 Plaq 120 mil Cr 0,9	Sintomas	C50.9 - Mama	Diagnostico
24	22	22	Boa tolerância ao Tratamento	Toxicidade	C50.9 - Mama	Diagnostico
25	23	23	Hematoma em tórax a E	Ectoscopia / Pele, anexos/ Mamas	C50.9 - Mama	Diagnostico
26	24	24	Impalpáveis	Linfonodos	C50.9 - Mama	Diagnostico
27	25	25	Ritmo cardíaco regular em dois tempos, BNF s/sopro	Cardiovascular	C50.9 - Mama	Diagnostico
28	26	26	MV + ARA	Respiratório	C50.9 - Mama	Diagnostico
29	27	27	Flácido, indolor, sem visceromegalias	Abdominal	C50.9 - Mama	Diagnostico
30	28	28	sem edema	Membros	C50.9 - Mama	Diagnostico
31	29	29	Ca de mama E T2N1M0 Triplo negativo EIIA R Osso em tratamento de primeira linha	Impressão	C50.9 - Mama	Diagnostico
32	30	30	Libero ciclo 18 Xeloda e Zometa	Conduta	C50.9 - Mama	Diagnostico
33	31	31	# Colectomia E por obstrução intestinal _ Adenocarcinoma G2, invasão perineural, T3N0M0 # 2x Roswell Park	Histórico Oncológico	C18.9 - Cólon	Diagnostico
34	32	32	Insônia	Sintomas	C18.9 - Cólon	Diagnostico
35	33	33	Boa tolerância a Qt	Toxicidade	C18.9 - Cólon	Diagnostico

Fonte: Elaborado pelo autor

com uma linha para cada evento clínico. Cada linha contém a anotação médica em texto livre (coluna "*SENTENCE*"), a categoria da anotação (coluna "*SETENCE_CATEGORY*") e os dados complementares (colunas "*ARGUMENT_2*" e "*ARGUMENT_2_CATEGORY*"). No caso dos dados analisados, se o evento clínico correspondente tiver mais de um diagnóstico ou protocolo, o texto e seu rótulo de anotação serão repetidos para cada um. Foram gerados três arquivos diferentes, um para cada clínica (pequena, média e grande), portanto, cada arquivo representa dados clínicos diferentes, permitindo o processamento individual.

Um processo de enriquecimento de *dataset* foi desenvolvido para alavancar o uso de dados estruturados disponíveis no sistema EHR. Nas formas personalizáveis, uma ou mais perguntas de texto livre ou tipos estruturados são possíveis. Os tipos estruturados contêm informações como diagnóstico de oncologia do ICD, protocolos de quimioterapia, medidas corporais do paciente, alergias, medicamentos prescritos e assim por diante. As seguintes informações estruturadas do paciente foram selecionadas para esse experimento: o diagnóstico de ICD do paciente e o protocolo de quimioterapia usado no tratamento do paciente.

A Tabela 14 apresenta uma comparação entre três bases de dados de clínicas diferentes. Essas informações foram extraídas do Banco de Dados do SGO. Esta tabela apresenta o número de notas médicas cadastradas por categoria profissional e entre parênteses a quantidade de profissionais que cadastraram essas notas. Nessa pesquisa, definimos as clínicas como Pequena, Média e Grande, diferenciando-as de acordo com seu porte.

Tabela 14: Comparativo de dados entre clínicas.

	Pequena	**Média**	**Grande**
Médicos	1,653 (2)	23,119 (18)	248,639 (210)
Enfermeiros	1,269 (1)	16,076 (5)	185,518 (43)
Farmacêuticos	386 (1)	1,133 (4)	36,072 (12)

Psicólogos	-	1,441 (4)	13,134 (17)
Nutricionistas	-	1,145 (2)	6,964 (9)
Fisioterapeutas	-	51 (1)	3,031 (5)
Dentistas	-	-	165 (4)
Profissionais de serviço social	-	-	84 (2)
Recepcionistas	-	75 (2)	-
Outros profissionais	-	-	166 (7)
Total de notas médicas	**3,308**	**43,040**	**493,773**
Total de pacientes distintos	**397**	**5,407**	**40,335**
Total de profissionais distintos	**4**	**36**	**309**

Elaborado pelo autor.

Pode ser observado no comparativo das bases de dados que médicos e enfermeiros são as duas categorias de profissionais que mais registram prontuários. Isso se deve à maior demanda por consultas do paciente e atendimento com esses profissionais. Quanto ao formato do *dataset*, foram avaliadas duas estruturas para exportar os dados, sendo as possibilidades as seguintes:

- Por evento clínico: São carregados os dados informados ao sistema EHR até a data do evento. Em outras palavras, os dados existentes quando o profissional de saúde registrava o evento clínico;

- Por paciente: Nesse caso, todos os eventos clínicos do paciente são reunidos em um texto único, portanto, serão carregados os dados do paciente atual.

O carregamento por evento clínico gera uma linha para cada evento clínico. Cada linha contém a anotação médica em texto livre, a categoria da anotação e os dados aprimorados. No caso dos dados, se o evento clínico correspondente tiver mais de um diagnóstico ou protocolo, o texto e seu rótulo de anotação são repetidos para cada um.

No caso do *dataset* por paciente, foi gerado um arquivo JSON (*JavaScript Object Notation*) para cada paciente, com todos os eventos clínicos associados ao mesmo. Cada arquivo JSON contém um texto grande com todos os eventos clínicos do paciente e uma lista de seus dados. O número de arquivos JSON varia de acordo com o número de pacientes que cada banco de dados tiver. O banco de dados da clínica pequena geraria em torno de 397 arquivos, a média 5.398 arquivos, e a grande, 39.326 arquivos JSON. O motivo para o uso de arquivos JSON em vez de arquivos do Excel é que o último possui um limite de 32.767 caracteres por célula, e o texto grande com todos os eventos clínicos do paciente frequentemente excede esse limite. A Figura 24 ilustra um exemplo de arquivo JSON utilizado.

Figura 24 – Um exemplo de um arquivo .JSON de paciente


```
{
   "MedicalNotes": [
      {
         "Note": "# Trombose MMII bilateral pré diagnóstico\n# PET_CT : Lesão de 6,8x4,5x5,0cm em rim direito SUV 12,2 sinais de
         invasão e trombose neoplásica de veia renal E e VCI SUV 15\nLinfonodo retroaórtico ( SUV 7,5) 2,4cm / \n# do 4100 Plaq
         251 mil Cr 1,4 UR 39 Glic 104 TGO 37 TGP 22 FA 100 Alb 3,7         DHL 329       Ca 8,9  TSH2,35\nVit D 25\nRNM abdome
         26/10/2017: Linfonodomegalia aortocaval 3,2cm ( doença estável) 21/02/18 doença estável. 13/09 redução para 2,7cm (
         15%)\n07/02/18 Tc de tórax: normal\nHipocorado +, hidratado, eupineico\nRCR 2 T BNF sem sopros\nMV + ARA\nFlácido,
         indolor, sem VMG\nedema de mmii +++\nCa de rim EIV\nLibero ciclo 10 Keytruda\nKeytruda\nC 10 d1: Foi preparado 200 mg
         de Keytruda em 100 ml de SF 0,9%. Foi entregue Eritromax 40000 UI para aplicação posterior.",
         "Arguments": [
            {
               "name": "Diagnostico",
               "value": "C64 - Rim"
            },
            {
               "name": "Protocolo",
               "value": "Eritropoetina"
            },
            {
               "name": "Protocolo",
               "value": "Gemzar"
            },
            {
               "name": "Protocolo",
               "value": "Keytruda"
            }
         ]
      }
   ]
}
```


Fonte: Elaborado pelo autor

5.1.2 *Dataset* Drug-Drug Interation

O *dataset* DDI (SEGURA BEDMAR; MARTÍNEZ; HERRERO ZAZO, 2013) é um *dataset* semanticamente anotado de documentos que descrevem as interações medicamentosas do banco de dados DrugBank e resumos do MedLine sobre o assunto. Esse *dataset* foi disponibilizado no DDIExtraction 2013 para execução de uma tarefa de desafio. O *dataset* do DDI consiste em 1.017 textos (784 textos do DrugBank e 233 resumos do MedLine) e foi anotado manualmente com um total de 18.491 substâncias farmacológicas e 5.021 interações medicamentosas. O *dataset* é distribuído em documentos XML (*Extensible Markup Language*) seguindo o formato unificado.

O *dataset* foi dividido a fim de construir os conjuntos de dados para o treinamento e avaliação dos diferentes sistemas participantes. Aproximadamente 77% dos documentos do *dataset* do DDI foram selecionados aleatoriamente para o conjunto de dados de treinamento e os restantes (142 textos do DrugBank e 91 resumos do MedLine) foram usados para o conjunto de dados de teste. O conjunto de dados de treinamento é o mesmo para ambas, pois contém entidades e anotações DDI.

O conjunto de dados de teste foi formado descartando documentos que continham anotações DDI. Os documentos restantes (ou seja, aqueles que contêm alguma interação) foram usados para criar o conjunto de dados de teste. Uma vez que as anotações de entidade não são removidas desses documentos, o conjunto de dados de teste também pode ser usado como conjunto de dados de treinamentos adicionais. A Figura 25 ilustra um exemplo do *dataset*.

Esse *dataset* foi selecionado para desenvolver previamente um experimento baseado na arquitetura *Transformer* para o REN e ER entre as mesmas, iniciando o experimento no domínio de medicamentos e, futuramente, promovendo a experiência necessária para a aplicação no domínio de textos oncológicos.

Figura 25 – Exemplo de documento do DDI

```
−<document id="DDI-DrugBank.d372">
 −<sentence id="DDI-DrugBank.d372.s0" text="Cytadren accelerates the metabolism of dexamethasone;">
    <entity id="DDI-DrugBank.d372.s0.e0" charOffset="0-7" type="brand" text="Cytadren"/>
    <entity id="DDI-DrugBank.d372.s0.e1" charOffset="39-51" type="drug" text="dexamethasone"/>
    <pair id="DDI-DrugBank.d372.s0.p0" e1="DDI-DrugBank.d372.s0.e0" e2="DDI-DrugBank.d372.s0.e1" ddi="true" type="mechanism"/>
  </sentence>
 −<sentence id="DDI-DrugBank.d372.s1" text="therefore, if glucocorticoid replacement is needed, hydrocortisone should be prescribed.">
    <entity id="DDI-DrugBank.d372.s1.e0" charOffset="14-27" type="group" text="glucocorticoid"/>
    <entity id="DDI-DrugBank.d372.s1.e1" charOffset="52-65" type="drug" text="hydrocortisone"/>
    <pair id="DDI-DrugBank.d372.s1.p0" e1="DDI-DrugBank.d372.s1.e0" e2="DDI-DrugBank.d372.s1.e1" ddi="false"/>
  </sentence>
 −<sentence id="DDI-DrugBank.d372.s2" text="Aminoglutethimide diminishes the effect of coumarin and warfarin.">
    <entity id="DDI-DrugBank.d372.s2.e0" charOffset="0-16" type="drug" text="Aminoglutethimide"/>
    <entity id="DDI-DrugBank.d372.s2.e1" charOffset="43-50" type="group" text="coumarin"/>
    <entity id="DDI-DrugBank.d372.s2.e2" charOffset="56-63" type="drug" text="warfarin"/>
    <pair id="DDI-DrugBank.d372.s2.p0" e1="DDI-DrugBank.d372.s2.e0" e2="DDI-DrugBank.d372.s2.e1" ddi="true" type="effect"/>
    <pair id="DDI-DrugBank.d372.s2.p1" e1="DDI-DrugBank.d372.s2.e0" e2="DDI-DrugBank.d372.s2.e2" ddi="true" type="effect"/>
    <pair id="DDI-DrugBank.d372.s2.p2" e1="DDI-DrugBank.d372.s2.e1" e2="DDI-DrugBank.d372.s2.e2" ddi="false"/>
  </sentence>
</document>
```

Fonte: Elaborado pelo autor

5.1.3 *Dataset* Gemed Onco - Exames

Para a construção do DGO-E, foram exportados 1.022 documentos de evolução de uma clínica oncológica. Esses documentos contêm dados reais prescritos por profissionais de saúde e foram exportados no formato JSON. Antes de utilizar os documentos, eles passaram por um processo de anonimização descrito na seção 4.4, no qual foram removidos quaisquer nomes ou identificações de profissionais de saúde ou pacientes. Um exemplo de documento de evolução é ilustrado na figura 26.

Figura 26 – Exemplo de documento de evolução em JSON

```
[
  {
    "document": "ca da mama direita\nha 6 meses percebeu nodulo na mam direita\nrealizou eco mama em setembro de
    2019 birads 3  nodulo 192x168x9mm qsm mama direitalinfonodomegalias axila direita com 20x11mm\neco mamas 29/04/
    2020 nodulos mama direita qsm com 29x22x14 mm a 5cm do mamilo e outro adjacente medial com 12x10x10mm outro
    nodulo adjacente no qsm da mama direita com 5x4x4mm linfonodomegalia axilar direita com 22x11mm 18x8
    17x13mm\nbx percutanea em 04/05/2020 carcinoma ductal invasor no nodulo maior ca mucinoso no nodulo menor
    carcinoma metastatico para linfonodo axilar\nhigida\nnodulos palpaveis no qsm da mama direita\nmamas
    pequenas\naxila direita com adenomegalias palpaveis\nsolicito cintilografia ossea  rx torax eco
    abdominal\nencaminho para o sus para qt neo adj\nparar o uso do anticoncepcional\nesta em qt neo adj no
    heg\nqx de dor na mandibula e maxila apos qt\nao exame sem sinais de infeccao na gengiva\ncd prescrevo miosan
    10 mg rivotril gotas paco"
  }
]
```

Fonte: Elaborado pelo autor

Após a exportação, foram realizadas reuniões com especialistas em oncologia para definir o foco da construção do conjunto de dados. Decidiu-se realizar um experimento com dados de exames, pois é possível identificar esses dados no texto e são importantes para um protótipo da empresa. Com base nisso, as seguintes entidades de interesse foram definidas como foco da anotação de dados:

- Exame: Anota o nome do exame. Exemplo: PSA;
- Resultado: Anota o valor do resultado de um exame. Exemplos: 193mil, 1547;

- Data: Anota uma data que identifique a data de realização de um exame. Exemplos: 18/07/2022, 07/2022;
- Membro: Anota um membro do corpo humano, buscando identificar o membro sobre o qual o exame foi realizado. Exemplos: Crânio, lombar;
- Tempo: Anota o tempo de uma ação, buscando identificar a quanto tempo um exame foi realizado. Exemplos: Há 1 mês, ano passado.

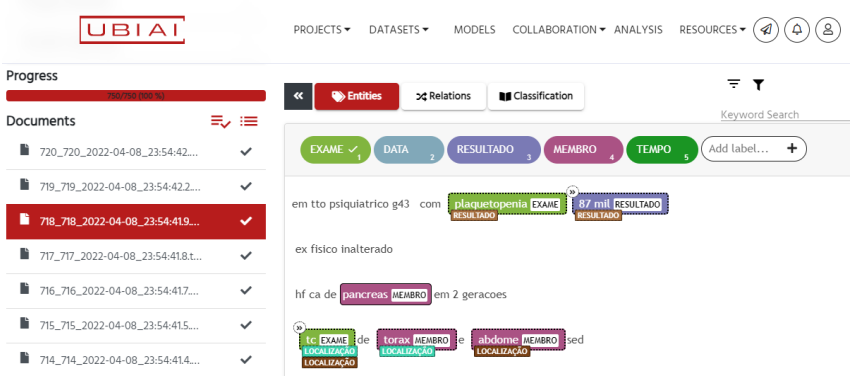
Quanto às relações de interesse para a anotação, as seguintes foram definidas:

- Resultado: Relaciona o resultado de um exame ao exame;
- Quando: Relaciona a data/tempo de um exame ao exame;
- Localização: Relaciona um membro ao exame.

Após definir os parâmetros necessários, deu-se início ao processo de construção das anotações para o conjunto de dados. Para isso, foram selecionados três alunos bolsistas que estão no final do curso de medicina da Unisinos e possuem conhecimentos na área, além de uma enfermeira e uma biomédica, ambas colaboradoras da Empresa Interprocess.

Foi utilizado o software UBIAI Tools[1] para realizar as anotações. Essa ferramenta possibilita a colaboração na realização das anotações e possui uma interface intuitiva. Com o UBIAI Tools, é possível realizar anotações manuais de entidades, relações e classificações, contando com recursos como Dicionário de Entidades, Correspondência Baseada em Regras e Dicionário de Relações para anotações automáticas. Para este estudo, optou-se por realizar anotações manuais, visando alcançar um alto nível de qualidade no conjunto de dados. A figura 27 apresenta uma imagem do software utilizado.

Figura 27 – Software UBIAI com um documento anotado



Fonte: Elaborado pelo autor

O conjunto dos 1.022 documentos anotados passou por um processo de revisão e validação realizado por especialistas na área de oncologia. Durante o processo de revisão e validação,

[1] https://ubiai.tools

os especialistas garantem a consistência e a precisão das anotações, verificando se as entidades e relações foram corretamente identificadas e marcadas nos documentos. Essa etapa é fundamental para garantir a qualidade e confiabilidade do conjunto de dados anotado, fornecendo informações valiosas para análises posteriores. Os resultados das anotações das entidades podem ser visualizados na tabela 15.

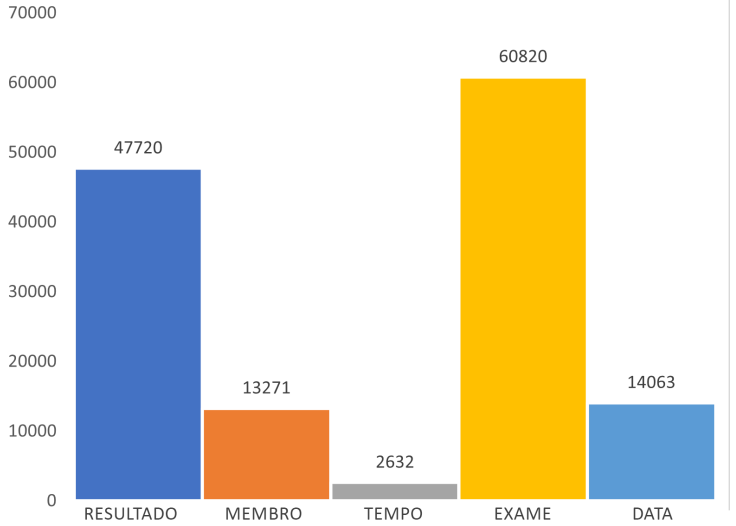
Tabela 15 – Entidades do DGO-E

Entidade	**Quantidade**
Resultado	48.499
Membro	14.110
Tempo	2.823
Exame	61.861
Data	15.919

Fonte: Elaborado pelo autor.

O *dataset* é composto por um número significativo de entidades anotadas, com destaque para as categorias ”Exame” e ”Resultado”, que possuem uma quantidade expressiva de anotações (61.861 e 91.355, respectivamente). Esses números mostram um ponto positivo, pois indica que o *dataset* abrange uma ampla variedade de informações relacionadas nos EHRs. Além disso, as entidades ”Membro” e ”Data” apresentam quantidades menores de anotações (14.110 e 15.919). Essas entidades são relevantes para o contexto médico e podem fornecer informações importantes sobre partes do corpo afetadas e datas relacionadas aos exames. O gráfico da figura 28 apresenta uma representação visual desses resultados.

Figura 28 – Comparativo de entidades de exames do DGO-E



Fonte: Elaborado pelo autor

No contexto da ER em nas evoluções oncológicas, o objetivo foi identificar as relações relevantes entre as entidades anotadas, como por exemplo, a relação entre um ”Exame” e seu

”Resultado”, ou a relação entre uma ”Data” e um ”Exame” específico. Os resultados das relações anotadas são apresentados na Tabela 16.

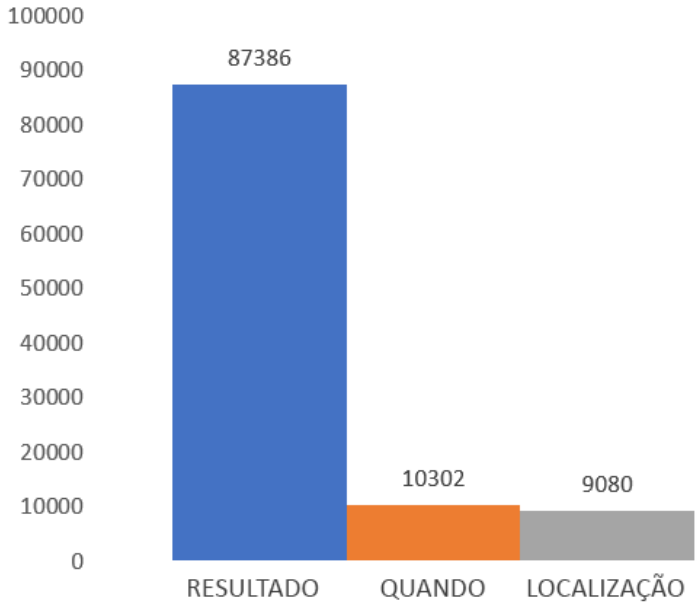
Tabela 16 – Relações do DGO-E

Relação	**Quantidade**
Resultado	91.355
Quando	9.517
Localização	10.844

Fonte: Elaborado pelo autor.

Quanto às relações anotadas no *dataset*, observa-se que a relação ”Resultado” possui uma quantidade expressiva de anotações (91.355), enquanto as relações ”Quando” e ”Localização” têm quantidades menores (9.517 e 10.844, respectivamente). Essas informações serão importantes para auxiliar na interpretação e análise dos resultados dos exames. Um comparativo das relações é ilustrado no gráfico da figura 29.

Figura 29 – Comparativo de relações de exames do DGO-E



Fonte: Elaborado pelo autor

Os resultados obtidos nas tabelas e gráficos destacam a distribuição das entidades e relações no conjunto de dados. Essas informações são essenciais para compreender a frequência de ocorrência das entidades e a relação entre elas. Por fim, o *dataset* é exportado no formato IOB (TOLEU; TOLEGEN; MAKAZHANOV, 2017). O formato IOB é amplamente utilizada na rotulação de dados para REN. Ele fornece uma estrutura consistente e padronizada para marcar as fronteiras das entidades em um texto. No formato IOB, cada palavra ou *token* em um texto é marcado como uma das três classes: *Beginning* (B), *Inside* (I) ou *Outside* (O). O marcador ”B” é atribuído à primeira palavra de uma entidade, o marcador ”I” é atribuído às

palavras subsequentes que pertencem à mesma entidade, e o marcador ”O” é atribuído a todas as palavras que estão fora de qualquer entidade. Esse formato permite uma análise precisa do texto. Ele evita a sobreposição ou a falta de rotulação de palavras relacionadas a uma entidade, garantindo que todas as partes relevantes sejam identificadas corretamente.

5.1.4 *Dataset* Gemed Onco - Características de Diagnóstico

Para a construção do DGO-CD, foi realizada uma parceria com a empresa Comsentimento, que permitiu que profissionais especializados em anotação e saúde apoiassem a realização das tarefas de rotulação. Após as reuniões de análise, decidiu-se direcionar os esforços para a criação de um conjunto de dados focado no diagnóstico de Câncer de Mama, visando fornecer informações valiosas nessa área crucial da saúde. As entidades que foram abordadas nesse conjunto de dados incluem:

- Característica do Diagnóstico: Anota atributos e informações relevantes sobre um determinado caso de câncer ou condição oncológica.
  Exemplo: Ca de cólon Ell **alto risco**.
- Órgão: Anota o órgão afetado pelo câncer.
  Exemplo: Paciente notou nódulo em **mama** Esquerda.
- Local do Órgão: Anota a localização precisa do órgão afetado.
  Exemplo: Identificou um nódulo em mama **Direita**.
- Estadiamento: Anota a classificação do estadiamento do câncer.
  Exemplo: Ca de ovário **EIV** resistente a platina.
- Tempo: Anota o tempo decorrido desde a detecção ou início dos sintoma.
  Exemplo: Paciente notou **há 6 meses** nódulo na mama E com crescimento moderado.

Para construir este *dataset*, foi decidido não anotar as relações, uma vez que essa não era a especialidade dos profissionais de saúde envolvidos. Para isso, foram exportados 600 documentos de evoluções oncológicas no formato JSON de uma clínica oncológica, já filtrados com diagnóstico de Câncer de Mama. Todos os documentos passaram por um processo de anonimização, detalhado na seção 4.4, a fim de garantir a privacidade dos pacientes e a conformidade com as diretrizes éticas.

Para realizar as anotações, foi utilizado o software MAE Annotation Tool[2]. O MAE (*Multi-document Annotation Environment*) é uma ferramenta de anotação de LN que oferece recursos abrangentes para a marcação de entidades em textos. O MAE possui recursos avançados, como a capacidade de trabalhar com vários documentos simultaneamente, facilitando o processamento de grandes conjuntos de dados.

[2]http://keighrim.github.io/mae-annotation/

Após a conclusão do processo de anotação, os 600 documentos anotados foram submetidos a uma revisão e validação realizadas por especialistas na área de oncologia. Essa etapa é essencial para garantir a qualidade e a consistência das anotações, assegurando que as informações relevantes tenham sido corretamente identificadas e rotuladas nos documentos. por fim, o *dataset* foi exportado no formato IOB. A tabela 17 apresenta a quantidade de entidades anotadas.

Tabela 17: Entidades do DGO-CD

Entidade	**Quantidade**
Característica do Diagnóstico	1096
Órgão	855
Local do Órgão	342
Estadiamento	1148
Tempo	116

Elaborado pelo autor.

Neste *dataset*, nota-se que há uma quantidade considerável de entidades anotadas, com destaque para a entidade ”Estadiamento” com 1.148 ocorrências. Essa entidade é de extrema importância no diagnóstico do câncer de mama, pois fornece informações valiosas sobre o estágio da doença e sua progressão. Além disso, outras entidades relevantes foram anotadas, como ”Característica do Diagnóstico” com 1.096 ocorrências e ”Órgão” com 855 ocorrências. Essas entidades fornecem informações sobre as características específicas da doença e o órgão afetado.

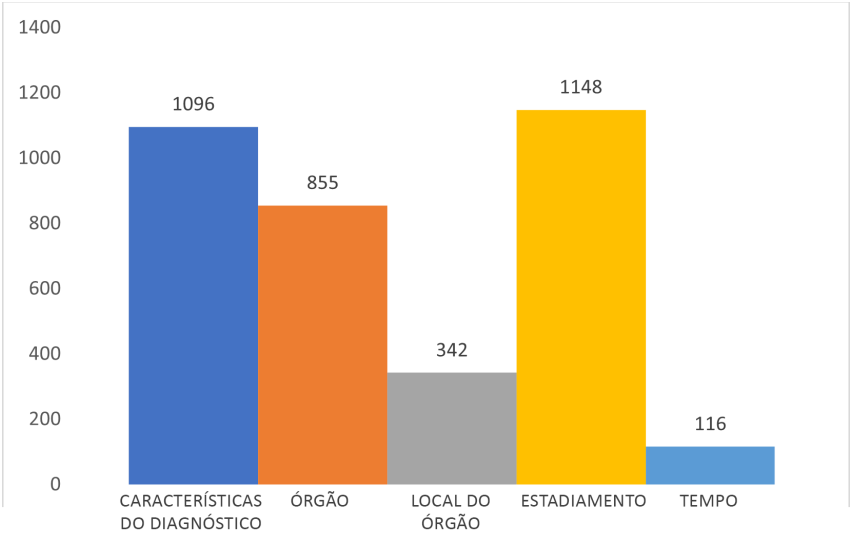
Por outro lado, a entidade ”Local do Órgão” apresentou um número menor de ocorrências, com 342 registros. Essa quantidade pode sugerir que a informação específica sobre a localização do órgão afetado pelo câncer de mama pode ser menos enfatizada nos textos médicos anotados. No entanto, é necessário considerar que a importância dessa entidade pode variar dependendo do contexto médico específico.

A entidade ”Tempo” registrou 116 ocorrências, fornecendo informações sobre o tempo relacionado ao diagnóstico do câncer de mama. Embora seja uma quantidade menor em comparação com outras entidades, o tempo é um fator relevante na análise do câncer de mama e pode fornecer informações importantes para o tratamento e prognóstico dos pacientes. Um gráfico comparativo das entidades é ilustrado na figura 30.

5.2 Experimento de classificação de textos

Foi desenvolvido um experimento de classificação utilizando arquiteturas de ML e DL. Informações estruturadas e de texto livre foram usadas para treinar os modelos de classificadores de texto e, posteriormente, para sugerir a categoria de novos textos usando os modelos treina-

Figura 30 – Comparativo de entidades do DGO-CD



Fonte: Elaborado pelo autor

dos. Neste experimento, foi utilizado o *dataset* DGP-P. A seguir são descritos os detalhes das etapas desenvolvidas.

5.2.1 Pré-processamento do *dataset*

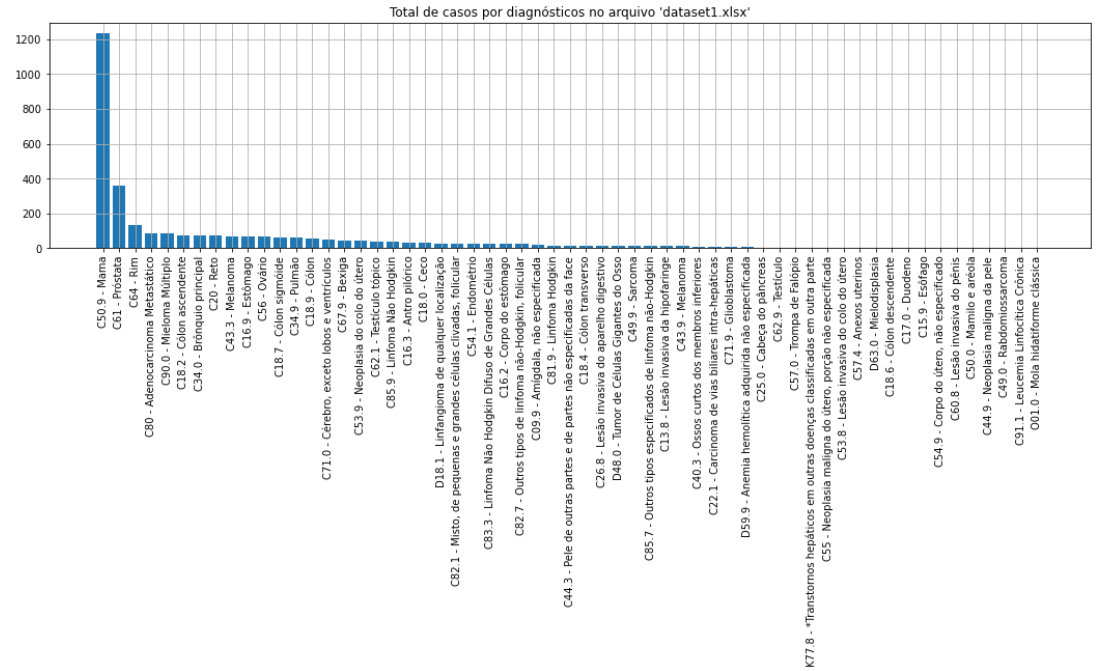
Após o processo de criação do DGP-P, foi necessário realizar um pré-processamento para adequar o texto. O mesmo pré-processamento foi realizado para os DGP-P por evento clínico e por paciente. Antes do pré-processamento do texto, um histograma de diagnóstico foi gerado (Figura 31). Foi possível observar que um pequeno grupo de diagnósticos concentra as ocorrências mais frequentes. Portanto, os diagnósticos com menos de 50 ocorrências foram unidos em um único grupo chamado ”Outros”.

Além disso, para avaliar o desempenho da RN de acordo com a esparsidade do conjunto de dados, foi criada uma versão do conjunto de dados com os 12 diagnósticos com mais ocorrência, ilustrado na Figura 32.

As seguintes tarefas foram executadas. Inicialmente a tokenização para a divisão do texto em *tokens* que correspondem a palavras. Em seguida a Filtragem de *stop-words* para a remoção das palavras mais comuns no idioma brasileiro, pontuação e caracteres especiais. Por fim, foi implementada a caixa dobrável para conversão de todas as palavras para minúsculas.

Além disso, uma análise manual adicional foi realizada. O texto foi avaliado para entender como ele poderia ser transformado para melhorar os algoritmos da RN. Foram removidos textos médicos repetidos em várias categorias de anotação. Foi realizado um experimento adicional para avaliar como essa etapa alavancou os resultados dos classificadores. Uma melhoria significativa foi alcançada com a aplicação desta etapa.

Figura 31 – Exemplo de histograma de diagnóstico com o total de ocorrências de diagnóstico por código ICD



Fonte: Elaborado pelo autor

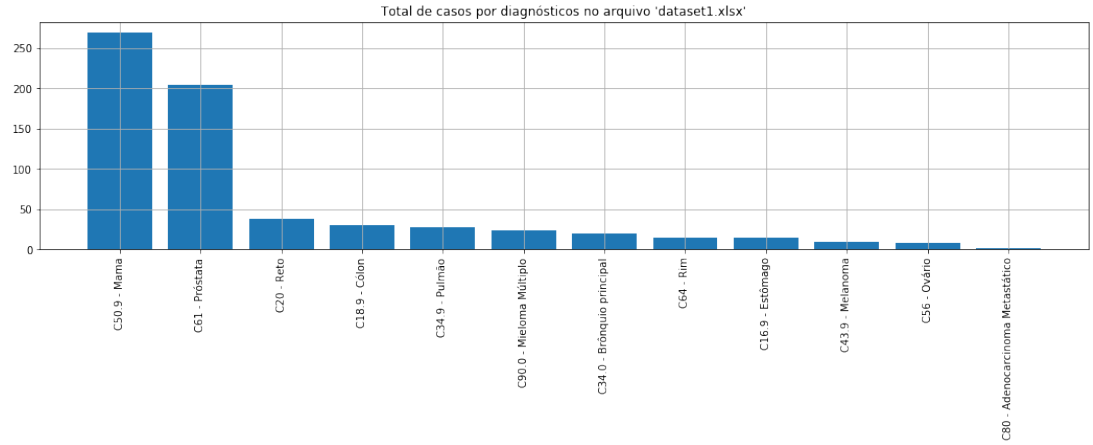
5.2.2 Extração de aspectos de interesse

O texto das anotações médicas deve ser transformado em uma estrutura que possa ser usada pelos classificadores. Por esse motivo, foi utilizado o método *Bag-of-Words* (BoW), que é uma representação que transforma um texto em vetores de comprimento fixo, contando quantas vezes cada palavra aparece. Essa representação é útil para ser usada pelos algoritmos de classificação. Foi utilizado o BoW para anotações médicas para extrair os elementos de interesse a serem usados no ML e no treinamento de aprendizagem profunda. Nos *datasets* por evento clínico, o BoW foi gerado para cada nota médica. Da mesma forma, nos *datasets* por paciente, foi gerado para cada paciente com todas as suas anotações médicas.

Antes de criar as notas médicas do BoW, o texto foi adequado no pré-processamento. Essa etapa de pré-processamento visa reduzir o número de palavras inúteis, caracteres especiais e pontuações, o que não faria diferença no treinamento dos modelos de classificadores. Também ajudou a reduzir o esforço computacional para criar o BoW.

O BoW aplicado às notas médicas resultou em uma representação esparsa, ou seja, a sequência vetorial de números que representam cada palavra continha muitos zeros. Foi aplicada a técnica de Análise de Componentes Principais (*Principal Component Analysis* - PCA) para reduzir a esparsidade dos dados. A técnica PCA converte um conjunto de observações de recursos possivelmente correlacionados em um conjunto de valores de recursos linearmente não correlacionados. Foi utilizado o PCA com 500 recursos.

Figura 32 – Diagnósticos com mais ocorrências.



Fonte: Elaborado pelo autor

5.2.3 Arquiteturas de *Machine Learning* e *Deep Learning*

Para a tarefa de classificação de texto, aplicou-se métodos de ML e DP, comparando os resultados. Vários algoritmos de classificação de ML foram aplicados para avaliar qual teria o melhor desempenho. Além disso, um algoritmo LSTM de DL também foi aplicado para comparar o ML tradicional e uma RNR de aprendizagem profunda. Os seguintes algoritmos de ML foram avaliados: MLP; Regressão Logística; Classificador de Árvore de Decisão; Classificador *Random Forest*; Classificador *Extra Tress*; *K-Nearest Neighbors* (KNN). Além disso, também foi realizado um experimento de aprendizagem profunda com redes LSTM.

Os conjuntos de dados foram divididos em dois grupos, para treinar e testar as RNs: um com 80% dos dados para treinar os modelos e o outro com 20% dos dados para testar os modelos. Os dados foram embaralhados para manter a proporção das categorias e, em seguida, foram divididos conforme mencionado acima. Os algoritmos de ML foram implementados usando o *scikit-learn*, e o LSTM de aprendizagem profunda foi realizado usando a biblioteca *keras*, ambos foram implementados em Python. No primeiro conjunto de experimentos, foram realizados sete testes, com os seguintes detalhes da arquitetura: Um MLP com uma camada oculta com 500 neurônios; um MLP com duas camadas ocultas com 800 e 500 neurônios; um classificador de regressão logística; uma árvore de decisão com no máximo vinte níveis e três amostras por folha; uma *Random Forest* com um máximo de vinte níveis e três amostras por folha; árvores extras com no máximo vinte níveis e três amostras por folha; um classificador KNN com um K. unitário. As quantidades de neurônios e camadas ocultas dos experimentos de MLP foram escolhidas para testes de acordo com os resultados positivos de trabalhos relacionados.

O segundo conjunto de experimentos foi realizado, desta vez com os 12 diagnósticos com mais ocorrências no conjunto de dados. Para esse experimento, foi selecionado o algoritmo de ML que melhor funcionava para comparar com um LSTM recorrente de aprendizagem, com a seguinte arquitetura: um MLP com duas camadas ocultas com 800 e 500 neurônios; um LSTM

com a biblioteca *keras* construída sobre o *Tensorflow* em uma implementação Python, na qual a parametrização usada era composta pelo tamanho do lote 128, taxa de abandono de 0,2, divisão de validação de 0,2, a medida de perda foi a entropia categórica. Além disso, para evitar o excesso de ajustes, foi utilizado o *EarlyStopping*. Nesses experimentos, como o foco principal foi dedicado à aplicação geral, foram avaliados usando tanto métricas padrão como precisão *F1 macro* e *Weighted scores*.

5.2.4 Avaliações com diferentes conjuntos de dados

Neste experimento foram utilizados os dois conjuntos de dados (por evento clínico e por paciente). Na etapa seguinte foi realizado um novo experimento envolvendo o conjunto de dados por paciente e os algoritmos MLP e LSTM. O conjunto de dados por paciente foi escolhido na segunda etapa, porque todas as anotações clínicas do paciente foram concatenadas em um único registro, com desempenho melhor considerando a capacidade do LSTM de processar sequências inteiras de dados.

Portanto, foram realizados dois experimentos principais. O primeiro dedicado à ML, no qual vários classificadores de ML foram experimentados e seu desempenho comparado. O segundo dedicados à Aprendizagem profunda, tendo sido realizado um experimento com uma RNR de aprendizagem profunda. Estes experimentos são descritos a seguir.

Nos Experimentos de ML, o DGP-P por evento clínico da clínica pequena foi usado para realizar as classificações. Este *dataset* contém 3.308 notas clínicas e 397 pacientes distintos. O conjunto de dados pré-processados foram usados primeiro com os classificadores de ML. Para realizar os experimentos, o conjunto de dados foi dividido aleatoriamente em duas partes: 80% para treinamento e 20% para teste. Um método foi usado para gerar essas duas partes e, para cada vez que era realizado, um conjunto diferente de dados de treinamento e teste era criado.

Quanto à precisão média, o *Macro F1* score e o *Weighted F1 score* de cada classificador, estes são apresentados na Tabela 18. Esses experimentos foram realizados para avaliar qual classificador de ML apresentaria o melhor desempenho. De acordo com a Tabela 18, o classificador MLP 2 alcançou a melhor precisão, os *scores Macro F1* e *F1 Weighted*.

Tabela 18: Resultado dos melhores experimentos

Método	**Precisão**	**Macro F1**	**F1 ponderada**
MLP 1 (1 camada oculta, 500 neurônios)	84.89%	84.21%	84.99%
MLP 2 (2 camadas ocultas, 800 e 500 neurônios)	87.62%	87.44%	87.70%
Regressão Logística	84.89%	82.75%	84.75%
Árvore de Decisão	71.86%	63.95%	71.98%
Random forest	80.23%	76.09%	79.53%
Extra trees	78.46%	76.71%	78.03%

KNN	85.05%	83.93%	85.20%

Elaborado pelo autor.

Um experimento adicional foi realizado para avaliar como a estrutura do conjunto de dados e a etapa de pré-processamento alavancaram o desempenho dos classificadores. Esse experimento utilizou o mesmo conjunto de dados da clínica identificada como pequena. As versões pré-processadas e brutas do conjunto de dados por evento clínico, além do conjunto de dados por paciente pré-processado, foram usadas com o melhor classificador de desempenho. De acordo com a Tabela 18, o classificador MLP 2 teve o melhor desempenho.

A Tabela 19 apresenta a precisão média do classificador MLP 2 com as versões pré-processadas e brutas do conjunto de dados por evento clínico, além do conjunto de dados por paciente pré-processado.

Tabela 19: Comparação do desempenho dos *datasets*

Dataset	**Precisão média**
Dataset por evento clínico bruto	26.1%
Dataset por evento clínico pré-processado	86.7%
Dataset por paciente pré-processado	93.9%

Elaborado pelo autor.

No conjunto de experimentos de Aprendizagem Profunda, os seguintes classificadores de ML e aprendizagem profunda foram testados: O classificador de ML com melhor desempenho no experimento anterior (o MLP 2); Uma RNR de aprendizagem profunda em LSTM.

O *dataset* por paciente da clínica pequena da versão pré-processada foi utilizado para realizar o experimento com os classificadores. O *dataset* por paciente foi escolhido porque todas as anotações clínicas do paciente foram concatenadas em um único registro, com melhor desempenho, considerando a capacidade do LSTM de processar sequências inteiras de dados. A Tabela 20 apresenta a precisão média, o Macro F1 score e o Weighted F1 score dos classificadores MLP 2 (2 camadas ocultas, 800 e 500 neurônios) e LSTM.

Tabela 20: Desempenho do MLP 2 e LSTM.

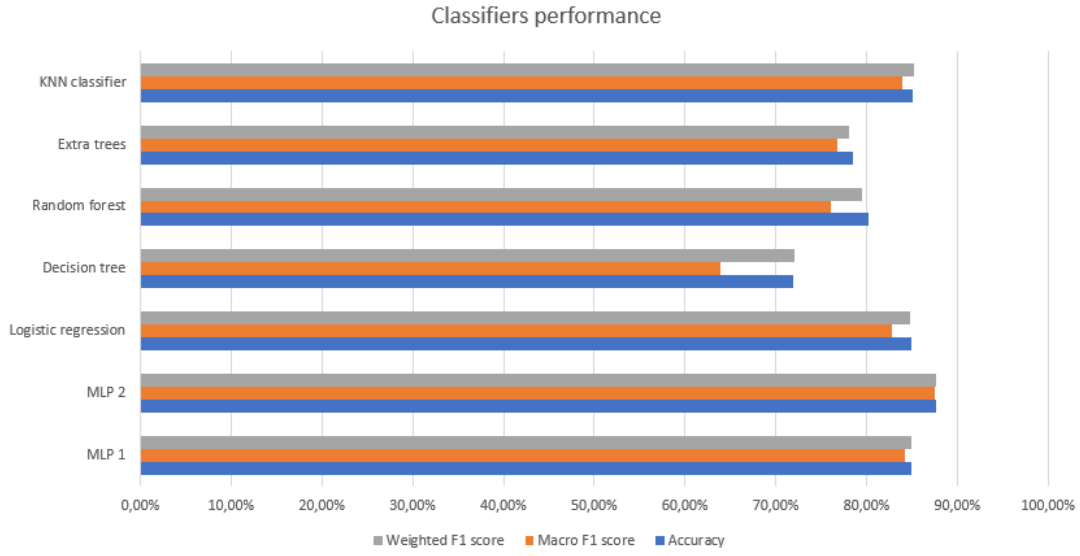
Método	**Precisão**	**Macro F1 score**	**Weighted F1 score**
MLP 2	93.90%	93.61%	93.99%
LSTM	84.81%	84.57%	84.93%

Elaborado pelo autor.

5.2.5 Avaliação dos resultados

Vários experimentos foram realizados para entender o comportamento dos classificadores selecionados de ML e aprendizagem profunda com os *datasets* criados e pré-processados. Primeiro, um conjunto de experimentos com sete classificadores de ML foram realizados com o *dataset* por evento clínico. Considerando a precisão média, os Macro F1 e F1 Weighted scores, o classificador que melhor se apresentou foi o MLP 2, como mostra a Figura 33.

Figura 33 – Gráfico de desempenho dos classificadores de ML.

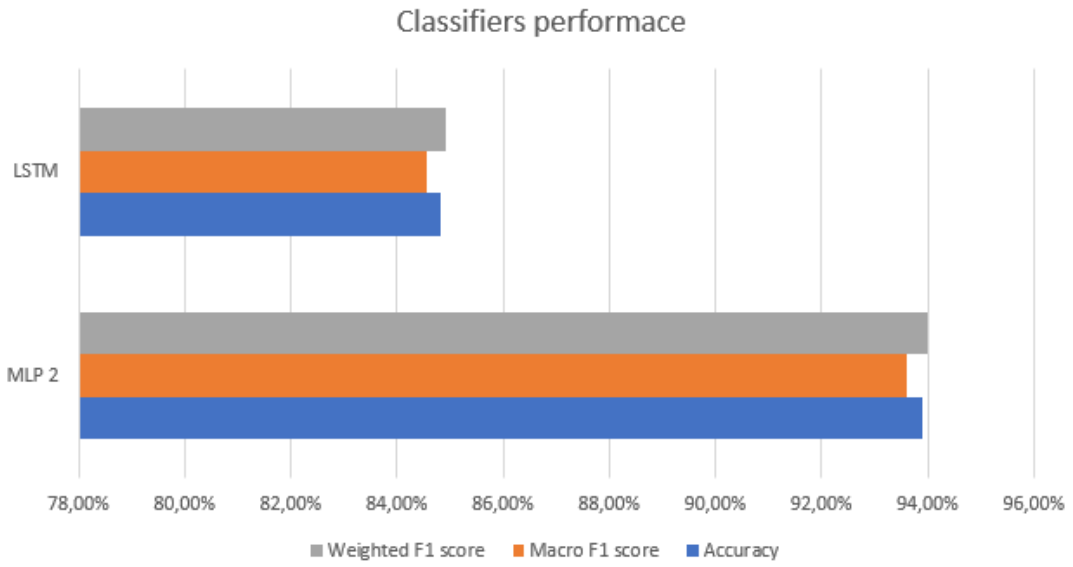


Fonte: Elaborado pelo autor

A preparação do DGP-P a ser usado nos classificadores de ML e DL foi um passo importante. Foi realizada um experimento com o classificador MLP 2, as versões pré-processadas e brutas do conjunto de dados por evento clínico e o conjunto de dados por paciente pré-processado. Ele mostra uma melhoria significativa do desempenho do classificador MLP 2 com o pré-processamento do conjunto de dados, nos conjuntos de dados por evento clínico e por paciente. Além disso, o *dataset* por paciente teve um desempenho melhor que o DGP-P por eventos clínicos. Por esse motivo, os próximos experimentos usaram o DGP-P por paciente.

Após a avaliação do classificador de ML de melhor resultado, o classificador MLP 2 foi selecionado. Foi realizado um novo experimento, para comparar o classificador MLP 2 com uma LSTM, com o *dataset* por paciente pré-processado. Como visto na Figura 34, o MLP 2 teve um desempenho melhor do que o classificador LSTM, mesmo que este último seja uma RN mais recente. A razão para esse resultado demonstrou que o experimento utilizou o menor *dataset* por paciente e os algoritmos de aprendizagem profunda tiveram melhor desempenho em grandes conjuntos de dados.

Figura 34 – A tabela de desempenho do melhor classificador de ML e classificador de aprendizagem profunda.



Fonte: Elaborado pelo autor

5.3 Reconhecimento de entidades e relações com o *dataset* DDI

Nesse experimento de REN, o modelo BERT foi utilizado para identificar as entidades presentes em um determinado texto, bem como seus tipos. Cada texto é processado pelo modelo BERT uma única vez. Devido ao processo de anotação de dados do DGO estar em andamento, não foi possível utilizar esse conjunto de dados para os experimentos. Portanto, optou-se por utilizar o conjunto de dados DDI (SEGURA BEDMAR; MARTÍNEZ; HERRERO ZAZO, 2013) para realizar este experimento. Os detalhes da arquitetura e dos experimentos são descritos a seguir:

5.3.1 Arquitetura do experimento de reconhecimento de entidades

Para construir a arquitetura, utilizou-se como base a documentação presente no Huggingface[3], tendo sido usado o modelo BertForTokenClassification, proposto em Devlin et al. (2018a). Esse modelo possui um amplo suporte em bibliotecas, incluindo um *token*izer Python chamado ”lento” e um *token*izer ”rápido”, além de ser compatível com Jax (via Flax), PyTorch e TensorFlow. No presente experimento, o modelo BERT foi treinado para realizar o REN e extrair características que representam o texto. Para isso, utilizou-se o conjunto de dados DDI.

O Quadro 2 apresenta um exemplo de texto. Alguns exemplos em relação do DDI são destacados a seguir.

Entidades:

[3] www.huggingface.co/transformers

Quadro 2 – Exemplo de texto medicamentoso

”*Prolonged recovery time may occur if barbiturates and/or narcotics are used concurrently with ketamine.*”

- Nome: “*barbiturates*”; Tipo: “grupo”;
- Nome: “*narcotics*”; Tipo: “grupo”;
- Nome: “*ketamine*”; Tipo: “droga”

Relações:

- “Efeito” (“*barbiturates*”, “*ketamine*”);
- “Efeito” (“*narcotics*”, “*ketamine*”);
- “Nenhuma” (“*barbiturates*”, “*narcotics*”).

O modelo utilizado, tanto para etapa de reconhecimento das entidades quanto da ER, foi o BERT. Utilizou-se uma rede neural profunda com diversas versões pré-treinadas disponíveis, que são ajustadas para as tarefas específicas com uso dos *datasets* anotados.

Através do uso do BERT, foi possível identificar as entidades de acordo ao classificar de cada palavra (“comum”, significa que é um termo comum, ou seja, não é uma entidade). A Tabela 21 ilustra alguns exemplos de classificação.

Tabela 21: Exemplos de Classificação

Termos da Entrada	**Classificação das Palavras**
Prolonged	comum
recovery	comum
time	comum
may	comum
occur	comum
if	comum
barbiturates	grupo
and/or	comum
narcotics	grupo
are	comum
used	comum
concurrently	comum
with	comum

Continua na próxima página

Tabela 21 – continua da página anterior

Termos da Entrada	Classificação das Palavras
ketamine	droga
.	comum

Elaborado pelo autor.

Diferentes modelos iniciais foram incluídos no experimento. A proposta foi utilizar o melhor resultado do experimento de reconhecimento das entidades para classificação das relações com a BiLSTM (descrito na seção 5.3.3). Os modelos BERT utilizados para os testes desse experimento são: "BERT-base-multilingual-cased" (SOUZA et al., 2021b) (SOUZA et al., 2021b), "BioBERT-PT" (SOUZA et al., 2021b) (JI; WEI; XU, 2020) e "BERTimbau" (LOPES; CORREA; FREITAS, 2021) (SOUZA; NOGUEIRA; LOTUFO, 2020), descritos a seguir:

- BERT multilingual ("bert-base-multilingual-cased"): pré-treinado na Wikipedia com textos em mais de 100 línguas, disponibilizado pela Google. Escolhido por ser o padrão para textos que não são em inglês nem em chinês;
- BERTimbau ("neuralmind/bert-base-portuguese-cased"): pré-treinado em textos em português, disponibilizado pela NeuralMind. Escolhido por geralmente apresentar resultados um pouco melhores que o BERT multilingual para textos em português;
- BioBERTpt[4]: É um BERT multilingual adaptado para textos médicos em português, disponibilizado por Schneider et al. (2020). Escolhido por ter sido adaptado a um domínio próximo do que está sendo trabalhado.

O modelo BERT possui 12 ou 24 camadas, dependendo da versão, e cada camada é composta por 12 ou 16 cabeças de atenção (componentes do BERT que ajudam a processar informações e identificar características importantes em um texto), que projetam as representações em um subespaço vetorial e identificam características distintas. O resultado de cada cabeça de atenção é concatenado e passado para a próxima camada por meio de mecanismos de auto-atenção, nos quais o modelo determina quais características são mais relevantes para serem propagadas.

Em resumo, o modelo possui as seguintes características:

- Função de ativação GELU em vez da função de ativação ReLU original da *Transformer*;
- Mecanismo de auto-atenção, que permite filtrar, combinar e transformar as características aprendidas pelo modelo em cada camada;

[4] pucpr/biobertpt-bio

- *embeddings* posicionais, que permitem a leitura do texto inteiro de uma só vez, em oposição às redes neurais recorrentes que precisam processar palavra por palavra, permitindo uma paralelização muito maior;
- Quebra do texto em sub-palavras (*tokenização WordPiece*), permitindo um vocabulário menor e mais abrangente do que os vocabulários baseados em palavras;
- Otimizador AdamW (*Adaptive Moment Estimation*).

Neste experimento, utilizou-se o modelo BERT sem nenhuma alteração, treinado para ler os textos do DDI e identificar as entidades presentes nele. O BERT possui uma limitação de entrada máxima de 512 sub-palavras. Como o DDI contém textos mais longos do que isso, processamos o texto em várias etapas com sobreposição e, em seguida, concatenamos as classificações. Embora existam versões como o *Transformer*XL (DAI et al., 2019) projetadas para lidar com textos mais longos, não foram localizadas, até o momento do experimento, versões adaptadas para o português e para o domínio médico.

5.3.2 Avaliação dos resultados do reconhecimento de entidades (DDI)

Na avaliação dos resultados, inicialmente analisou-se um experimento que considera cada palavra individualmente no REN, seguido de um segundo resultado que foi implementar um algoritmo que decide se quer uni-las ou não. Ambos são descritos a seguir.

No caso do reconhecimento de palavras individuais, para o cálculo do desempenho desse resultado ainda é considerada individualmente cada palavra. Por exemplo, caso haja a entidade ”dipirona sódica” e o experimento identificar como ”remédio” e ”não-entidade”, considera-se como 1 acerto e 1 erro, enquanto, no melhor cenário, deveria considerar como 1 erro ou 1 acerto parcial.

Cada exemplo trabalhado no *dataset* é composto por um conjunto de quatro pares chave-valor. As chaves são ”*fold*”, ”*tokens*”, ”*true_types*” e ”*predicted_types*”:

- *Tokens*: são os termos de cada texto.
- *True_types*: são os tipos corretos das entidades.
- *Predicted_types*: são as predições do BERT, juntamente com a probabilidade que ele deu para cada classe.
- *Fold*: é o *fold* do exemplo.

Para o experimento, utilizou-se o procedimento de validação cruzada, onde separou-se o *dataset* em 5 subconjuntos, treinando 5 modelos diferentes. Para cada modelo, foi utilizado 1 subconjunto como teste, 1 subconjunto como validação, e 3 subconjuntos como treinamento. Os resultados relatados são apenas para os conjuntos de teste.

A tabela 22 mostra o desempenho do resultado desse experimento. A coluna Suporte apresenta quantos exemplos de cada classe que estão presentes no *dataset*.

Tabela 22: Desempenho por palavras individuais.

	Precisão	***Recall***	***F1 score***	**Suporte**
brand	94,65%	95,80%	95,22%	1571
Drug	95,34%	94,38%	94,86%	10276
Drug_n	74,60%	73,75%	74,17%	1063
group	89,11%	92,37%	90,71%	6334
Outra	99,40%	99,30%	99,35%	137681
Accuracy			98,49%	156925
Macro avg	90,62%	91,12%	90,86%	156925
Weighted avg	98,50%	98,49%	98,49%	156925

Elaborado pelo autor.

Conforme visto na tabela 22, o experimento apresenta bons resultados, chegando a 98,49% de *Accuracy* nos *Tokens*.

No caso de reconhecimento de palavras em grupo, para obter esses resultados trabalhou-se no problema de encontrar os limites de cada entidade dentro dos textos. O formato é similar ao anterior, contendo duas principais alterações.

A primeira consiste em adicionar uma classe especial para detectar o início de um tipo de entidade. Dessa forma, é possível encontrar uma quebra quando duas entidades diferentes estiverem separadas só por um espaço em branco. Sendo assim, nas métricas de desempenho por "*token*", também aparecerá classes com um prefixo "inicio" (Tabelas 23 e 24). Na segunda alteração o procedimento dedica-se à detecção das entidades por completo. Colocou-se também uma matriz de confusão para que possam ser medido os erros que o experimento mais comete (Figuras 36 e 37).

Tabela 23: Desempenho de *tokens* do experimento.

	Precisão	***Recall***	***F1 score***	**Suporte**
brand	96,06%	92,78%	94,39%	1551
Drug	95,63%	93,43%	94,52%	9970
Drug_n	75,69%	65,58%	70,27%	921
group	89,44%	88,00%	88,71%	5898
Inicio_*brand*	21,52%	85,00%	34,34%	20
Inicio_*drug*	47,49%	80,39%	59,71%	306

Inicio_ *drug_n*	58,00%	81,69%	67,84%	142
Inicio_*group*	53,79%	87,84%	66,72%	436
Outra	99,39%	99,32%	99,35%	137681
Accuracy			98,17%	156925
Macro avg	70,78%	86,00%	75,10%	156925
Weighted avg	98,33%	98,17%	98,22%	156925

Elaborado pelo autor.

Tabela 24: Desempenho de entidades do experimento.

	Precisão	***Recall***	***F1 score***	**Suporte**
brand	98,61%	97,87%	98,24%	2254
Drug	98,78%	97,74%	98,25%	15026
Drug_n	90,75%	70,38%	79,28%	655
group	97,76%	95,19%	96,46%	5323
Outra	0	0	0	0
Accuracy			98,17%	156925
Macro avg	70,78%	86,00%	75,10%	156925
Weighted avg	98,33%	98,17%	98,22%	156925

Elaborado pelo autor.

Mesmo com a aplicação em grupos de palavras, o experimento continuou apresentando um bom resultado, exibindo uma *accuracy* de 98,17%. Para visualizar melhor esse resultado, um exemplo de texto medicamentoso é descrito no Quadro 3.

Quadro 3 – Exemplo de texto medicamentoso

Other Drugs: Based on the results ofdruginteraction studies, no dosage adjustment is recommended when SUSTIVA (efavirenz) is given with the following: aluminum/magnesium hydroxide antacids, azithromycin, cetirizine, famotidine, fluconazole, lamivudine, lorazepam, nelfinavir, paroxetine, and zidovudine.

A Tabela 25 ilustra a frase do Quadro 3 de acordo com as entidades reconhecidas e seus resultados. Para cada *Token*, há uma *Tru_types*, que representa a classe correta daquela palavra. Por fim, é mostrada a classe que o modelo reconheceu *(predicted_types)* e a probabilidade de certeza que ele apresentou pra cada classe.

Tabela 25: Exemplos de Classificação

Ordem	***tokens***	***True_types***	***Predicted_types***
0	*Other*	outra	[outra, 99,96%]
1	*Drugs*	outra	[outra, 99,95%]
2	*:*	outra	[outra, 99,97%]
3	*Based*	outra	[outra, 99,97%]
4	*on*	outra	[outra, 99,97%]
5	*the*	outra	[outra, 99,97%]
6	*results*	outra	[outra, 99,97%]
7	*of*	outra	[outra, 99,97%]
8	*drug*	outra	[outra, 99,96%]
9	*interaction*	outra	[outra, 99,97%]
10	*studies*	outra	[outra, 99,97%]
11	*,*	outra	[outra, 99,97%]
12	*no*	outra	[outra, 99,97%]
13	*dosage*	outra	[outra, 99,97%]
14	*adjustment*	outra	[outra, 99,97%]
15	*is*	outra	[outra, 99,97%]
16	*recommended*	outra	[outra, 99,97%]
17	*when*	outra	[outra, 99,97%]
18	*SUSTIVA*	*brand*	[*brand*, 74,81%]
19	*(*	outra	[outra, 99,96%]
20	*efavirenz*	*drug*	[*drug*, 63,64%]
21	*)*	outra	[outra, 99,96%]
22	*is*	outra	[outra, 99,97%]
23	*given*	outra	[outra, 99,97%]
24	*with*	outra	[outra, 99,97%]
25	*the*	outra	[outra, 99,97%]
26	*following*	outra	[outra, 99,96%]
27	*:*	outra	[outra, 99,97%]
28	*aluminum*	*group*	[*drug*, 64,70%]
29	*/*	*group*	[outra, 97,21%]
30	*magnesium*	*group*	[*drug*, 63,64%]
31	hydroxide	*group*	[outra, 46,32%]
32	antacids	*group*	[*group*, 68,51%]
33	,	outra	[outra, 99,97%]
34	azithromycin	*drug*	[*drug*, 79,10%]

Continua na próxima página

Tabela 25 – continua da página anterior

Ordem	*tokens*	*True_types*	*Predicted_types*
35	,	outra	[outra, 99,97%]
36	cetirizine	*drug*	[*drug*, 65,95%]
37	,	outra	[outra, 99,97%]
38	*famotidine*	*drug*	[*drug*, 74,24%]
39	,	outra	[outra, 99,97%]
40	*fluconazole*	*drug*	[*drug*, 79,23%]
41	,	outra	[outra, 99,97%]
42	*lamivudine*	*drug*	[*drug*, 74,25%]
43	,	outra	[outra, 99,97%]
44	*lorazepam*	*drug*	[*drug*, 74,27%]
45	,	outra	[outra, 99,97%]
46	*nelfinavir*	*drug*	[*drug*, 65,95%]
47	,	outra	[outra, 99,97%]
48	*paroxetine*	*drug*	[*drug*, 74,08%]
49	,	outra	[outra, 99,97%]
50	*and*	outra	[outra, 99,97%]
51	*zidovudine*	*drug*	[*drug*, 74,13%]
52	.	outra	[outra, 99,97%]

Elaborado pelo autor.

Para comparar, a figura 35 ilustra um gráfico comparativo de *F1-Score* entre palavras individuais, *tokens* e Entidades.

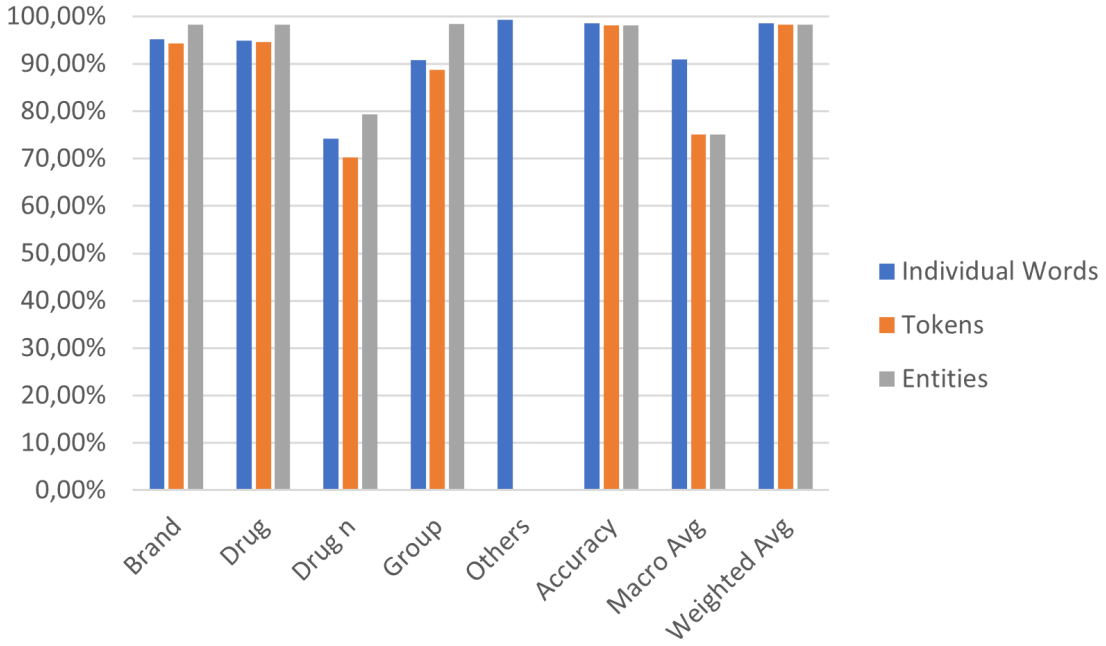
Foram elaboradas duas matrizes de confusão para ilustrar quais são os erros mais frequentes e exibir onde o modelo está se confundindo. A Figura 36 ilustra a Matriz de Confusão em quantidades e a Figura 37 em porcentagens. As linhas são os rótulos do *dataset* e a coluna indica o resultado apresentado pelo BERT.

Em seguida, esses resultados foram treinados na BiLSTM, para que pudessem ser aplicados experimentos de ER, a fim de extrair a relação entre os *tokens*/entidades, descritos a seguir.

5.3.3 Experimento de extração de relações com o *dataset* DDI

Neste experimento, foi aplicada uma abordagem com BiLSTM para identificar pares (ou trios, etc.) de entidades e seus tipos, com o objetivo de classificar as relações existentes entre elas. Com base nos tipos das entidades presentes no texto e nas entidades selecionadas para a relação, a classificação é realizada. A BiLSTM é executada várias vezes separadamente para

Figura 35 – Comparativo de resultados



Fonte: Elaborado pelo autor

cada par, trio ou grupo de entidades. Os *tokens* especiais são usados para indicar o início e o fim das entidades, que são reconhecidas pelo modelo BERT.

Para implementar essa abordagem, foi utilizada a biblioteca *PyTorch* da API (*Application Programming Interface*) *Hugging Face*. Foi adicionado um classificador de nível de *token* no topo dos modelos BERT. No experimento, foram utilizados dados do conjunto de dados DDI, principalmente relacionados ao experimento anterior. No exemplo apresentado, três pares de entidades foram classificados, além de um trio. A Tabela 26 ilustra esse exemplo.

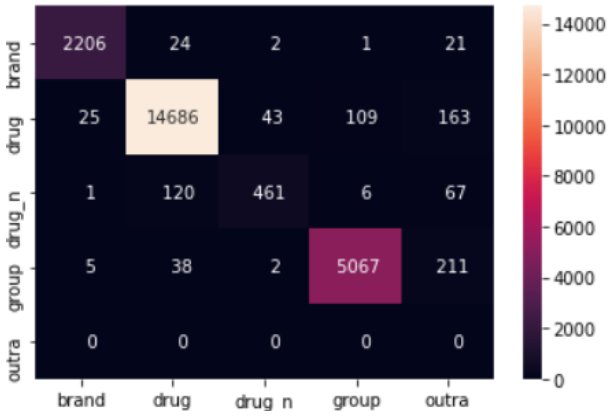
Em relação à arquitetura para a ER, uma modificação foi feita no BERT original. Antes da etapa de classificação, na última camada de processamento, a atenção é zerada para todas as palavras que não fazem parte do par de entidades para o qual a relação está sendo classificada. Ou seja, o texto é processado para cada par de entidades e, no final, apenas as representações das entidades em questão são mantidas.

Essa arquitetura apresentou melhores resultados do que o uso de RNR, que era a primeira escolha nesta pesquisa. O BERT utilizado para a ER é inicializado com os mesmos pesos do BERT treinado para o REN. Dessa forma, o treinamento anterior é aproveitado para obter um melhor desempenho nessa tarefa.

5.3.4 Avaliação dos resultados de extração de relações (DDI)

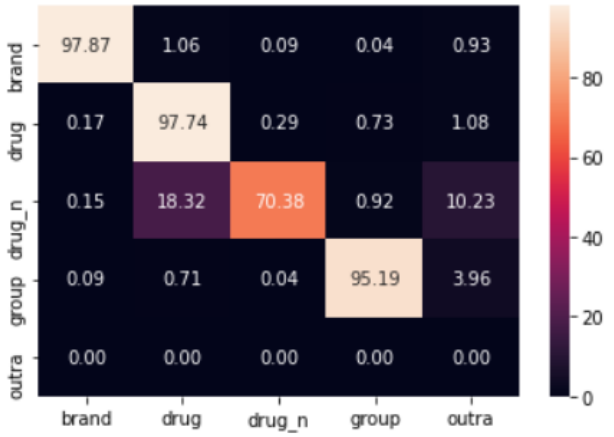
Foi realizada uma modificação no BERT para permitir a entrada adicional no segundo modelo (ER), no qual as posições das entidades são passadas como informação. Assim, a representação final do texto utiliza apenas as representações das entidades. O primeiro BERT (REN)

Figura 36 – Matriz de confusão (números absolutos)



Fonte: Elaborado pelo autor

Figura 37 – Matriz de confusão (porcentagem relativa à quantidade de rótulos)



Fonte: Elaborado pelo autor

encontra as entidades e, em seguida, o segundo BERT (ER) informa a relação entre cada par de entidades, permitindo a verificação de relações entre três, quatro ou mais entidades. No conjunto de dados DDI, que foi utilizado como base para a implementação, as relações conhecidas têm no máximo duas entidades.

Das 27.479 relações do conjunto de dados, 3.381 (12,3%) não tiveram suas entidades encontradas, sendo consideradas erros. Dos 87,7% restantes, apresentamos a seguir o desempenho do modelo. Considerando 12,3% como erro, o ajuste do *F1-score* de 92,00% para 87,7% dos dados resultou em um desempenho final de 80,68%. De acordo com o site "*Papers with Code*" [5], o melhor resultado relatado na literatura é de 84,08%, o que aproxima a arquitetura proposta do estado da arte. A Tabela 27 apresenta os resultados obtidos.

[5]https://paperswithcode.com/sota/drug-drug-interaction-extraction-on-ddi

Tabela 26 – Exemplos de relação

Texto de Entrada	**Relação**
Prolonged recovery time may occur if <INI_ENTIDADE> barbiturates <FIM_ENTIDADE> and/or <INI_ENTIDADE> narcotics <FIM_ENTIDADE> are used concurrently with ketamine."	Nenhuma (o experimento terá que falar que não há relação)
Prolonged recovery time may occur if <INI_ENTIDADE> barbiturates <FIM_ENTIDADE> and/or narcotics are used concurrently with <INI_ENTIDADE> ketamine <FIM_ENTIDADE>."	Efeito
Prolonged recovery time may occur if <INI_ENTIDADE> barbiturates <FIM_ENTIDADE> and/or <INI_ENTIDADE> narcotics <FIM_ENTIDADE> are used concurrently with ketamine."	Efeito
Prolonged recovery time may occur if <INI_ENTIDADE> barbiturates <FIM_ENTIDADE> and/or <INI_ENTIDADE> narcotics <FIM_ENTIDADE> are used concurrently with <INI_ENTIDADE> ketamine <FIM_ENTIDADE>."	Nenhuma

Fonte: Elaborado pelo autor.

Tabela 27 – Resultado por relações

	Precision	*recall*	*F1-score*	Support
Advise	0.6798	0.7771	0.7252	776
Effect	0.7605	0.6980	0.7279	1424
Int	0.4752	0.2275	0.3504	173
Mechanism	0.6406	0.7445	0.6887	1233
None	0.9623	0.9566	0.9595	20492
Accuracy			0.9198	24098
Macro avg	0.7037	0.6907	0.6903	24098
Weighted avg	0.9214	0.9198	0.9200	24098

5.4 Extração de entidades e relações com *datasets* DGO-E e DGO-CD

Após a conclusão do processo de anotação dos *datasets* DGO-E e DGO-CD (descritos nas seções 5.1.3 e 5.1.4), iniciou-se um novo ciclo de experimentos, destinados a avaliar as possibilidades de extração de entidades e relações com base no material anotado, originado do sistema utilizado no estudo de caso. Estes experimentos e seus resultados são descritos a seguir.

5.4.1 Experimento de Reconhecimento de Entidades (DGO-E)

Inicialmente, foi realizado um teste de treinamento utilizando a mesma estrutura descrita na seção 5.3. Os resultados obtidos não foram considerados positivos o suficiente. Sendo assim, iniciou-se um processo de ajuste do modelo utilizado.

O processo de construção do modelo de inteligência artificial para o REN foi realizada

utilizando inicialmente a biblioteca *Spacy*[6], implementada na versão Python 3.10. O *Spacy* é uma biblioteca de PLN que oferece recursos avançados para análise e extração de informações de texto. Essa biblioteca foi utilizada para preparar e processar os dados durante o treinamento.

Foram realizados dois treinamentos distintos, cada um com 20 épocas, permitindo que o modelo aprendesse gradualmente a identificar as entidades presentes nos dados. A escolha da RN foi o modelo BERT, uma arquitetura baseada em *Transformers*. O treinamento foi realizado em um ambiente de aprendizado supervisionado, onde os dados de treinamento possuem textos rotulados com suas respectivas classes.

No primeiro treinamento, foram utilizadas 716 amostras anotadas, correspondendo a 70% do *dataset* disponível para o REN. Os dados foram lidos a partir de arquivos JSON que contêm os textos e suas respectivas classes, e foram organizados em *dataframes* (uma estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas) do Pandas para facilitar o processamento e manipulação.

Antes do treinamento, foi realizada uma etapa de limpeza e tratamento dos dados, visando remover possíveis *outliers*, dados nulos ou corrompidos. No entanto, nenhum problema desse tipo foi encontrado nos dados utilizados. Além disso, foi aplicada a *tokenização* nos textos, dividindo-os em unidades menores, como palavras ou subpalavras. Em seguida, os textos foram pré-processados utilizando o *token*izer do BERT, convertendo-os em sequências de *tokens* e adicionando *tokens* especiais para marcar o início e o fim das sequências. As sequências de *tokens* foram ajustadas para terem o mesmo tamanho máximo especificado. O modelo foi carregado a partir de um modelo pré-treinado utilizando a biblioteca *Transformers*.

Devido à limitação de tamanho de entrada do BERT, foi necessário fazer alterações em relação ao experimento com o DDI (descrito na seção 5.3). No experimento anterior, os primeiros 512 *tokens* eram utilizados, descartando o restante. Modificou-se o algoritmo para percorrer o texto em blocos de 128 *tokens*, considerando a classificação de cada *token* no momento em que ele estava mais próximo do centro. Em resumo, o tamanho máximo da sequência (S) foi configurado para 512 com um passo (P) de 128.

Em um *token* 400, por exemplo. Na primeira iteração, ele estava a uma distância de 144 *tokens* do centro da leitura (que varia de 0 a 511). Na segunda iteração, ele estava a 16 *tokens* do centro (o texto avançou 128 *tokens*). Na terceira iteração, a distância era de 112 *tokens*, na quarta iteração era de 240 *tokens* e, na quinta iteração, o *token* 400 já não fazia mais parte do texto lido pelo BERT. Portanto, para o *token* 400 será considerado a saída dada na segunda execução.

Outra alteração em relação ao com o DDI (onde foi utilizado 20 épocas), aqui o treinamento foi realizado em 50 épocas, com um tamanho de lote (*batch size*) de 16. A otimização foi feita usando o algoritmo AdamW (LOSHCHILOV; HUTTER, 2017) com os parâmetros de β 1 = 0,9 e β 2 = 0,999. Foi aplicado um decaimento de peso (*weight decay*) de 0,1 e a taxa de aprendizado inicial foi definida como 1e-5, com um decaimento linear e um período de

[6]https://spacy.io

aquecimento (*warmup*) nos primeiros 100 passos.

Durante o treinamento, o processo de *fine-tuning* foi aplicado. O modelo é treinado com o *dataset* DGO-E e seus exemplos rotulados para a tarefa específica. Neste caso substitui-se a camada de classificação final do BERT por uma nova camada adequada à tarefa, como uma camada de classificação de entidades. O *fine-tuning* envolve o treinamento do modelo com os dados anotados, utilizando algoritmos de otimização, para ajustar os pesos do modelo com o objetivo de minimizar a função de perda, que mede a discrepância entre as previsões do modelo e os rótulos reais das entidades.

Durante o treinamento, os dados de treinamento foram fornecidos em lotes (*batches*) para o modelo, e um otimizador foi utilizado para ajustar os pesos do modelo com base na função de perda calculada entre as previsões e os rótulos verdadeiros das entidades. O treinamento foi realizado por 50 épocas, sendo que cada época percorreu todos os lotes de treinamento. Foi utilizada a biblioteca *Transformers* do *Hugging Face* para carregar o modelo BERT pré-treinado e realizar o *fine-tuning*.

Durante o treinamento, todos os textos anotados foram percorridos e verificados. Em seguida, foi feita a separação das entidades presentes nos textos, criando um dicionário que mapeia essas informações. As entidades dos textos anotados foram adicionadas ao *pipeline* do modelo, permitindo que ele aprendesse a reconhecer e classificar corretamente as entidades e relações. Além disso, foi realizado o pré-carregamento de um modelo já treinado fornecido pela biblioteca *Spacy*. Isso auxiliou no processo do PLN, fornecendo recursos adicionais ao modelo e contribuindo para um melhor desempenho.

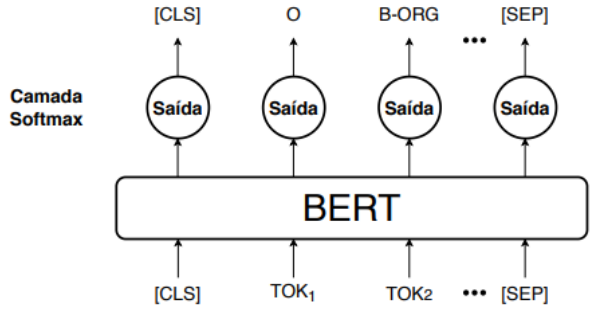
Durante o treinamento, foi aplicada aleatoriedade nos dados, evitando um possível *overfitting* do modelo. Essa abordagem ajuda a garantir que o modelo seja capaz de generalizar para novos dados e não se torne excessivamente dependente dos exemplos de treinamento. As entidades foram processadas em conjunto, permitindo que o modelo aprendesse a identificar as entidades nomeadas. Ao longo do processo de treinamento, foram utilizadas diversas bibliotecas, como *re, random, numpy, pandas, matplotlib, sklearn, tensorflow* e *keras*, que forneceram funcionalidades adicionais e facilitaram o desenvolvimento e treinamento do modelo.

Após o treinamento, o modelo foi avaliado utilizando os dados de validação. Foram obtidas as predições do modelo para os textos de validação, e métricas de avaliação, como acurácia e matriz de confusão, foram calculadas para avaliar o desempenho do modelo. Além disso, foi gerado um relatório de classificação contendo métricas como precisão, *recall* e *f1-score* para cada classe.

O modelo treinado pode ser utilizado para realizar inferências em novos textos. Os textos de teste são processados e as predições são obtidas a partir do modelo. Além das predições, também são extraídos *embeddings* dos textos usando a representação do *token* [CLS] do BERT, capturando as informações contextuais do texto. A Figura 38 apresenta uma ilustração do BERT no REN.

Os resultados do treinamento, incluindo os pesos do modelo para cada época, estatísticas de

Figura 38 – BERT no REN



Fonte: Pelo autor

treinamento e avaliação, predições e *embeddings*, são armazenados em arquivos para posterior análise. O processo de treinamento foi controlado por diversos parâmetros, como o tamanho máximo dos *tokens*, tamanho do lote, taxa de aprendizado, número de épocas, inicialização do BERT, equilíbrio de perda (para lidar com desequilíbrio de classes) e outros parâmetros específicos da tarefa. Com base nessas etapas, o modelo de inteligência artificial foi construído, treinado com dados anotados e ajustado para reconhecer entidades. O objetivo desse modelo é fornecer uma ferramenta precisa e eficiente para análise e extração de informações relevantes em texto, auxiliando em tarefas de PLN e facilitando a compreensão e o uso de dados não estruturados.

5.4.2 Reconhecimento de Relações (DGO-E)

Além do uso do modelo BERT para REN, foi realizado o treinamento de uma BILSTM para o reconhecimento de relações na tarefa de ER. Inicialmente os dados foram preparados para o treinamento. Da mesma forma que no caso do modelo BERT, foi feita a transformação dos dados em um formato adequado, como um *DataFrame*, facilitando o processamento subsequente. Foi fundamental gerar uma estrutura de dados para representar corretamente as relações entre as entidades presentes no texto.

Em seguida, as palavras do texto foram codificadas em representações numéricas. Essa codificação foi realizada com o auxílio das técnicas de *Word embeddings Word2Vec* e *GloVe*. As representações numéricas capturam informações semânticas das palavras, permitindo que o modelo aprenda as relações entre elas.

A arquitetura da BILSTM foi construída com camadas de células LSTM. A LSTM é uma unidade de processamento recorrente capaz de aprender e lembrar informações de longo prazo. A utilização da estrutura bidirecional implica em uma camada processando a sequência no sentido direto (*forward*) e outra no sentido inverso (*backward*). Com isso, o modelo foi capaz de capturar informações contextuais tanto do passado quanto do futuro em relação a cada palavra na sequência, melhorando a compreensão global do contexto.

Os dados preparados e codificados foram então utilizados para o treinamento da BILSTM. A rede foi treinada para aprender a extrair as relações entre as entidades presentes nos dados de treinamento. Durante o treinamento, uma função de perda foi definida para medir a discrepância entre as relações extraídas pela BILSTM e as relações reais presentes nos dados. A escolha adequada da função de perda é essencial para orientar o treinamento da rede e minimizar os erros.

Para inicializar os pesos da BILSTM, foi aplicada uma inicialização aleatória próxima de zero, conforme sugerido pela inicialização Xavier (XU; WANG; HE, 2018). Essa abordagem auxilia na prevenção de problemas de saturação ou explosão dos gradientes durante o treinamento. Os pesos da BILSTM são ajustados utilizando o algoritmo de *backpropagation* (LI et al., 2023), que propaga os gradientes da função de perda pela rede. Esse processo de ajuste de pesos busca reduzir a perda e aprimorar o desempenho da rede na ER. O treinamento foi realizado em várias épocas, ou seja, iterações completas pelos dados de treinamento. A cada época, os pesos da rede são atualizados com base nos gradientes calculados, permitindo que a rede aprenda as relações entre as entidades de forma mais precisa.

Para guiar o processo de atualização dos pesos da rede, é utilizado o otimizador Adam. O Adam combina as vantagens de dois outros otimizadores, o RMSprop e o *momentum* do SGD (*Stochastic Gradient Descent*). O otimizador Adam utiliza estimativas adaptativas do momento de primeira ordem (média móvel dos gradientes) e do momento de segunda ordem (média móvel do quadrado dos gradientes) para ajustar os pesos da rede. Essas estimativas adaptativas permitem um ajuste mais eficiente da taxa de aprendizado para cada parâmetro da rede.

O Adam possui hiperparâmetros, como a taxa de aprendizado (*learning rate*), o decaimento do aprendizado (*learning rate decay*), o momento ($\beta 1$) e o decaimento do momento ($\beta 2$). A taxa de aprendizado controla a velocidade de atualização dos pesos durante o treinamento. O decaimento do aprendizado permite reduzir gradualmente a taxa de aprendizado à medida que o treinamento avança. O momento controla a contribuição do momento de primeira ordem no processo de atualização dos pesos, enquanto o decaimento do momento permite reduzir gradualmente essa contribuição. Sendo assim, Para otimizar a BiLSTM, foram adotados os seguintes parâmetros: $\beta 1 = 0{,}9$, $\beta 2 = 0{,}999$ e ϵ = 1e-7. A taxa de aprendizado inicial foi definida como 1e-3, com um decaimento de 1e-5. Realizamos experimentos para determinar os valores de S = 128 e P = 64, os quais mostraram um desempenho superior para o modelo. A arquitetura de base da BiLSTM foi testada com esses parâmetros, enquanto os demais hiperparâmetros foram ajustados. O treinamento da BiLSTM foi realizado por 75 épocas.

Para a regularização da BILSTM, foram aplicadas as técnicas de regularização L1 e L2, que adicionam termos à função de perda para penalizar pesos grandes, juntamente com a técnica de *Dropout*, que desativa aleatoriamente um conjunto de unidades durante o treinamento, reduzindo a coadaptação entre elas. Essas técnicas de regularização foram utilizadas para evitar o *overfitting* e melhorar a capacidade de generalização do modelo. Durante o treinamento, os dados foram divididos em *batches* (ou *mini-batches*) visando a eficiência computacional. Essa

divisão permite realizar cálculos em paralelo e otimizar o uso dos recursos disponíveis.

5.4.3 Reconhecimento de entidades com DGO-CD

Para avaliar o desempenho do modelo, foi conduzido um treinamento adicional utilizando o conjunto de dados DGO-CD sem realizar quaisquer modificações. Dado que o DGO-CD contém apenas entidades anotadas, a avaliação foi focada exclusivamente na arquitetura de REN, conforme descrito na seção 5.3.1. A seguir, são apresentados os resultados obtidos nos experimentos.

5.4.4 Avaliação dos resultados

Para avaliar os resultados do DGO-E, foram utilizadas 305 evoluções anotadas (30% do *dataset*) e realizados testes no modelo. Os resultados de Precisão, *Recall*, F1 *Score*, Suporte, Acurácia, Média Macro e Média Ponderada estão ilustrados na Tabela 28.

Tabela 28: Desempenho de REN do DGO-E.

	Precisão	***Recall***	***F1 score***	**Suporte**
Exame	78,92%	77,27%	78,77%	14648
Membro	78,26%	77,74%	78,25%	3311
Resultado	70,01%	70,38%	70,59%	11504
Data	77,76%	75,14%	76,33%	3501
Tempo	73,33%	72,17%	73,56%	592
Accuracy			78,24%	35994
Macro avg	72,78%	76,01%	75,10%	35994
Weighted avg	78,32%	78,76%	78,66%	35994

Durante o treinamento, foi realizada uma avaliação do desempenho da BILSTM em um conjunto de validação, que consiste em dados não utilizados durante o treinamento (30% do DGO-E). A saída da BiLSTM é comparada com os rótulos reais das relações presentes nos exemplos de treinamento. Para medir a discrepância entre as previsões do modelo e os rótulos verdadeiros, foi utilizada uma função de perda, neste caso a entropia cruzada. Os resultados obtidos são apresentados na Tabela 29.

Tabela 29: Desempenho de reconhecimento de relações do DGO-E.

	Precisão	***Recall***	***F1 score***	**Suporte**
Resultado	76,61%	75,87%	75,24%	20983

Quando	73,45%	73,55%	74,12%	2100
Localização	76,75%	75,01%	75,28%	2501
Accuracy			76,17%	25584
Macro avg	76,78%	76,00%	75,10%	25584
Weighted avg	76,33%	78,17%	78,22%	25584

Elaborado pelo autor.

Para avaliar o resultado, treinou-se o modelo sem nenhuma modiciação com o DGO-CD. A tabela 30 apresenta os resultados de REN do DGO-CD.

Tabela 30: Desempenho de REN do DGO-CD.

	Precisão	***Recall***	***F1 score***	**Suporte**
Característica do Diagnóstico	70,22%	70,37%	71,24%	230
Órgão	73,71%	72,74%	72,98%	189
Local do Órgão	70,12%	70,17%	70,99%	71
Estadiamento	71,73%	70,13%	71,00%	246
Tempo	71,70%	71,19%	71,46%	24
Accuracy			72,87%	760
Macro avg	72,78%	72,00%	72,10%	760
Weighted avg	72,33%	72,17%	72,33%	760

Elaborado pelo autor.

A seguir são apresentados os resultados do REN e ER nos conjuntos de dados DGO-E e DGO-CD, utilizando o modelo BERT.

Na Tabela 28, são apresentados os resultados do REN no conjunto de dados DGO-E. Podemos observar que as entidades ”Exame”, ”Membro”, ”Resultado”, ”Data” e ”Tempo” foram reconhecidas com precisões variando entre 70,01% e 78,92%. O *recall* dessas entidades varia de 70,38% a 77,27%, enquanto o *F1-score* varia de 70,59% a 78,77%. Esses valores indicam que o modelo apresentou um desempenho razoável na detecção e classificação dessas entidades. O suporte, que representa o número de instâncias de cada entidade no conjunto de dados, varia de 3.311 a 14.648. O desempenho geral do modelo, medido pela acurácia, é de 78,24%.

Na Tabela 29, são apresentados os resultados do reconhecimento de relações no conjunto de dados DGO-E. As relações ”Resultado”, ”Quando” e ”Localização” alcançaram precisões de 73,45% a 76,75%, *recalls* de 73,55% a 75,87% e *F1-scores* de 74,12% a 75,28%. Esses resultados indicam que o modelo obteve um desempenho moderado na detecção e classificação

das relações nesse conjunto de dados. O suporte das relações varia de 2.100 a 20.983. A acurácia geral do modelo é de 76,17%.

Na Tabela 30, são apresentados os resultados do RENs no conjunto de dados DGO-CD. As entidades ”Característica do Diagnóstico”, ”Órgão”, ”Local do Órgão”, ”Estadiamento” e ”Tempo” alcançaram precisões de 70,12% a 73,71%, *recalls* de 70,17% a 72,74% e *F1-scores* de 70,99% a 72,98%. Esses resultados indicam que o modelo teve um desempenho razoável na detecção e classificação dessas entidades no conjunto de dados DGO-CD. O suporte das entidades varia de 24 a 230. A acurácia geral do modelo é de 72,87%.

No geral, os resultados obtidos são satisfatórios. Ao analisar as métricas, pode-se observar que as categorias apresentam desempenho semelhante, com valores de precisão, *recall* e *F1-score* próximos. A maioria das categorias alcançou valores acima de 70% nessas métricas. Isso indica que o modelo é capaz de identificar corretamente as entidades e relações de maneira consistente.

Em termos gerais, o modelo BERT demonstrou um bom desempenho no REN e ER nos dois domínios avaliados. É importante destacar que esses resultados devem ser interpretados considerando as características específicas dos conjuntos de dados utilizados. No contexto do domínio médico, o RNE e a ER pode ser desafiadora devido à complexidade e diversidade dos textos médicos. Portanto, os resultados obtidos podem ser considerados satisfatórios, levando em conta a dificuldade da tarefa.

No entanto, é importante considerar que esses resultados são específicos para os conjuntos de dados e tarefas em questão. Eles não são diretamente comparáveis com outros experimentos na área devido às características particulares dos conjuntos de dados DGO-E e DGO-CD, que são específicos do domínio da saúde e contêm dados exclusivos da empresa e dos pacientes.

Ao contrário dos modelos presentes na literatura, que geralmente são pré-treinados em grandes conjuntos de dados gerais, como a Wikipédia ou dados fictícios, o uso de conjuntos de dados próprios, anotados por especialistas da área da saúde, permite fornecer contexto específico do domínio para a RN. Isso representa uma contribuição significativa para a pesquisa. Embora os resultados não sejam diretamente comparáveis aos estudos anteriores devido às razões mencionadas, eles contribuem para o avanço da área e podem servir como referência em futuros estudos comparativos.

Além disso, é importante considerar que os resultados podem ser influenciados pela quantidade de dados disponíveis para treinamento. Se houver mais dados anotados, é possível que o desempenho do modelo melhore, pois terá mais informações para aprender e generalizar.

5.5 Experimento com Ontologia

Neste item são descritos aspectos da ontologia de domínio utilizada para experimentos de integração de dados obtidos nos procedimentos de REN e ER.

5.5.1 Construção e teste da Ontologia

A construção da ontologia foi iniciada em trabalho anterior e publicada em Schwertner et al. (2019), sendo continuada nessa pesquisa. A seguir são descritos os procedimentos utilizados na sua criação e posterior adaptação.

Inicialmente, a fim de analisar os termos relevantes do domínio, foram exportados do SGO um total de 3.309 documentos XML contendo consultas médicas e anotações de profissionais de saúde. Esses documentos totalizaram 212.829 palavras em 28.178 linhas. A partir do estudo desses conteúdos, foram identificados conjuntos de informações que podem ser obtidas automaticamente por meio da extração de informações. Essas informações complementares são frequentemente necessárias para os profissionais de saúde em seu trabalho diário. Um exemplo claro dessas situações é observado em frases que descrevem a percepção de sintomas ou o uso de medicamentos. Um procedimento manual de estudo e contato com profissionais especialistas em saúde e tecnologia foi realizado, resultando em um conjunto de frases significativas e conceitos próprios que serviram como ponto de partida para este trabalho.

Com a participação de profissionais que utilizam o SGO, foi possível definir um conjunto típico de sentenças de interesse. Foram realizadas entrevistas que tinham como objetivo identificar um conjunto crítico de questões relevantes e fornecer as bases para a modelagem do conhecimento necessário a ser representado. Com os principais conceitos e relações definidos na base de conhecimento, uma análise textual manual foi realizada para preencher essas informações. Esse processo foi adotado para garantir que a base de conhecimento fosse precisa e confiável, servindo como ponto de partida para futuras iniciativas de extração automática de informações textuais. Algumas das sentenças descritas são exemplificadas a seguir:

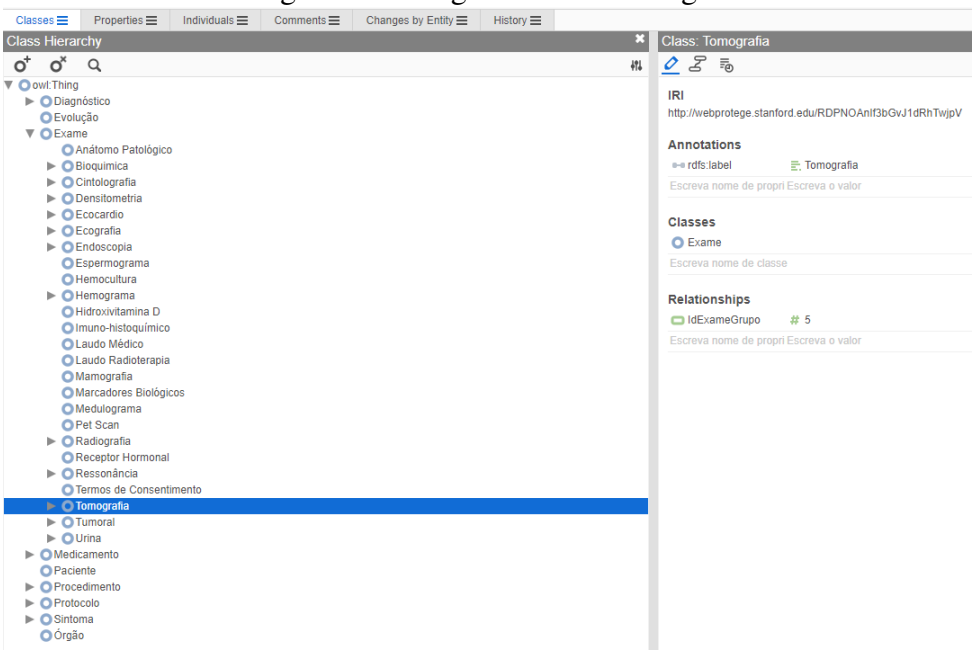
- Quais os sintomas que o paciente **PacienteNome** relata?
- Qual o diagnóstico do paciente **PacienteNome**?
- Quais pacientes possuem o diagnóstico **Diagnostico**?
- Quais sintomas os pacientes com diagnóstico **Diagnostico** relatam?
- O sintoma **sintoma** está associado a quais diagnósticos?
- Quais sintomas aparecem no órgão **orgão**?
- Há quanto tempo paciente notou **sintoma** no órgão **orgão**?
- Quais são os protocolos solicitados para o paciente **PacienteNome**?
- Quais os medicamentos que o paciente **PacienteNome** faz uso?
- Quais pacientes usam o medicamento **medicamento**?

A partir dessas sentenças, foi criada a estrutura de conhecimento necessária para armazenar os conceitos e relações essenciais para responder a essas perguntas. Uma base de conhecimento armazena os conceitos e relações definidos, bem como as instâncias desses conceitos e relações, através de um processo de extração automática de informações textuais que identifica essas instâncias. A análise das informações textuais permitiu identificar os principais conceitos em cada frase, que foram usados para criar a base de conhecimento. Além disso, informações sobre a estrutura típica das sentenças foram valiosas para o componente de resposta às perguntas.

Neste momento, optou-se por não utilizar formalismos padrão em saúde, como o openEHR2 (MIN et al., 2018) e o SNOMED CT® (IBRAHIM et al., 2014). Essas ontologias são abrangentes e foram projetadas para representar todo o conhecimento na área de saúde e suas relações. Em vez disso, optou-se por descrever uma ontologia mais flexível que represente os conceitos e relações identificados nas sentenças selecionadas, a fim de investigar a integração do sistema de resposta a perguntas precvisto para ser explorado em conjunto com a ontologia. A estrutura da ontologia determina o formato das consultas e as possibilidades de inferência, e, portanto, uma nova ontologia foi criada para ter maior flexibilidade nesse aspecto.

A ontologia foi desenvolvida usando o software Protégé[7], sendo composta por 181 classes, 14 propriedades de dados e 12 propriedades de objetos. A Figura 39 ilustra um fragmento da ontologia.

Figura 39 – Fragmento da Ontologia



Fonte: Elaborado pelo autor

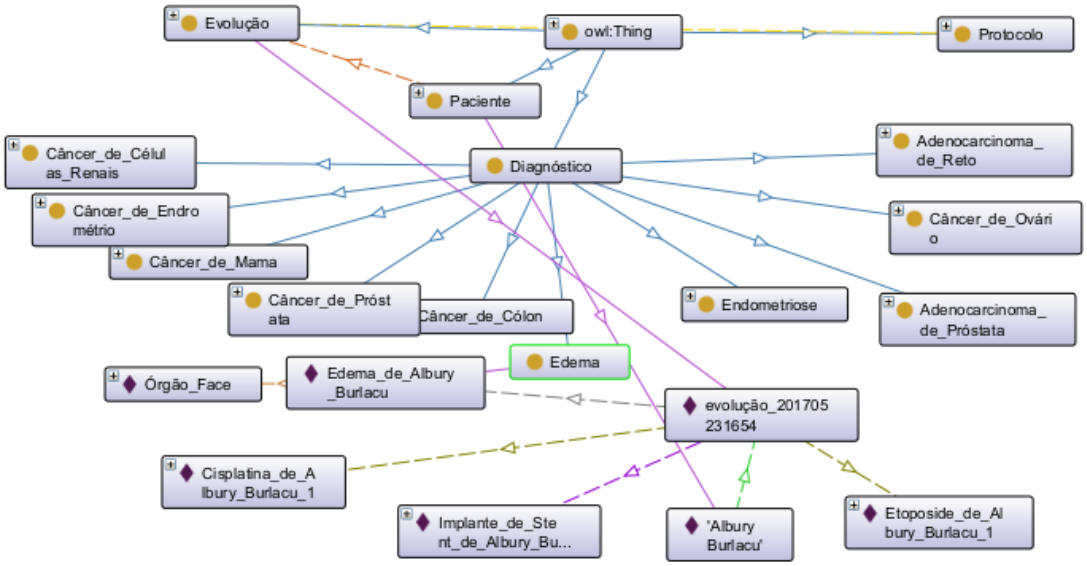
Conforme ilustrado na Figura 39, a ontologia descreve várias classes, como diagnóstico, evolução, exame, medicamento, paciente, procedimento, protocolo, sintoma e órgão, que são relevantes para o domínio da oncologia.

[7]https://protege.stanford.edu/

Além disso, a ontologia define propriedades de objetos que estabelecem relações entre as diferentes entidades, como a relação entre pacientes e sintomas, pacientes e medicamentos, medicamentos e unidades de medida, exames ou procedimentos e localizações no corpo, pacientes e diagnósticos, pacientes e exames ou procedimentos, pacientes e resultados de exames, e exames e unidades de medida. Também são definidas propriedades de dados que representam características específicas dos dados, como identificadores de exames, características de exames, ciclos de tratamento, datas de eventos, estadiamento do câncer, tipos de exames, intensidade de sintomas, localizações no corpo, identificadores de pacientes, períodos de tempo, posologia de medicamentos e resultados de exames.

A ontologia fornece uma estrutura organizada para representar e relacionar dados relevantes à oncologia, permitindo uma melhor compreensão e análise dos dados clínicos. O apêndice C apresenta a ontologia completa e o apêndice B contém uma lista de relações e suas consultas na ontologia. Para exemplificar, na Figura 40 são destacados os conceitos centrais de “Evolução” e ”Edema” com suas principais relações, com exemplos de algumas instâncias da ontologia criada.

Figura 40 – Conceitos centrais de Edema



Fonte: Elaborado pelo autor

A ontologia construída representa informações relevantes para o campo da oncologia, especificamente para esse domínio. Ela é composta por classes de entidades, propriedades de objetos e propriedades de dados que descrevem e relacionam aspectos importantes relacionados ao diagnóstico, tratamento e acompanhamento de pacientes com câncer. As propriedades de objeto fornecem relações significativas entre essas classes, possibilitando estabelecer ligações importantes entre pacientes, sintomas, medicamentos, exames e outros elementos relevantes para o tratamento e diagnóstico do câncer. Além disso, as propriedades de dados definem atributos específicos para cada classe, permitindo a captura precisa e organizada de informações essenciais, como identificadores únicos, datas de eventos, intensidade de sintomas e resultados

de exames.

A abordagem detalhada e bem estruturada da ontologia pode facilitar a integração dos dados clínicos, auxiliando profissionais de saúde na interpretação, análise e tomada de decisões para pacientes com câncer. Com as relações bem definidas, a ontologia pode servir como base para o desenvolvimento de sistemas de apoio à decisão médica, contribuindo para melhorias no tratamento e monitoramento de pacientes com câncer.

5.5.2 Integração de dados

A integração do resultado dos experimentos de REN e ER com a ontologia permite estruturar e relacionar os dados extraídos dos textos não estruturados do SGO, deixando-os disponíveis para utilização em componentes do SGO.

Durante o processo de integração, as entidades e relações identificadas pelo modelo de inteligência artificial são mapeadas e cadastradas na ontologia. Os dados extraídos, como entidades e suas relações, são representados como indivíduos (*Individuals*) e associados às suas respectivas classes e relacionamentos na ontologia.

A integração com a ontologia foi realizada utilizando o software Protégé e bibliotecas desenvolvidas neste trabalho. Essas ferramentas auxiliam no cadastro e manipulação dos dados na ontologia, permitindo que as entidades e relações extraídas sejam registradas adequadamente. Foram realizados testes de inserção manual e automática.

O modelo utiliza o experimento com o DGO-E para extrair as entidades e relações dos exames, dados estes que em seguida ao processo de extração encontram-se disponíveis para o cadastro na ontologia. Por padrão, é criada uma instância da ontologia com as informações da evolução, como o ID GEMED do paciente (não extraído neste experimento devido à anonimização), a data da evolução e os exames extraídos por meio da relação ”Apresenta”. Os exames extraídos e suas relações são inseridos em uma nova instância. Por padrão, os nomes das instâncias são compostos pelo nome do exame ou evolução, a data, a hora e o ID do paciente no GEMED (ignorado por enquanto devido à anonimização). Optou-se por testar a integração com o DGO-CD em um trabalho futuro, uma vez que o mesmo não possui as relações anotadas.

A integração automática da ontologia com o modelo foi desenvolvida em Python 3.10 utilizando a biblioteca *Owlready2*[8]. O *Owlready2* é uma biblioteca que permite a programação orientada a ontologia em Python, fornecendo um acesso transparente a ontologias OWL. Ele inclui um *quadstore* (um tipo de banco de dados otimizado para armazenar e consultar dados no formato RDF (*Resource Description Framework)*, adequada para o armazenamento e consulta de grandes quantidades de dados) otimizado baseado no *SQLite3* (uma biblioteca de banco de dados relacional embutida em código aberto), que oferece bom desempenho e consumo eficiente de memória, permitindo lidar com ontologias grandes. Além disso, o *Owlready2* possui integração com o UMLS (*Unified Medical Language System*) e terminologias médicas por meio

[8]https://owlready2.readthedocs.io/en/latest

do submódulo *PyMedTermino2*.

O algoritmo também é integrado com o raciocinador *HermiT* (HE et al., 2023), que é um raciocinador lógico desenvolvido para processar ontologias OWL e realizar inferências sobre os conceitos e relações definidos nelas. Ele é utilizado para suportar o raciocínio automatizado, aplicando lógica de descrição para representar conhecimento e realizar inferências na ontologia. Ele segue as especificações da linguagem de descrição de ontologias OWL e é capaz de realizar tarefas como classificação de instâncias, inferência de propriedades, detecção de inconsistências e inferência de equivalência.

Dessa forma, a integração do modelo com a ontologia possibilita a utilização de conhecimento ontológico para enriquecer as análises e inferências realizadas pelo modelo. Além disso, o uso da ontologia estruturada e das capacidades de raciocínio automatizado oferecidas pelo *HermiT* auxilia na interpretação semântica do conhecimento representado nas ontologias e facilita a manipulação dos dados gerados pelo modelo de forma coerente e estruturada.

Após a obtenção das saídas do modelo, a ontologia OWL é carregada no algoritmo. Em seguida, ocorre o mapeamento das saídas para os conceitos e propriedades definidos na ontologia. Isso envolve a criação de instâncias e o estabelecimento de relações entre elas. Através da *Owlready2*, é possível acessar e consultar os conceitos, propriedades e relações da ontologia, permitindo que os dados do modelo sejam inseridos no formato estruturado definido pela ontologia.

O quadro 4 apresenta a visão geral do algoritmo de integração proposto, do qual chama a função IntegracaoExame.

Quadro 4 – Visão geral do algoritmo de integração de dados proposto

```
Algoritmo IntegracaoExames()
    VAR exames[exame, data, resultado, membro, tempo, relacao] : Vetor
    VAR id_paciente: Inteiro
    Inicio
        ontologia = abrir_ontologia()
        VAR nome_exame, nome_membro, valor_resultado, data_exame, tempo_exame : String
        VAR data_extracao = data_hora_atual() : Data
         Individual ontologia.evolucao = criarEvolucao(data_extracao, id_paciente)
                Label ontologia.evolucao.nome = "evolução_"+ data_extracao + id_paciente
                Types ontologia.evolucao.Class = Evolução
                Relationships ontologia.evolucao.data = data_extracao
                Relationships ontologia.evolucao.paciente_id = id_paciente
        Para cada exame em exames[], faça
            Individual ontologia.novoexame = criarExame(exame, data, resultado, membro,
tempo, relacao, id_paciente, data_extracao)
                nome_exame = exame
                nome_membro = nome
                valor_resultado = resultado
                data_exame = data
                tempo_exame = tempo
                Label novoexame.nome = nome_exame + data_extracao + id_paciente
                Types novoexame.Class = "nome do exame"
                Relationships ontologia.evolucao.apresenta = novoexame.nome
                   Se entidade nome_membro == membro E relacao exame == verdadeiro faça
                       Relationships ontologia.novoexame.local_corpo = membro
                    Se entidade valor_resultado == resultado E relacao exame == verdadeiro faça
                       Relationships ontologia.novoexame.resultado = valor_resultado
                    Se entidade data_exame == data OU tempo_exame == tempo E relacao exame
== verdadeiro faça
                       Relationships ontologia.novoexame.data = data_exame
        Fim Para
        salvar_ontologia()
    Fim
```

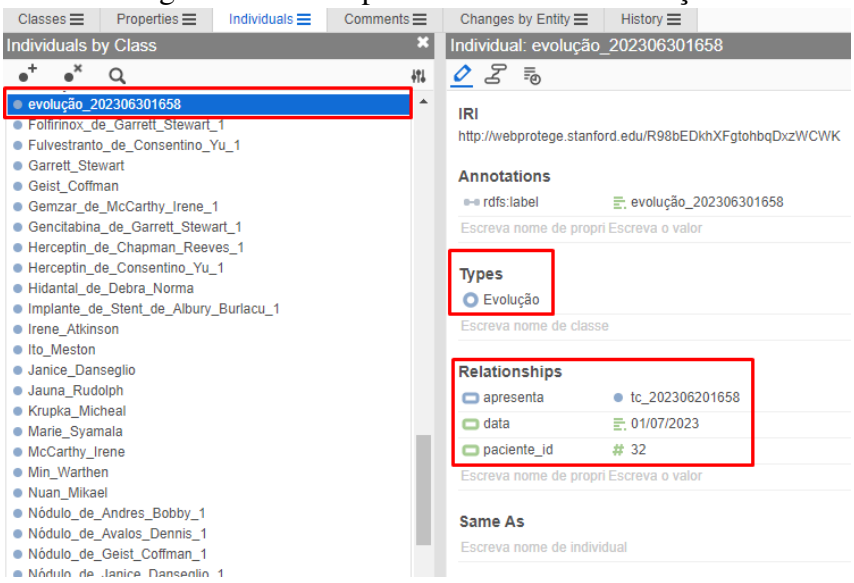
Para exemplificar este processo, pode-se considerar a evolução (extraída do SGO) descrita no Quadro 5 em LN, extraída de uma clínica oncológica. Neste quadro temos os itens comentados a seguir. O item ”Patient” corresponde ao Termo substituido pelo nome do paciente no processo de anonimização. O Termo ”Há 23 anos” descreve uma entidade de tempo, sem nenhuma relação com exame. O termo ”Há 5 meses” também é uma entidade de tempo, sem nenhuma relação com exame. O item ”TC” corresponde a uma entidade de Exame, que se refere a sigla para Tomografia Computacional. O item ”Massa” é uma entidade de Resultado, com relação ao exame TC. Por fim, o termo ”65cm” é uma entidade de Resultado, com relação ao exame TC.

Quadro 5 – Exemplo de evolução oncológica

Ca ”Patient” mama operado há 23 anos há 5 meses com mal estar tc ”Patient” torax massa ”Patient” 65 cm

Após a integração com a ontologia, a Figura 41 ilustra um exemplo do resultado da criação da evolução mencionada anteriormente. Foi criado um novo indivíduo chamado ”evolução_202306301650”, classificado como um tipo ”Evolução”. Suas relações incluem a data (data da evolução), o paciente_id (do SGO) e o exame extraído da evolução (TC). O exame (TC) foi inserido na ontologia como um novo indivíduo, descrito nesta figura com o identificador ”tc_202306201658” e detalhado conforme ilustrado na Figura 42.

Figura 41 – Exemplo de instância de evolução

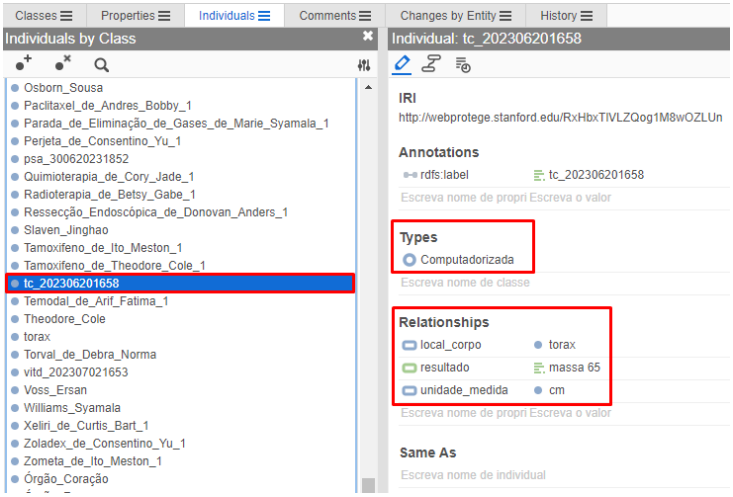


Fonte: Elaborado pelo autor

De acordo com a Figura 42, o processo de integração de dados cria um novo indivíduo chamado ”tc_202306201659”, visto que ”tc” é o nome do exame identificado na extração. O algoritmo marca o indivíduo como tipo ”Computadorizada” (subclasse de Tomografia). Suas relações incluem as demais entidades que foram extraídas do texto e que possuem relação com

o exame ”tc”: ”local_corpo” com o valor ”tórax”, ”resultado” com o valor ”massa 65” e ”unidade_medida” com ”cm” (unidade de medida do resultado).

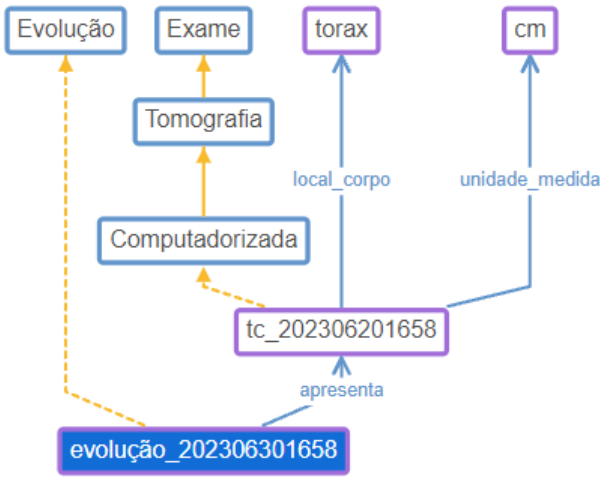
Figura 42 – Exemplo de instância de exame



Fonte: Elaborado pelo autor

A Figura 43 ilustra o relacionamento deste exemplo no formato de grafo. O fluxo de relações descreve a associação entre uma evolução específica e um exame específico de tomografia computadorizada na área da oncologia, conforme extraído do Quadro 5.

Figura 43 – Fluxo das relações



Fonte: Elaborado pelo autor

Um dos principais benefícios da integração do modelo com a ontologia é o rastreamento detalhado das evoluções ao longo do tempo. Com a incorporação da ontologia, é possível registrar e acompanhar de forma sistemática a progressão do estado de saúde dos pacientes com câncer. Isso proporciona aos profissionais de saúde uma visão histórica completa da condição do paciente, permitindo uma análise mais precisa e informada das mudanças no quadro clínico ao longo do tratamento. A hierarquia de exames definida na ontologia permite uma classificação mais clara e compreensível dos diversos exames utilizados no diagnóstico e tratamento

do câncer, facilitando a identificação de padrões e correlações entre os resultados de diferentes testes. Além disso, com a estrutura ontológica bem definida, é possível integrar informações de diferentes fontes e domínios, permitindo uma análise mais abrangente dos dados de evoluções oncológicas. Isso favorece o desenvolvimento de tratamentos personalizados e a identificação de tendências e padrões que podem melhorar os resultados dos pacientes.

5.5.3 Avaliação dos resultados

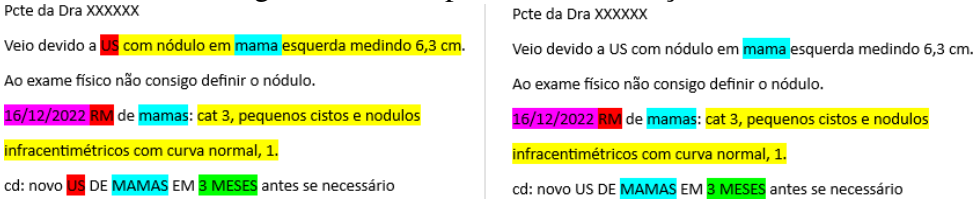
Nesta seção, são apresentadas as etapas de avaliação realizadas para o experimento do DGO-E com a ontologia. O estudo foi conduzido em um estudo de cenários com profissionais da saúde, incluindo biomédicos, enfermeiros especialistas em saúde e estudantes de medicina. Após uma preparação adequada, foram realizadas duas avaliações distintas. A primeira avaliação teve como objetivo calcular o nível de acurácia do modelo em relação ao julgamento dos profissionais, conforme descrito na seção 5.5.3.1. A segunda avaliação visou compreender a aceitação dos especialistas em relação à integração dos dados não estruturados em uma ontologia, e os detalhes dessa avaliação estão apresentados na seção 5.5.3.2.

5.5.3.1 Avaliação da integração

Para a avaliação de integração, foram exportadas 22 evoluções aleatoriamente do SGO. Um profissional da área de saúde foi encarregado de avaliar os textos e realizar manualmente a marcação das entidades presentes em cada um dos documentos. Posteriormente, essas mesmas evoluções foram inseridas no modelo, permitindo a comparação do nível de acerto do modelo em relação ao especialista.

A Figura 44 ilustra um exemplo comparativo da evolução 12, onde são apresentados os resultados da marcação realizada pelo profissional da saúde e as entidades reconhecidas pelo modelo. No lado esquerdo da figura, encontram-se as marcações feitas pelo profissional da saúde, enquanto no lado direito, são indicadas as entidades identificadas pelo modelo. As entidades são identificadas por meio de tarjas coloridas: o vermelho representa os ”exames”, o amarelo os ”resultados”, o azul os ”membros”, o rosa as ”datas” e o verde os ”tempos”.

Figura 44 – Comparativo da evolução 12



Fonte: Elaborado pelo autor

Conforme ilustrado na Figura 44, o modelo apresentou um desempenho de 70% na identi-

ficação das entidades marcadas pelo profissional da saúde, com notáveis acertos nas categorias de Membro, Tempo e Data. No entanto, o modelo não conseguiu identificar corretamente os exames de ”US” (Ultrassom), assim como o resultado ”nódulo em mama esquerda medindo 6,3cm.”. A tabela 31 apresenta o resultado comparativo total entre as marcações realizadas pelo profissional da saúde e as efetuadas pelo modelo. A primeira coluna indica o número de cada evolução, cujos detalhes podem ser encontrados no Apêndice D deste trabalho. Nas colunas ”exame”, ”resultado”, ”tempo”, ”membro” e ”data”, os valores à esquerda representam a quantidade de marcações realizadas pelo profissional da saúde, enquanto os valores à direita indicam a quantidade de marcações feitas pelo modelo. Adicionalmente, é apresentada uma média de acurácia das marcações realizadas no documento indicado em cada linha.

Tabela 31 – Comparativo de acertividade do modelo

	Exame		**Resultado**		**Tempo**		**Membro**		**Data**		**Acurácia**
1	5	3	15	4	0	0	0	0	18	18	65,79%
2	30	21	29	9	0	0	12	12	14	14	65,88%
3	15	9	13	4	0	0	5	3	8	8	58,54%
4	2	2	2	0	0	0	1	1	1	1	66,67%
5	6	3	6	4	0	0	2	2	6	5	70,00%
6	5	2	5	2	0	0	0	0	8	7	61,11%
7	2	0	2	1	0	0	3	1	2	2	44,44%
8	2	0	6	3	0	0	2	2	14	14	79,17%
9	28	17	28	8	0	0	12	11	16	16	61,90%
10	2	1	2	0	0	0	1	0	7	6	58,33%
11	4	2	3	1	0	0	1	1	5	2	46,15%
12	3	1	2	1	1	1	3	3	1	1	70,00%
13	2	1	6	3	0	0	2	1	3	2	53,85%
14	3	0	3	1	0	0	1	1	10	9	64,71%
15	21	9	20	6	0	0	7	7	15	15	58,73%
16	2	2	3	1	0	0	1	0	8	8	78,57%
17	5	2	5	3	0	0	3	3	3	3	68,75%
18	5	3	5	1	0	0	5	4	4	4	63,16%
19	2	1	2	0	0	0	1	1	1	1	50,00%
20	5	2	5	1	0	0	0	0	7	6	52,94%
21	16	10	16	5	0	0	7	5	6	6	57,78%
22	2	2	2	2	0	0	1	1	2	2	100,00%
Média		55,68%		33,33%		100%		84,28%		94,33%	**73,52%**

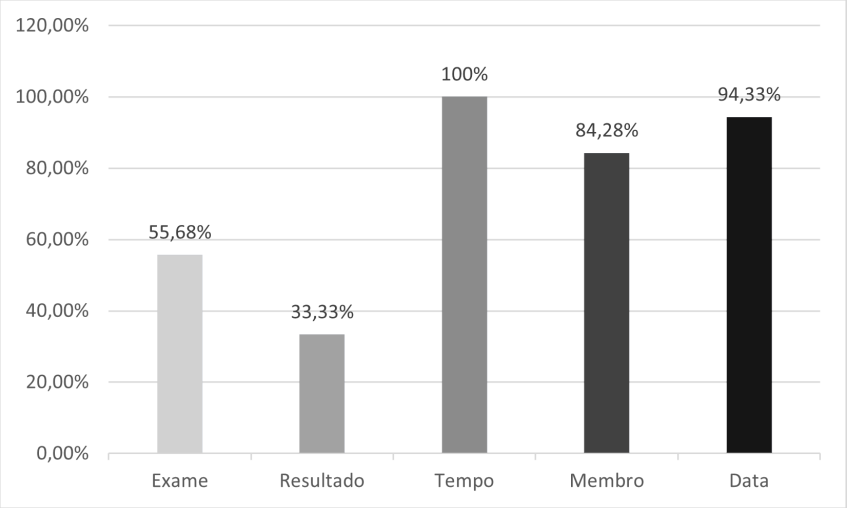
Fonte: Elaborado pelo autor.

O modelo demonstrou uma média geral de acertos de 73,52%, indicando um desempenho razoável em relação às marcações realizadas pelo profissional de saúde. Evoluções específicas também foram examinadas, revelando casos em que o modelo obteve resultados discrepantes em comparação com as marcações realizadas pelos profissionais de saúde. Ao analisar as categorias específicas, observa-se que o modelo teve a menor acurácia na marcação dos ”Resultados” (33,33%), indicando dificuldades em reconhecer e extrair informações detalhadas de

exames, que geralmente são complexos, expressos em termos técnicos e abreviações específicas. Por exemplo, na evolução 9, o modelo acertou apenas 8 das 28 marcações de ”Resultado” realizadas pelo profissional de saúde, indicando uma dificuldade particular nesse contexto. Os resultados identificados pelo médico muitas vezes decorrem de uma longa frase do texto sem determinados padrões. Este é um ponto que pode ser melhor explorado em trabalhos futuros.

Por outro lado, obteve a maior acurácia na marcação do ”Tempo” (100%), porém foi marcado apenas 1 amostra dessa categoria. As marcações ”Membro” e ”Data” apresentaram uma acurácia mais próxima da média geral (84,28% e 94,33%, respectivamente), indicando uma maior consistência do modelo nessas categorias. Nos ”Exames” foram trabalhados com muitas siglas, mas mesmo assim o modelo teve um desempenho mediano, apresentando 55,68% de acertos. Uma evolução em que o modelo obteve 100% de acurácia em todas as categorias é a evolução 22. Entretanto, esse resultado pode ser atribuído ao fato de que essa evolução possui apenas 2 marcações em cada categoria, tornando-a um caso mais simples para o modelo lidar. Por fim, a Figura 45 ilustra um comparativo por categoria.

Figura 45 – Comparativo de acertividade por categoria



Fonte: Elaborado pelo autor

Outro aspecto importante levantado na análise é a relação entre a complexidade das informações presentes nas evoluções e o desempenho do modelo. Evoluções escritas em formato de texto corrido apresentaram melhores resultados, enquanto aquelas extraídas de tabelas ou com dados sem continuidade tiveram desempenho inferior (O quadro 6 apresenta um exemplo desse problema), o que indica possíveis desafios para o modelo em relação à estrutura e continuidade dos dados.

Quadro 6 – Exemplo da extração de tabelas

06/10/2017 30/08/2018 12/2018 07/2019 02/2020 10/12/2020 22/04/2021 13/09/2021 24/01/2022 04/2022 30/06/2022 12/2022
PSAT 1,26 » RT » 1,00 > 0,21 0,13 0,05 0,05 0,08 0,12 0,23 0,34 0,47 0,45 0,66

Após a análise de acertividade do modelo, as evoluções foram integradas na ontologia. A seguir é exemplificado este processo com um exemplo específico. A evolução descrita no Quadro 7 apresenta um dos exemplos inseridos.

Quadro 7 – Exemplo de evolução oncológica

adenocarcinoma de reto medio tratado com cirurgia 26/12/2018 ap adenocarcinoma tubular g2 ial epn - linf 0/11 xelox 7x ate 01/07/2019 dpoc d alzheimer- donila duo 10/20 sem sintomas cea 252 vit d 249 seguimento oncologico ad til.

Após o processamento destes dados pelo modelo, as entidades presentes no exemplo do Quadro 7 são extraídas. A Figura 46 apresenta o resultado da extração. Nela podem ser observadas as entidades ”reto” (membro), ”26/12/2018” (data), ”01/07/2019” (data), ”cea” (exame), ”252” (resultado), ”vid d” (exame), ”249” (resultado) e as demais (*other*), das quais foram reconhecidas e extraídas pelo modelo. Em seguida, as relações são identificadas e enviadas para a ontologia.

Figura 46 – Resultado da frase inserida no modelo

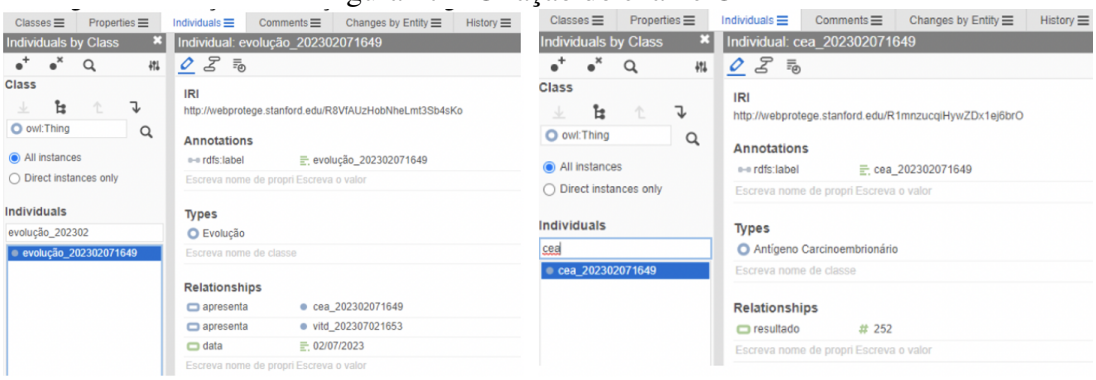
```
Entidades encontradas: [('adenocarcinoma', 'OTHER'), ('de', 'OTHER'), ('reto', 'MEMBRO'), ('medio', 'OTHER'), ('tratado', 'O
THER'), ('com', 'OTHER'), ('cirurgia', 'OTHER'), ('26/12/2018', 'DATA'), ('ap', 'OTHER'), ('adenocarcinoma', 'OTHER'), ('tub
ular', 'OTHER'), ('g2', 'OTHER'), ('ial', 'OTHER'), ('epn', 'OTHER'), ('-', 'OTHER'), ('linf', 'OTHER'), ('0/11', 'OTHER'),
('xelox', 'OTHER'), ('7x', 'OTHER'), ('ate', 'OTHER'), ('01/07/2019', 'DATA'), ('dpoc', 'OTHER'), ('d', 'OTHER'), ('alzheime
'), ('01/07/2019', 'DATA'), ('dpoc', 'OTHER'), ('d', 'OTHER'), ('alzheimer-', 'OTHER'), ('donila', 'OTHER'), ('duo', 'OTHER'
), ('10/20', 'OTHER'), ('sem', 'OTHER'), ('sintomas', 'OTHER'), ('cea', 'EXAME'), ('252', 'RESULTADO'), ('vit d', 'EXAME'),
('249', 'RESULTADO'), ('seguimento', 'OTHER'), ('oncologico', 'OTHER'), ('ad', 'OTHER'), ('til', 'OTHER')]
```

Fonte: Elaborado pelo autor

A Figura 47 ilustra a evolução criada na ontologia, de acordo com os métodos implementados. O item da esquerda ilustra a criação de um indivíduo de evolução na ontologia, classificado como ”Evolução”, que inclui os exames ”CEA” e ”VIT D”, juntamente com a data da evolução. Além disso, foram criados dois indivíduos separados para cada exame extraído do texto, bem como para os respectivos resultados. A Figura 47 ilustra, no item da direita, um exemplo da criação do exame ”CEA”, identificado como ”Antígeno Carcinoembrionário”. Por fim, a Figura 48 apresenta as relações da evolução extraída em formato de grafo.

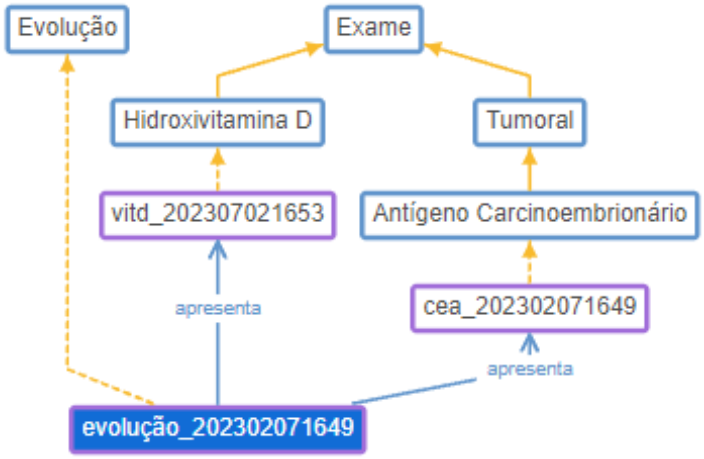
A ontologia permite uma estruturação mais precisa e semântica dos dados extraídos, além de possibilitar a criação de relacionamentos entre as entidades, enriquecendo a análise e permitindo consultas avançadas. Com a integração, foi possível obter uma visão mais completa e estruturada dos dados não estruturados, facilitando a recuperação e o entendimento dessas informações.

Figura 47 – Criação do exame CEA



Fonte: Elaborado pelo autor

Figura 48 – Fluxo das relações da evolução extraída

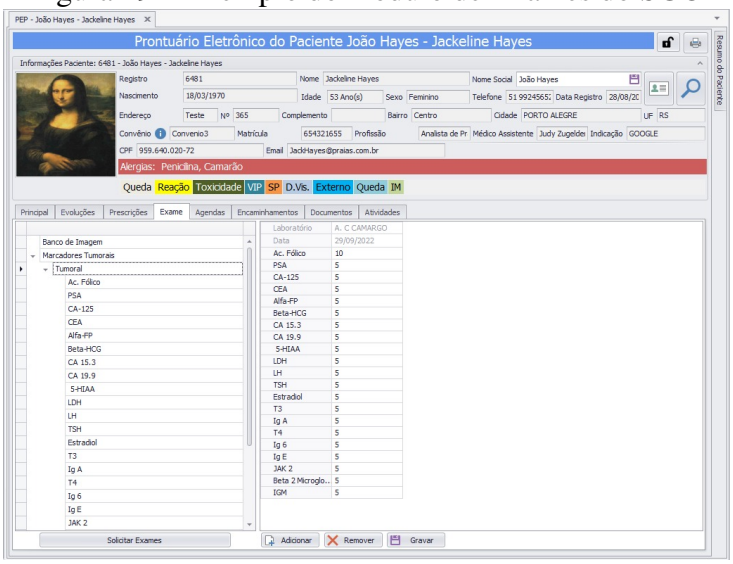


Fonte: Elaborado pelo autor

Uma proposta preliminar para a implementação de um protótipo, com base em reuniões de alinhamento com a Interprocess, consiste em automatizar o processo de estruturação das informações extraídas dos exames escritos pelos médicos durante a evolução do paciente, a fim de preencher automaticamente o Sistema de Gestão de Saúde (SGO). Atualmente, esse processo é realizado manualmente pelos profissionais de saúde. A Figura 49 apresenta um exemplo do módulo de exames do SGO. Esse módulo faz parte do EHR do paciente, que contém também informações principais, evoluções, prescrições, agendas, encaminhamentos, documentos, atividades e dados pessoais. No módulo de exame, o SGO apresenta uma lista dos exames feitos pelo paciente, bem como seus resultados, o laboratório e a data do exame.

A proposta consiste em utilizar um modelo de extração de informações para analisar os exames médicos registrados e, com base nesses dados, estruturar as informações relevantes. Dessa forma, os exames serão apresentados em uma lista organizada. Isso eliminará a necessidade de preenchimento manual e possibilitará uma visualização mais clara e acessível para os profissionais de saúde. Essa abordagem visa otimizar o fluxo de trabalho, economizar tempo dos profissionais de saúde e reduzir possíveis erros de preenchimento.

Figura 49 – Exemplo do módulo de Exames do SGO



Fonte: Elaborado pelo autor

É importante destacar que o protótipo integrado com o SGO será uma primeira versão, sujeita a ajustes e refinamentos com base no feedback e nas necessidades identificadas durante sua utilização. A parceria com a Interprocess permitirá que a implementação seja realizada de forma colaborativa. Com a adoção desse protótipo, a expectativa é de que o SGO se torne ainda mais completo e eficaz, contribuindo para aprimorar o atendimento aos pacientes e o gerenciamento das informações clínicas de forma segura e integrada.

5.5.3.2 Avaliação da aceitação pelos profissionais

A ontologia foi avaliada com base em um modelo de cenários. Ao descrever o modelo de cenários, um avaliador descreve sistematicamente os cenários de interação do usuário com um sistema e avalia as ações necessárias para completar cada um dos cenários executados pelo usuário com um modelo cognitivo (SUGIMURA; ISHIGAKI, 2005). Devido aos objetivos descritos em termos de cenários e levando em conta as respostas obtidas, o avaliador pode efetivamente identificar questões críticas que podem prejudicar o alcance dos objetivos, bem como identificar aspectos de aceitação do uso dos recursos no sistema.

O questionário foi respondido pelos cinco especialistas da área da saúde citados anteriormente. Antes das avaliações, esses profissionais receberam uma apresentação detalhada sobre o sistema desenvolvido, uma explicação sobre ontologias e exemplos de execução e resultados do modelo. A avaliação de aceitação buscou compreender a aceitação dos especialistas quanto à integração dos dados não estruturados em uma ontologia. Para isso, eles responderam a um questionário especialmente elaborado para colher evidências necessárias. Os resultados do questionário e as respostas obtidas estão apresentados de forma clara na Tabela 32. As colunas desta tabela utilizam siglas, sendo seu significados os seguintes: DT para Discordo Totalmente,

D para Discordo, NDC para Nem Concordo nem Discordo, C para Concordo, CT para Concordo Totalmente. Quando as perguntas citam o modelo e seu uso, os respondentes receberam a informação de que se trata do modelo mais geral de extração de dados não estruturados e sua integração na ontologia.

Tabela 32 – Resultados das Perguntas

Perguntas	**DT**	**D**	**NDC**	**C**	**CT**
Conheço algum software que realiza este tipo de tarefa.	60%	40%	20%	0%	0%
O uso do modelo pode facilitar a formalização do conhecimento da área da saúde.	0%	0%	0%	0%	100%
O uso do modelo pode viabilizar o desenvolvimento de novas aplicações na área de saúde.	0%	0%	0%	0%	100%
O modelo proposto é útil para a área da saúde.	0%	0%	0%	20%	80%
O uso da ontologia como base para estruturação de dados pode simplificar o entendimento e a visualização das informações da área da saúde.	0%	0%	0%	0%	100%
Se necessitar de informações estruturadas, usaria o modelo.	0%	0%	0%	20%	80%
É fácil de entender a estruturação dos dados com o modelo.	0%	0%	0%	20%	80%

Fonte: Elaborado pelo autor

Conforme a Tabela 32, é possível afirmar que 60% dos participantes discordam totalmente quando questionados sobre o conhecimento de algum software capaz de realizar a tarefa proposta, o que indica que a maioria não conhece um softwares que realiza a tarefa proposta. Isso aponta para uma falta de conhecimento e adoção nessa área específica, o que pode representar uma oportunidade para a pesquisa. Os 40% restantes demonstraram discordância, o que sugere a existência de um pequeno grupo de participantes com algum conhecimento prévio sobre esse tipo de tecnologia.

Quanto à possibilidade do modelo facilitar a formalização do conhecimento na área da saúde, 100% dos participantes concordam totalmente. Isso indica que há um reconhecimento significativo entre os participantes de que o uso do modelo pode contribuir para a formalização do conhecimento nesse campo específico da saúde. Essa percepção positiva sugere um potencial interesse e aceitação do modelo para essa finalidade.

Em relação à capacidade do modelo de viabilizar o desenvolvimento de novas aplicações na área da saúde, 100% dos participantes concordam totalmente. Essa alta concordância reflete um forte consenso de que o modelo proposto pode ser uma ferramenta viável para o desenvolvimento de novas aplicações na área da saúde

Na avaliação da utilidade do modelo, novamente 20% dos participantes concordam que o modelo proposto é útil para a área da saúde, enquanto 80% concordam totalmente. Essa alta concordância indica que a grande maioria dos participantes considera o modelo útil para a área

da saúde, reforçando a relevância do modelo proposto na visão dos especialistas.

No que diz respeito ao uso da ontologia como base para a estruturação de dados e sua capacidade de simplificar o entendimento e visualização das informações na área da saúde, 100% dos participantes concordam totalmente. Esses resultados mostra o potencial da ontologia como base para simplificar o entendimento e visualização de informações na área da saúde, tornando-a um tópico relevante a ser explorado.

Quando questionados se utilizariam o modelo caso necessitassem de informações estruturadas, 20% dos participantes concordam, enquanto 80% concordam totalmente. Isso indica que uma parcela substancial dos participantes estaria disposta a utilizar o modelo em caso de necessidade de informações estruturadas, o que demonstra um interesse considerável na aplicação do modelo.

Por fim, na última pergunta sobre a facilidade de entender a estruturação dos dados com o modelo, 20% dos participantes concordam, enquanto 80% concordam totalmente. Isso sugere que a maioria dos participantes considera a estruturação dos dados com o modelo como algo de fácil compreensão, o que é útil para um uso diário do software.

Em resumo, os resultados mostram um reconhecimento positivo do potencial do modelo proposto, principalmente em relação à facilidade de entendimento, utilidade na área de saúde, viabilidade para o desenvolvimento de aplicações e simplificação da estruturação dos dados. Esses *insights* podem auxiliar na implantação do modelo no SGO, destacando pontos de discussão relevantes.

6 CONCLUSÃO

Este estudo apresentou um novo modelo para REN e ER de dados não estruturados para população de uma ontologia. O estudo foi conduzido com a revisão bibliográfica dos formalismos e trabalhos relacionados, a partir dos quais os objetivos e a questão de pesquisa foram formulados, bem como a estrutura e contexto geral do modelo. O estudo de caso e os experimentos implementaram arquiteturas baseadas na abordagem de *Transformers* e com o modelo de linguagem BERT.

Este modelo foi implementado como caso de estudo no SGO e seus resultados serão usados para auxílio na área da saúde oncológica. O modelo foi avaliado por especialistas na área da saúde e também a nível computacional. A parceria com a Interprocess junto ao programa Doutorado Acadêmico em Inovação, além da oportunidade de gerar inovação no setor empresarial, possibilita uma aproximação em relação ao conhecimento técnico de especialistas em saúde, fazendo com que a aproximação com especialistas da área da saúde e da área de informática médica torne-se um diferencial para essa pesquisa.

Para a realização do trabalho foram efetuados estudos do estado da arte e o estudo do embasamento teórico, bem como o aprofundamento nos requisitos do estudo de caso envolvendo o SGO. Com base neste contexto, foram realizados experimentos com técnicas de REN e ER em textos médicos e a população de uma ontologia com os dados estruturados. O modelo desenvolvido alcançou resultados de 78,24% de precisão no domínio de exames e 72,87% no domínio de diagnósticos no REN e ER. Além disso, uma ontologia com foco em oncologia foi construída e integrada ao modelo, englobando aproximadamente 181 classes, 14 propriedades de dados, 12 propriedades de objetos e mais de 200 indivíduos. A avaliação com especialistas da área de saúde obteve uma taxa de acerto de 73,52% em relação a análise deles, e a pesquisa de usabilidade mostrou uma excelente aceitação.

Desta forma, a questão de pesquisa foi devidamente abordada e respondida no decorrer deste trabalho. O modelo desenvolvido alcançou o objetivo de permitir a extração de relações de dados não estruturados, com a finalidade de estruturá-los e gerar uma base de conhecimento para uso na área da saúde. O modelo proposto combina técnicas de PLN, DL e ontologias, proporcionando uma abordagem eficaz para o REN e a ER em dados não estruturados provenientes de um sistema EHR.

No que diz respeito aos objetivos específicos, cada um deles foi alcançado com sucesso:

- Realizou-se estudos aprofundados sobre sistemas médicos, PLN, DL e ontologias, permitindo adquirir o conhecimento necessário para o desenvolvimento do modelo proposto;

- Realizou-se uma avaliação detalhada dos trabalhos correlatos disponíveis na literatura, analisando as diferentes abordagens utilizadas para o reconhecimento e resolução de entidades em dados não estruturados, o que auxiliou na definição e aprimoramento do modelo;

- Realizou-se experimentos com dados na área da saúde, trabalhando com um conjunto de dados em português relacionado à oncologia. Esses experimentos foram fundamentais para a validação e avaliação do modelo.
- Anonimizou-se e anotou-se o conjunto de dados em português, seguindo as diretrizes de ética e privacidade. Essa etapa foi essencial para garantir a confidencialidade das informações dos pacientes e permitir a utilização dos dados no desenvolvimento do modelo.
- Desenvolveu-se um modelo detalhado para a realização do REN e ER, levando em consideração as características específicas dos dados não estruturados da área da saúde. Esse modelo foi construído com base em práticas e técnicas identificadas durante os estudos.
- Desenvolveu-se um protótipo funcional para avaliação do modelo proposto. O protótipo foi implementado e permitiu a demonstração prática da eficácia do modelo na extração e estruturação de dados não estruturados.
- Avaliou-se os resultados dos experimentos com especialistas da área da saúde, obtendo feedbacks valiosos sobre a precisão e relevância das relações extraídas pelo modelo. Essa validação com especialistas contribuiu para a confirmação da eficácia e aplicabilidade do modelo proposto na prática clínica.

Em resumo, o modelo desenvolvido neste trabalho alcançou plenamente os objetivos propostos, fornecendo uma abordagem abrangente e eficiente para o REN e ER não estruturados na área da saúde. Os resultados obtidos são promissores e abrem caminho para aplicações futuras que possam beneficiar a área médica, fornecendo uma base de conhecimento estruturada e confiável para auxiliar profissionais de saúde e pesquisadores.

Através dos resultados, pode-se observar o potencial e a importância deste trabalho para a área da saúde, bem como identificar que há uma gama de possibilidades a serem trabalhadas para uma extração eficaz de relações e aspectos a avaliar sobre a forma de estruturar esses dados recebidos em LN.

6.1 Contribuições

Este estudo possui contribuições em dois aspectos interligados. Primeiramente, destaca-se a sua aplicabilidade em sistemas EHR voltados para oncologia, por meio da ampliação da capacidade desses sistemas de utilizarem dados não estruturados.

Em segundo lugar, destaca-se a experimentação e a proposta de avanços na computação aplicada a RNs de REN e ER. O modelo traz contribuições quanto à avaliação do desempenho do BERT e BiLSTM para REN e ER em textos oncológicos em português no contexto de EHRs explorando diferentes modelos de BERT, incluindo o ”BERT-base-multilingual-cased”, ”BioBERT-PT” e ”BERTimbau”. É importante destacar o *fine-tuning* realizado com os *datasets* DGO-E e DGO-CD, que desempenhou um papel fundamental na adaptação e melhoria do

desempenho dos modelos BERT e BiLSTM para as tarefas específicas de REN e ER, permitindo que os modelos fossem ajustados e calibrados para lidar com as características e nuances presentes nos textos reais relacionados à área oncológica em português. Esse processo contribuiu para melhorar a precisão, resultando em um desempenho mais eficiente e acurado nos experimentos realizados.

O presente trabalho foi realizado como um estudo de caso em uma empresa especializada em Oncologia. Foram realizadas análises detalhadas de um sistema amplamente adotado em clínicas oncológicas, resultando em uma contribuição na definição, implementação e testes de componentes que permitem a utilização de dados não estruturados do sistema para a construção de uma base de conhecimento.

Desenvolveu-se também um algoritmo para anonimização de dados. Este algoritmo foi utilizado na criação de três *datasets* com dados do SGO, sendo eles ambientados com dados em português em texto livre e na área da oncologia. Outro diferencial deste trabalho é a criação de conjuntos de dados inéditos, contendo informações reais de entidades e relações de evolução médica, com um total de 1.622 documentos anotados, compreendendo 146.769 entidades e 111.716 relações.

Adicionalmente, destaca-se a análise e adaptação de uma ontologia de domínio para representar os dados estruturados provenientes desse estudo de caso, sendo esse mais um dos diferenciais do trabalho.

Por fim, foram conduzidos experimentos utilizando abordagens de REN e ER em texto, obtendo resultados promissores nos modelos utilizados. A distinção desse estudo reside no treinamento com dados reais de oncologia e na construção de uma base de conhecimento por meio da ontologia.

Quanto a publicações científicas, dois artigos foram publicados ao longo do período, sendo:

- Um artigo com os resultados experimentais de classificação (descrito na seção 5.2), do qual foi publicado no BRACIS[1] (*Brazilian Conference on Intelligent Systems*) sob o título "*Clinical oncology textual notes analysis using machine learning and deep learning*".

- Um artigo com os experimentos prévios de REN com o *dataset* DDI, publicado no SBCAS[2] (Simpósio Brasileiro de Computação Aplicada à Saúde) sob o título "Avaliação de modelos para extração de dados não estruturados de um sistema EHR para atender a estrutura final de uma ontologia".

Também foram submetidos artigos que estão aguardando revisão, sendo eles:

- Um artigo da revisão sistemática (descrito na seção 3), que foi enviado para o *Journal* JCBM[3] (*Computers in Biology and Medicine*) sob o título "*Deep Learning And Natural*

[1] https://www.bracis.dcc.ufmg.br/home
[2] http://www.midiacom.uff.br/sbcas2023
[3] https://www.sciencedirect.com/journal/computers-in-biology-and-medicine

Language Processing Applied To Health Data, Information Extraction, And Ontologies: A Systematic Literature Review";

- Um artigo com os experimento de REN com ER completos utilizando o DDI (descrito na seção 5.3) foi enviado ao *Journal* JBHI[4] (*IEEE Journal of biomedical and health informatics*) sob o título "*Building Ontologies from unstructured EHR data with Natural Language Processing*";

- Por fim, um artigo com os experimentos de REN e ER com o DGO-E e DGO-CD foi enviado ao JAI[5] (*Journal Applied Intelligence*) sob o título "*Named Entity Recognition and Relations Extraction from unstructured data of an EHR system to population of an ontology*".

6.2 Limitações do Trabalho

Embora este estudo tenha alcançado bons resultados, existem algumas limitações a serem consideradas. Uma delas é o tamanho do *dataset* utilizado. Embora o trabalho tenha criado conjuntos de dados inéditos com informações reais de entidades e relações médicas, o número total de documentos anotados foi de 1.622. Um *dataset* maior poderia fornecer mais exemplos para treinar os modelos e avaliar o sistema, aumentando sua capacidade de generalização e melhorando seu desempenho.

Além disso, a natureza dos dados reais e não estruturados extraídos de um software apresentou desafios adicionais. Os documentos continham erros de português, abreviações, siglas e terminologias específicas da área da saúde, o que pode dificultar o processamento e a extração de informações corretas. Essa variabilidade na qualidade e na estrutura dos dados pode afetar a precisão e a confiabilidade do modelo.

Outra limitação é a falta de diversidade nos tipos de documentos anotados. Embora o estudo tenha se concentrado em dados relacionados à oncologia, especificamente de exames e características de diagnósticos e protocolos, a inclusão de outros contextos médicos poderia enriquecer o modelo e torná-lo mais abrangente. A variação nos tipos de documentos, como prontuários médicos, relatórios de laboratório e históricos de pacientes, traria desafios adicionais, mas também refletiria melhor a realidade dos dados encontrados na prática clínica.

Além disso, é importante mencionar que o modelo proposto depende fortemente da disponibilidade e da qualidade dos dados anotados utilizados no treinamento. A anotação manual de dados requer tempo, especialistas na saúde e conhecimento do domínio. A obtenção de dados anotados em larga escala pode ser um desafio e a qualidade das anotações pode variar dependendo dos anotadores. Portanto, é necessário um esforço contínuo para melhorar a qualidade e a disponibilidade dos conjuntos de dados anotados para aprimorar o desempenho do sistema.

[4]https://www.embs.org/jbhi/
[5]https://www.springer.com/journal/10489

Por fim, é importante ressaltar que as limitações mencionadas não são exclusivas deste estudo. Trabalhar com dados reais e não estruturados é um desafio comum em muitas áreas de pesquisa em PLN e ML. A falta de padronização e a variabilidade dos dados podem impactar o desempenho e a generalização dos modelos desenvolvidos. No entanto, essas limitações também oferecem oportunidades para pesquisas futuras e melhorias contínuas na área.

6.3 Trabalhos futuros

Como trabalhos futuros, primeiramente, pretende-se expandir o DGO-E com novas anotações de entidades e relações para a área da oncologia. Com o aumento da quantidade de dados disponíveis, seria possível obter melhorias no desempenho do modelo. Outra possibilidade é o aumento do DGO-CD com a anotação de relações das entidades. Após a expansão dos *datasets*, seria possível realizar um novo treinamento na arquitetura BiLSTM utilizando os dados anotados. Essa abordagem poderia aprimorar a extração de informações e a compreensão das relações das evoluções do EHR.

Outra etapa importante prevista é a integração do modelo desenvolvido com o SGO. Essa integração permitirá utilizar o modelo em um ambiente real de EHR, fornecendo suporte automatizado na extração e estruturação dos dados não estruturadas, além de melhorar a sua qualidade gradativamente com os novos dados.

Por fim, considera-se interessante explorar a ontologia desenvolvida neste estudo de forma mais abrangente. Em vez de se concentrar apenas em entidades e relações específicas, poderiam ser extraídas informações da ontologia para gerar perguntas relacionadas à oncologia. Isso possibilitaria a criação de um sistema capaz de interagir com os usuários, fornecendo respostas e informações relevantes com base nos conceitos da ontologia.

REFERÊNCIAS

ABEYSINGHE, R. et al. Ssif: subsumption-based sub-term inference framework to audit gene ontology. **Bioinformatics**, [S.l.], v. 36, n. 10, p. 3207–3214, 2020.

ADEKKANATTU, P. et al. Comorbidity and healthcare utilization in patients with treatment resistant depression: a large-scale retrospective cohort analysis using electronic health records. **Journal of Affective Disorders**, [S.l.], v. 324, p. 102–113, 2023.

ADEL, E. et al. A unified fuzzy ontology for distributed electronic health record semantic interoperability. In: **U-healthcare monitoring systems**. [S.l.]: Elsevier, 2019. p. 353–395.

ADEL, E. et al. Ontology-based electronic health record semantic interoperability: a survey. In: **U-healthcare monitoring systems**. [S.l.]: Elsevier, 2019. p. 315–352.

ADLUNG, L. et al. Machine learning in clinical decision making. **Med**, [S.l.], 2021.

ALEMI, F. et al. Bayesian processing of context-dependent text: reasons for appointments can improve detection of influenza. **Medical Decision Making**, [S.l.], v. 32, n. 2, p. E1–E9, 2012.

ALFATTNI, G.; PEEK, N.; NENADIC, G. Extraction of temporal relations from clinical free text: a systematic review of current approaches. **Journal of Biomedical Informatics**, [S.l.], p. 103488, 2020.

AMATO, F. et al. A semantic system for diagnoses suggestion and clinical record management. In: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION NETWORKING AND APPLICATIONS WORKSHOPS (WAINA), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 133–138.

ANDREA, B.; FRANCO, T. Mining bayesian networks out of ontologies. **Journal of Intelligent Information Systems**, [S.l.], v. 38, n. 2, p. 507–532, 2012.

ARSENE, O.; DUMITRACHE, I.; MIHU, I. Medicine expert system dynamic bayesian network and ontology based. **Expert Systems with Applications**, [S.l.], v. 38, n. 12, p. 15253–15261, 2011.

BADAR, M.; HARIS, M.; FATIMA, A. Application of deep learning for retinal image analysis: a review. **Computer Science Review**, [S.l.], v. 35, p. 100203, 2020.

BALZANO, W.; DEL SORBO, M. R. Setra: a smart framework for gps trajectories' segmentation. In: INTERNATIONAL CONFERENCE ON INTELLIGENT NETWORKING AND COLLABORATIVE SYSTEMS, 2014., 2014. **Anais...** [S.l.: s.n.], 2014. p. 362–368.

BANDY, J.; VINCENT, N. Addressing"documentation debt"in machine learning research: a retrospective datasheet for bookcorpus. **arXiv preprint arXiv:2105.05241**, [S.l.], 2021.

BECKER, M.; BÖCKMANN, B. Personalized guideline-based treatment recommendations using natural language processing techniques. In: **Informatics for health**: connected citizen-led wellness and population health. [S.l.]: IOS Press, 2017. p. 271–275.

BEG, S. et al. Wearable smart devices in cancer diagnosis and remote clinical trial monitoring: transforming the healthcare applications. **Drug discovery today**, [S.l.], 2022.

BENÍCIO, D. H. P. **Aplicação de mineração de texto e processamento de linguagem natural em prontuários eletrônicos de pacientes para extração e transformação de texto em dado estruturado**. 2020. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Rio Grande do Norte, 2020.

BERGMANN, U. **Evolução de cenários através de um mecanismo de rastreamento baseado em transformações**. 2003. Tese (Doutorado em Ciência da Computação) — PhD Thesis, 2003.

BERTSIMAS, D.; WIBERG, H. Machine learning in oncology: methods, applications, and challenges. **JCO Clinical Cancer Informatics**, [S.l.], v. 4, p. 885–894, 2020.

BHASALE, A. The wrong diagnosis: identifying causes of potentially adverse events in general practice using incident monitoring. **Family Practice**, [S.l.], v. 15, n. 4, p. 308–318, 1998.

BOSSELUT, A. et al. Comet: commonsense transformers for automatic knowledge graph construction. **arXiv preprint arXiv:1906.05317**, [S.l.], 2019.

BOUKKOURI, H. E. et al. Characterbert: reconciling elmo and bert for word-level open-vocabulary representations from characters. **arXiv preprint arXiv:2010.10392**, [S.l.], 2020.

BRAGA, A. V. et al. Machine learning: o uso da inteligência artificial na medicina. **Brazilian Journal of Development**, [S.l.], v. 5, n. 9, p. 16407–16413, 2019.

BRASIL, M. da saúde do. **Como surge o câncer?** Disponível em: <https://www.inca.gov.br/como-surge-o-cancer>. 04/04/2021.

BUCUR, A. et al. Clinical decision support framework for validation of multiscale models and personalization of treatment in oncology. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING, 13., 2013. **Anais...** [S.l.: s.n.], 2013. p. 1–4.

CAMMAROTA, G. et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. **Nature reviews gastroenterology & hepatology**, [S.l.], v. 17, n. 10, p. 635–648, 2020.

CESAR, L. B.; MANSO-CALLEJO, M.-Á.; CIRA, C.-I. Bert (bidirectional encoder representations from transformers) for missing data imputation in solar irradiance time series. **Engineering Proceedings**, [S.l.], v. 39, n. 1, p. 26, 2023.

CHEN, L. et al. Omdp: an ontology-based model for diagnosis and treatment of diabetes patients in remote healthcare systems. **International Journal of Distributed Sensor Networks**, [S.l.], v. 15, n. 5, p. 1550147719847112, 2019.

CHOI, E. et al. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. **arXiv preprint arXiv:1608.05745**, [S.l.], 2016.

CHOI, E. et al. Multi-layer representation learning for medical concepts. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 1495–1504.

CHOI, E. et al. Using recurrent neural network models for early detection of heart failure onset. **Journal of the American Medical Informatics Association**, [S.l.], v. 24, n. 2, p. 361–370, 2017.

CHRISTOPOULOU, F. et al. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. **Journal of the American Medical Informatics Association**, [S.l.], v. 27, n. 1, p. 39–46, 2020.

CONG, R. H. About the studying on the uncertainty of ontology on the bayesian network. In: APPLIED MECHANICS AND MATERIALS, 2014. **Anais...** [S.l.: s.n.], 2014. v. 513, p. 1717–1721.

COOPER, I. D. What is a "mapping study?". **Journal of the Medical Library Association: JMLA**, [S.l.], v. 104, n. 1, p. 76, 2016.

COTA, G. et al. The landscape of ontology reuse approaches. **Applications and Practices in Ontology Design, Extraction, and Reasoning**, [S.l.], v. 49, p. 21, 2020.

CUOCOLO, R. et al. Machine learning in oncology: a clinical appraisal. **Cancer letters**, [S.l.], v. 481, p. 55–62, 2020.

DAI, Z. et al. Transformer-xl: attentive language models beyond a fixed-length context. **arXiv preprint arXiv:1901.02860**, [S.l.], 2019.

DAWUD, A. M.; YURTKAN, K.; OZTOPRAK, H. Application of deep learning in neuroradiology: brain haemorrhage classification using transfer learning. **Computational Intelligence and Neuroscience**, [S.l.], v. 2019, 2019.

DEMNER-FUSHMAN, D.; CHAPMAN, W. W.; MCDONALD, C. J. What can natural language processing do for clinical decision support? **Journal of biomedical informatics**, [S.l.], v. 42, n. 5, p. 760–772, 2009.

DENG, L.; LIU, Y. **Deep learning in natural language processing**. [S.l.]: Springer, 2018.

DEVLIN, J. et al. Bert: pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, [S.l.], 2018.

DEVLIN, J. et al. Bert: pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, [S.l.], 2018.

DEVLIN, J. et al. Bert: pre-training of deep bidirectional transformers for language understanding. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, VOLUME 1 (LONG AND SHORT PAPERS), 2019., 2019. **Proceedings...** [S.l.: s.n.], 2019. v. 1, p. 4171–4186.

DEY, A. K. Understanding and using context. **Personal and ubiquitous computing**, [S.l.], v. 5, p. 4–7, 2001.

DHOLE, G.; UKE, N. Nlp based retrieval of medical information for diagnosis of human diseases. **Int J Renew Energy Technol**, [S.l.], v. 3, n. 10, p. 243e8, 2014.

DONG, C. et al. Character-based lstm-crf with radical-level features for chinese named entity recognition. In: NATURAL LANGUAGE UNDERSTANDING AND INTELLIGENT APPLICATIONS: 5TH CCF CONFERENCE ON NATURAL LANGUAGE PROCESSING AND CHINESE COMPUTING, NLPCC 2018, AND 24TH INTERNATIONAL CONFERENCE ON COMPUTER PROCESSING OF ORIENTAL LANGUAGES, ICCPOL 2018, KUNMING, CHINA, DECEMBER 2–6, 2018, PROCEEDINGS 24, 2018. **Anais...** [S.l.: s.n.], 2018. p. 239–250.

DUAN, L.; STREET, W. N.; XU, E. Healthcare information systems: data mining methods in the creation of a clinical recommender system. **Enterprise Information Systems**, [S.l.], v. 5, n. 2, p. 169–181, 2011.

ERSHOFF, B. D. et al. Training and validation of deep neural networks for the prediction of 90-day post-liver transplant mortality using unos registry data. In: TRANSPLANTATION PROCEEDINGS, 2020. **Anais...** [S.l.: s.n.], 2020. v. 52, n. 1, p. 246–258.

ESTEVA, A. et al. A guide to deep learning in healthcare. **Nature medicine**, [S.l.], v. 25, n. 1, p. 24–29, 2019.

FABREGAT, H.; ARAUJO, L.; MARTINEZ-ROMO, J. Deep neural models for extracting entities and relationships in the new rdd corpus relating disabilities and rare diseases. **Computer methods and programs in biomedicine**, [S.l.], v. 164, p. 121–129, 2018.

FANG, R. et al. Computational health informatics in the big data age: a survey. **ACM Computing Surveys (CSUR)**, [S.l.], v. 49, n. 1, p. 1–36, 2016.

FENG, M.; XIANG, B.; ZHOU, B. Distributed deep learning for question answering. In: ACM INTERNATIONAL ON CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 25., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 2413–2416.

FENZ, S. An ontology-based approach for constructing bayesian networks. **Data & Knowledge Engineering**, [S.l.], v. 73, p. 73–88, 2012.

FERNANDES et al. A framework for preditivos e participativa e-systems. , [S.l.], p. 123–129, 2009.

FORD, E. et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. **Journal of the American Medical Informatics Association**, [S.l.], v. 23, n. 5, p. 1007–1015, 2016.

GEORGE, G.; LAL, A. M. Review of ontology-based recommender systems in e-learning. **Computers & Education**, [S.l.], v. 142, p. 103642, 2019.

GOLDSTEIN, B. A. et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. **Journal of the American Medical Informatics Association**, [S.l.], v. 24, n. 1, p. 198–208, 2017.

GONZALEZ-HERNANDEZ, G. et al. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. **Yearbook of medical informatics**, [S.l.], v. 26, n. 1, p. 214, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT Press, 2016.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural Networks**, [S.l.], v. 18, n. 5-6, p. 602–610, 2005.

GROVES, P. et al. The'big data'revolution in healthcare: accelerating value and innovation. , [S.l.], 2016.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? **International journal of human-computer studies**, [S.l.], v. 43, n. 5-6, p. 907–928, 1995.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: **Handbook on ontologies**. [S.l.]: Springer, 2009. p. 1–17.

GUO, Z.; ZHANG, Y.; LU, W. Attention guided graph convolutional networks for relation extraction. **arXiv preprint arXiv:1906.07510**, [S.l.], 2019.

HE, Y. et al. Deeponto: a python package for ontology engineering with deep learning. **arXiv preprint arXiv:2307.03067**, [S.l.], 2023.

HEART, T.; BEN-ASSULI, O.; SHABTAI, I. A review of phr, emr and ehr integration: a more personalized healthcare and public health policy. **Health Policy and Technology**, [S.l.], v. 6, n. 1, p. 20–25, 2017.

HENDRICKX, I. et al. Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. In: INTERNATIONAL WORKSHOP ON SEMANTIC EVALUATION, 5., 2010. **Proceedings...** [S.l.: s.n.], 2010. p. 33–38.

HERNANDES, E. et al. Avaliação da ferramenta start utilizando o modelo tam e o paradigma gqm. In: EXPERIMENTAL SOFTWARE ENGINEERING LATIN AMERICAN WORKSHOP (ESELAW 2010), 7., 2010. **Proceedings...** [S.l.: s.n.], 2010. p. 30.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, [S.l.], v. 9, n. 8, p. 1735–1780, 1997.

HODGES, S. **The counseling practicum and internship manual**: a resource for graduate counseling programs. [S.l.]: New York, NY: Springer Publishing Company, 2011.

HOOLEY, I. et al. Pns98 natural language processing (nlp)-based detection of transgender and gender non-conforming patients in electronic health record (ehr)-derived data. **Value in Health**, [S.l.], v. 24, p. S190, 2021.

HUANG, K.; ALTOSAAR, J.; RANGANATH, R. Clinicalbert: modeling clinical notes and predicting hospital readmission. **arXiv preprint arXiv:1904.05342**, [S.l.], 2019.

HUANG, P. et al. Prediction of lung cancer risk at follow-up screening with low-dose ct: a training and validation study of a deep learning method. **The Lancet Digital Health**, [S.l.], v. 1, n. 7, p. e353–e362, 2019.

HUANG, Z.; DONG, W.; DUAN, H. A probabilistic topic model for clinical risk stratification from electronic health records. **Journal of Biomedical Informatics**, [S.l.], v. 58, p. 28–36, 2015.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. **Automated machine learning**: methods, systems, challenges. [S.l.]: Springer Nature, 2019.

IBRAHIM, A. et al. Analysis of the suitability of existing medical ontologies for building a scalable semantic interoperability solution supporting multi-site collaboration in oncology. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING, 2014., 2014. **Anais...** [S.l.: s.n.], 2014. p. 204–211.

ISLAM, M. T. et al. Extracting biomarker information applying natural language processing and machine learning. In: INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICAL ENGINEERING, 2010., 2010. **Anais...** [S.l.: s.n.], 2010. p. 1–4.

JAGANNATHA, A. N.; YU, H.; LIU, F. Structured prediction models for rnn based sequence labeling in clinical text. In: WORKSHOP ON BIOMEDICAL NATURAL LANGUAGE PROCESSING, 15., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 112–120.

JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. **Nature Reviews Genetics**, [S.l.], v. 13, n. 6, p. 395–405, 2012.

JI, Z.; WEI, Q.; XU, H. Bert-based ranking for biomedical entity normalization. **AMIA Summits on Translational Science Proceedings**, [S.l.], v. 2020, p. 269, 2020.

JOHNSON, M. et al. Google's multilingual neural machine translation system: enabling zero-shot translation. **Transactions of the Association for Computational Linguistics**, [S.l.], v. 5, p. 339–351, 2017.

JOSHI, M. et al. Spanbert: improving pre-training by representing and predicting spans. **Transactions of the Association for Computational Linguistics**, [S.l.], v. 8, p. 64–77, 2020.

JUHN, Y.; LIU, H. Artificial intelligence approaches using natural language processing to advance ehr-based clinical research. **Journal of Allergy and Clinical Immunology**, [S.l.], v. 145, n. 2, p. 463–469, 2020.

KAISER, L. et al. Model-based reinforcement learning for atari. **arXiv preprint arXiv:1903.00374**, [S.l.], 2019.

KALET, A. M. et al. Developing bayesian networks from a dependency-layered ontology: a proof-of-concept in radiation oncology. **Medical physics**, [S.l.], v. 44, n. 8, p. 4350–4359, 2017.

KELLEHER, J. D. **Deep learning**. [S.l.]: MIT press, 2019.

KHAN, S.; SHAMSI, J. A. Health quest: a generalized clinical decision support system with multi-label classification. **Journal of King Saud University-Computer and Information Sciences**, [S.l.], v. 33, n. 1, p. 45–53, 2021.

KITCHENHAM, B. A.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.]: Keele University and Durham University Joint Report, 2007. (EBSE 2007-001).

KÖCHE, J. C. **Fundamentos de metodologia científica**. [S.l.]: Editora Vozes, 2016.

KOLECK, T. A. et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. **Journal of the American Medical Informatics Association**, [S.l.], v. 26, n. 4, p. 364–379, 2019.

KOROTEEV, M. Bert: a review of applications in natural language processing and understanding. **arXiv preprint arXiv:2103.11943**, [S.l.], 2021.

KREIMEYER, K. et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. **Journal of biomedical informatics**, [S.l.], v. 73, p. 14–29, 2017.

LAI, K. et al. A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures. **Information Processing & Management**, [S.l.], v. 59, n. 1, p. 102735, 2022.

LAMPLE, G.; CONNEAU, A. Cross-lingual language model pretraining. **arXiv preprint arXiv:1901.07291**, [S.l.], 2019.

LAMPLE, G. et al. Neural architectures for named entity recognition. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2016., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 260–270.

LAMURIAS, A. et al. Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies. **BMC bioinformatics**, [S.l.], v. 20, n. 1, p. 1–12, 2019.

LAMY, J.-B. et al. Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. **Artificial intelligence in medicine**, [S.l.], v. 94, p. 42–53, 2019.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, [S.l.], v. 521, n. 7553, p. 436, 2015.

LEE, H. et al. Quote recommendation in dialogue using deep neural network. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 39., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 957–960.

LEE, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, [S.l.], v. 36, n. 4, p. 1234–1240, 2020.

LEHMAN, E. et al. Does bert pretrained on clinical notes reveal sensitive data? **arXiv preprint arXiv:2104.07762**, [S.l.], 2021.

LEWIS, M. et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. **arXiv preprint arXiv:1910.13461**, [S.l.], 2019.

LI, F. et al. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study. **JMIR medical informatics**, [S.l.], v. 7, n. 3, p. e14830, 2020.

LI, F.; YU, H. An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models. **Journal of the American Medical Informatics Association**, [S.l.], v. 26, n. 7, p. 646–654, 2019.

LI, P. et al. Multilingual named entity recognition using bilstm-crf with pretrained language models. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING, 2022. **Anais...** [S.l.: s.n.], 2022. p. 276–287.

LI, S. et al. Long-term structural health monitoring for bridge based on back propagation neural network and long and short-term memory. **Structural Health Monitoring**, [S.l.], v. 22, n. 4, p. 2325–2345, 2023.

LI, Y. et al. Behrt: transformer for electronic health records. **Scientific reports**, [S.l.], v. 10, n. 1, p. 1–12, 2020.

LI, Z. et al. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. **BMC medical informatics and decision making**, [S.l.], v. 19, n. 1, p. 1–8, 2019.

LIN, Y. et al. Long-distance disorder-disorder relation extraction with bootstrapped noisy data. **Journal of Biomedical Informatics**, [S.l.], v. 109, p. 103529, 2020.

LIU, Y. et al. Roberta: a robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, [S.l.], 2019.

LIU, Y. et al. Roberta: a robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, [S.l.], 2019.

LONDHE, V. Y.; BHASIN, B. Artificial intelligence and its potential in oncology. **Drug discovery today**, [S.l.], v. 24, n. 1, p. 228–232, 2019.

LOPES, É.; CORREA, U.; FREITAS, L. A. de. Exploring bert for aspect extraction in portuguese language. In: THE INTERNATIONAL FLAIRS CONFERENCE PROCEEDINGS, 2021. **Anais...** [S.l.: s.n.], 2021. v. 34, n. 1.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, [S.l.], 2017.

MA, X.; HOVY, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 1: LONG PAPERS), 54., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 1064–1074.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, [S.l.], v. 9, p. 381–386, 2020.

MAHMOUD, N.; ELBEH, H.; ABDLKADER, H. M. Ontology learning based on word embeddings for text big data extraction. In: INTERNATIONAL COMPUTER ENGINEERING CONFERENCE (ICENCO), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 183–188.

MANLEY, H. et al. Dynamic changes of convolutional neural network-based mammographic breast cancer risk score among women undergoing chemoprevention treatment. **Clinical Breast Cancer**, [S.l.], 2020.

MANNING, C.; SCHUTZE, H. **Foundations of statistical natural language processing**. [S.l.]: MIT press, 1999.

MATE, S. et al. A method for the graphical modeling of relative temporal constraints. **Journal of biomedical informatics**, [S.l.], v. 100, p. 103314, 2019.

MCCANN, B. et al. Learned in translation: contextualized word vectors. **arXiv preprint arXiv:1708.00107**, [S.l.], 2017.

MEHTA, N.; PANDIT, A. Concurrence of big data analytics and healthcare: a systematic review. **International journal of medical informatics**, [S.l.], v. 114, p. 57–65, 2018.

MEYER, A. et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. **The Lancet Respiratory Medicine**, [S.l.], v. 6, n. 12, p. 905–914, 2018.

MICELI, M.; SCHUESSLER, M.; YANG, T. Between subjectivity and imposition: power dynamics in data annotation for computer vision. **Proceedings of the ACM on Human-Computer Interaction**, [S.l.], v. 4, n. CSCW2, p. 1–25, 2020.

MIN, L. et al. An openehr based approach to improve the semantic interoperability of clinical data registry. **BMC Medical Informatics and Decision Making**, [S.l.], v. 18, n. 1, p. 15, Mar 2018.

MINTZ, M. et al. Distant supervision for relation extraction without labeled data. **ACM Transactions on Information Systems (TOIS)**, [S.l.], v. 27, n. 2, p. 12, 2009.

MORGAN, S.; RANASINGHE, T.; ZAMPIERI, M. Wlv-rit at germeval 2021: multitask learning with transformers to detect toxic, engaging, and fact-claiming comments. **arXiv preprint arXiv:2108.00057**, [S.l.], 2021.

MOZAFARI, M.; FARAHBAKHSH, R.; CRESPI, N. A bert-based transfer learning approach for hate speech detection in online social media. In: COMPLEX NETWORKS AND THEIR APPLICATIONS VIII: VOLUME 1 PROCEEDINGS OF THE EIGHTH INTERNATIONAL CONFERENCE ON COMPLEX NETWORKS AND THEIR APPLICATIONS COMPLEX NETWORKS 2019 8, 2020. **Anais...** [S.l.: s.n.], 2020. p. 928–940.

MUNIR, K. et al. Cancer diagnosis using deep learning: a bibliographic review. **Cancers**, [S.l.], v. 11, n. 9, p. 1235, 2019.

MUSEN, M. A. The protégé project: a look back and a look forward. **AI matters**, [S.l.], v. 1, n. 4, p. 4–12, 2015.

NÉVÉOL, A.; ZWEIGENBAUM, P. et al. Making sense of big textual data for health care: findings from the section on clinical natural language processing. **Yearbook of medical informatics**, [S.l.], v. 26, n. 1, p. 228, 2017.

NGIAM, K. Y.; KHOR, W. Big data and machine learning algorithms for health-care delivery. **The Lancet Oncology**, [S.l.], v. 20, n. 5, p. e262–e273, 2019.

PEHLEVAN, C.; CHKLOVSKII, D. B. Neuroscience-inspired online unsupervised learning algorithms: artificial neural networks. **IEEE Signal Processing Magazine**, [S.l.], v. 36, n. 6, p. 88–96, 2019.

PENG, W.; YE, Z.-S.; CHEN, N. Bayesian deep-learning-based health prognostics toward prognostics uncertainty. **IEEE Transactions on Industrial Electronics**, [S.l.], v. 67, n. 3, p. 2283–2293, 2019.

PEREIRA, A. L. D. et al. Proteção de dados e segurança informática no setor da saúde: o papel dos responsáveis pela proteção de dados no direito da união europeia. **Cadernos Ibero-Americanos de Direito Sanitário**, [S.l.], v. 10, n. 2, p. 211–232, 2021.

PERER, A.; WANG, F.; HU, J. Mining and exploring care pathways from electronic medical records with visual analytics. **Journal of biomedical informatics**, [S.l.], v. 56, p. 369–378, 2015.

PERLIS, R. et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. **Psychological medicine**, [S.l.], v. 42, n. 1, 2012.

PETERS, M. E. et al. Deep contextualized word representations. **arXiv preprint arXiv:1802.05365**, [S.l.], 2018.

POLPINIJ, J. The cancerology ontology: designed to support the search of evidence-based oncology from biomedical literatures. In: INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS (CBMS), 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p. 1–6.

PRADHAN, S.; ELHADAD, N.; SOUTH, B. R. Evaluating the state-of-the-art in automatic de-identification. **Journal of the American Medical Informatics Association**, [S.l.], v. 20, n. 5, p. 843–848, 2013.

PRISMA. **Preferred reporting items for systematic reviews and meta-analyses**. Disponível em: <http://prisma-statement.org/PRISMAStatement/Checklist.aspx>. Acessado em 08/04/2021.

PURUSHOTHAM, S. et al. Benchmarking deep learning models on large healthcare datasets. **Journal of biomedical informatics**, [S.l.], v. 83, p. 112–134, 2018.

QI, J.; DING, L.; LIM, S. Ontology-based knowledge representation of urban heat island mitigation strategies. **Sustainable Cities and Society**, [S.l.], v. 52, p. 101875, 2020.

QI, P. et al. Stanza: a python natural language processing toolkit for many human languages. **arXiv preprint arXiv:2003.07082**, [S.l.], 2020.

QIU, X. et al. Pre-trained models for natural language processing: a survey. **Science China Technological Sciences**, [S.l.], p. 1–26, 2020.

RADFORD, A. et al. Language models are unsupervised multitask learners. **OpenAI blog**, [S.l.], v. 1, n. 8, p. 9, 2019.

RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **The Journal of Machine Learning Research**, [S.l.], v. 21, n. 1, p. 5485–5551, 2020.

RAMESH, S. et al. Bert based transformers lead the way in extraction of health information from social media. In: SIXTH SOCIAL MEDIA MINING FOR HEALTH (# SMM4H) WORKSHOP AND SHARED TASK, 2021. **Proceedings...** [S.l.: s.n.], 2021. p. 33–38.

RASMUSSEN, M. H. et al. Bot: the building topology ontology of the w3c linked building data group. **Semantic Web**, [S.l.], n. Preprint, p. 1–19, 2021.

RASMY, L. et al. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. **NPJ digital medicine**, [S.l.], v. 4, n. 1, p. 1–13, 2021.

RAZZAQUE, A.; HAMDAN, A. Artificial intelligence based multinational corporate model for ehr interoperability on an e-health platform. In: **Artificial intelligence for sustainable development**: theory, practice and future applications. [S.l.]: Springer, 2021. p. 71–81.

RUDRAPATNA, V. A.; BUTTE, A. J. Opportunities and challenges in using real-world data for health care. **The Journal of Clinical Investigation**, [S.l.], v. 130, n. 2, p. 565–574, 2 2020.

RUSSELL, S.; NORVIG, P. **Artificial intelligence**: a modern approach. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2021.

SABRA, S. et al. Performance evaluation for semantic-based risk factors extraction from clinical narratives. In: IEEE 8TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE (CCWC), 2018., 2018. **Anais...** [S.l.: s.n.], 2018. p. 695–701.

SALATINO, A. A. et al. The computer science ontology: a comprehensive automatically-generated taxonomy of research areas. **Data Intelligence**, [S.l.], v. 2, n. 3, p. 379–416, 2020.

SANH, V. et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. **arXiv preprint arXiv:1910.01108**, [S.l.], 2019.

SARAIVA, R. et al. Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning. **Expert Systems with Applications**, [S.l.], v. 61, p. 192–202, 2016.

SATYANARAYANAN, M. Pervasive computing: vision and challenges. **IEEE Personal communications**, [S.l.], v. 8, n. 4, p. 10–17, 2001.

SCHNEIDER, E. T. R. et al. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: CLINICAL NATURAL LANGUAGE PROCESSING WORKSHOP, 3., 2020. **Proceedings...** [S.l.: s.n.], 2020. p. 65–72.

SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE Transactions on Signal Processing**, [S.l.], v. 45, n. 11, p. 2673–2681, 1997.

SCHWALBE, N.; WAHL, B. Artificial intelligence and the future of global health. **The Lancet**, [S.l.], v. 395, n. 10236, p. 1579–1586, 2020.

SCHWERTNER, M. A. **Exploring text classification methods in oncological medical notes using machine learning and deep learning**. 2020. Dissertação (mestrado) — Universidade do Val do Rio dos Sinos, São Leopoldo, 2020.

SCHWERTNER, M. A. et al. Fostering natural language question answering over knowledge bases in oncology ehr. **32nd IEEE CBMS International Symposium on Computer-Based Medical Systems**, [S.l.], 2019. Available at `http://www.cbms2019.org/`.

SCHWERTNER, M. A.; RIGO, S. J. Sistemas de apoio à decisão clínica e extração de informação na oncologia. **XVI Congresso Brasileiro de Informática em Saúde - CBIS 2018**, [S.l.], 2018. Available at `http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/issue/view/86`.

SCIENCE, W. of. **Web of science**. Disponível em: <http://www.apps-webofknowledge.ez101.periodicos.capes.gov.br>. 04/04/2021.

SCIENCEDIRECT. **Sciencedirect, health and medical journals**. Disponível em: <https://www.sciencedirect.com/>. 06/04/2021.

SEGURA BEDMAR, I.; MARTÍNEZ, P.; HERRERO ZAZO, M. Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In: OF THE 3RD , 2013. **Anais...** [S.l.: s.n.], 2013.

SETTLES, B. Active learning literature survey. **Computer sciences technical report 1648, University of Wisconsin-Madison**, [S.l.], 2009.

SEZGIN, E.; ÖZKAN, S. A systematic literature review on health recommender systems. In: E-HEALTH AND BIOENGINEERING CONFERENCE (EHB), 2013., 2013. **Anais...** [S.l.: s.n.], 2013. p. 1–4.

SHARMA, S.; KUMAR, A.; RANA, V. Ontology based informational retrieval system on the semantic web: semantic web mining. In: INTERNATIONAL CONFERENCE ON NEXT GENERATION COMPUTING AND INFORMATION SYSTEMS (ICNGCIS), 2017., 2017. **Anais...** [S.l.: s.n.], 2017. p. 35–37.

SHICKEL, B. et al. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. **IEEE journal of biomedical and health informatics**, [S.l.], v. 22, n. 5, p. 1589–1604, 2017.

SHIMIZU, H.; NAKAYAMA, K. I. Artificial intelligence in oncology. **Cancer science**, [S.l.], v. 111, n. 5, p. 1452, 2020.

SILVA, A. V. **Um modelo de classificação para o reconhecimento de entidades nomeadas**. 2020. Tese (Doutorado em Ciência da Computação) — Universidade de São Paulo, 2020.

SILVA, D. P. et al. Inteligência artificial aplicada a um simulador na Área da saúde. xvi congresso brasileiro de informática em saúde. **Congresso Brasileiro de Informática em Saúde**, [S.l.], p. 641–654, 2018.

SILVA, F. M. d. et al. Reconhecimento de padroes por compressÃo aplicado ao processamento da linguagem natural. , [S.l.], 2019.

SOCIETY, M. P. **Medical records**. Available at `https://www.medicalprotection.org/uk/articles/eng-medical-records`.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, [S.l.], v. 45, n. 4, p. 427–437, 2009.

SOLARES, J. R. A. et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. **Journal of biomedical informatics**, [S.l.], v. 101, p. 103337, 2020.

SOUZA, A. S.; DURAN, A.; VIEIRA, V. Uma ontologia de domínio para a metodologia de aprendizagem baseada em problemas. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE), 2014. **Anais...** [S.l.: s.n.], 2014. v. 25, n. 1, p. 1253.

SOUZA, B. S. W. de et al. Panorama do tema qualidade de vida nas publicações científicas em oncologia nos últimos 10 anos: um estudo de análise de redes e processamento de linguagem natural. **Brazilian Journal of Health Review**, [S.l.], v. 4, n. 2, p. 9103–9731, 2021.

SOUZA, E. P. d. et al. Aplicações do deep learning para diagnóstico de doenças e identificação de insetos vetores. **Saúde em Debate**, [S.l.], v. 43, p. 147–154, 2020.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 2020. **Anais...** [S.l.: s.n.], 2020. p. 403–417.

SOUZA, J. V. A. de et al. A multilabel approach to portuguese clinical named entity recognition. **Journal of Health Informatics**, [S.l.], v. 12, 2021.

STEPHENS, Z. D. et al. Big data: astronomical or genomical? **PLoS biology**, [S.l.], v. 13, n. 7, p. e1002195, 2015.

SUÁREZ-PANIAGUA, V. et al. A two-stage deep learning approach for extracting entities and relationships from medical texts. **Journal of biomedical informatics**, [S.l.], v. 99, p. 103285, 2019.

SUGIMURA, V.; ISHIGAKI, V. K. New web-usability evaluation method: scenario-based walkthrough. **Fujitsu Sci. Tech. J**, [S.l.], v. 41, n. 1, p. 105–114, 2005.

SUTHERLAND, A. et al. Incidence and prevalence of intravenous medication errors in the uk: a systematic review. **European Journal of Hospital Pharmacy**, [S.l.], v. 27, n. 1, p. 3–8, 2020.

TJONG KIM SANG, E. F.; DE MEULDER, F. Introduction to the conll-2003 shared task: language-independent named entity recognition. In: NATURAL LANGUAGE LEARNING AT HLT-NAACL 2003-VOLUME 4, 2003. **Proceedings...** [S.l.: s.n.], 2003. p. 142–147.

TOLEU, A.; TOLEGEN, G.; MAKAZHANOV, A. Character-based deep learning models for token and sentence segmentation. , [S.l.], 2017.

TOPAZ, M. et al. Automated identification of wound information in clinical notes of patients with heart diseases: developing and validating a natural language processing application. **International journal of nursing studies**, [S.l.], v. 64, p. 25–31, 2016.

TURCHIN, A.; MASHARSKY, S.; ZITNIK, M. Comparison of bert implementations for natural language processing of narrative medical documents. **Informatics in Medicine Unlocked**, [S.l.], v. 36, p. 101139, 2023.

URBANSKYY, S. M. C. **Uma ontologia para representação de conhecimento sobre boas práticas nas infecções relacionadas à assistência à saúde**. 2018. Dissertação Mestrado em Ensino na Saúde — Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, 2018.

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, [S.l.], v. 30, 2017.

WANG, Y. et al. Nested named entity recognition: a survey. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, [S.l.], v. 16, n. 6, p. 1–29, 2022.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação**. [S.l.]: Elsevier, 2009. v. 2.

WEI, Q. et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. **Journal of the American Medical Informatics Association**, [S.l.], v. 27, n. 1, p. 13–21, 2020.

WIKSTRÖM, K. et al. Electronic health records as valuable data sources in the health care quality improvement process. **Health services research and managerial epidemiology**, [S.l.], v. 6, p. 2333392819852879, 2019.

WOLF, T. et al. Transformers: state-of-the-art natural language processing. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING: SYSTEM DEMONSTRATIONS, 2020., 2020. **Proceedings...** [S.l.: s.n.], 2020. p. 38–45.

WU, S.; HE, Y. Enriching pre-trained language model with entity information for relation classification. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 28., 2019. **Proceedings...** [S.l.: s.n.], 2019. p. 2361–2364.

XPLORE, I. **Ieee xplore digital library**. Disponível em: <http://ieeexplore.ieee.org/xpl/aboutUs.jsp>. 08/04/2021.

XU, G.; WANG, C.; HE, X. Improving clinical named entity recognition with global neural attention. In: WEB AND BIG DATA: SECOND INTERNATIONAL JOINT CONFERENCE, APWEB-WAIM 2018, MACAU, CHINA, JULY 23-25, 2018, PROCEEDINGS, PART II 2, 2018. **Anais...** [S.l.: s.n.], 2018. p. 264–279.

XUE, K. et al. Fine-tuning bert for joint entity and relation extraction in chinese medical text. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2019., 2019. **Anais...** [S.l.: s.n.], 2019. p. 892–897.

YAN, L. et al. Development of a novel asset management system for power transformers based on ontology. In: IEEE PES ASIA-PACIFIC POWER AND ENERGY ENGINEERING CONFERENCE (APPEEC), 2013., 2013. **Anais...** [S.l.: s.n.], 2013. p. 1–6.

YANG, X. et al. Madex: a system for detecting medications, adverse drug events, and their relations from clinical notes. **Drug safety**, [S.l.], v. 42, n. 1, p. 123–133, 2019.

YANG, X. et al. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. **Journal of the American Medical Informatics Association**, [S.l.], v. 27, n. 1, p. 65–72, 2020.

YANG, Z. et al. Xlnet: generalized autoregressive pretraining for language understanding. **Advances in neural information processing systems**, [S.l.], v. 32, 2019.

YOON, W. et al. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. **BMC bioinformatics**, [S.l.], v. 20, n. 10, p. 55–65, 2019.

YOSINSKI, J. et al. How transferable are features in deep neural networks? In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2014. **Anais...** [S.l.: s.n.], 2014. p. 3320–3328.

YU, Z. et al. Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods. **BMC Medical Informatics and Decision Making**, [S.l.], v. 22, n. 3, p. 1–9, 2022.

ZANATTA, E. J. et al. Modelando ontologias a partir de diretrizes clínicas: diagnóstico e tratamento da cefaléia. In: ONTOBRAS-MOST, 2012. **Anais...** [S.l.: s.n.], 2012. p. 272–277.

ZANETTI, H. A.; BONACIN, R. Uma metodologia baseada em semiótica para elaboração e análise de práticas de ensino de programação com robótica pedagógica. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION (SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO-SBIE), 2014. **Anais...** [S.l.: s.n.], 2014. v. 25, n. 1, p. 1233.

ZENG, D. et al. Relation classification via convolutional deep neural network. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 25., 2014. **Proceedings...** [S.l.: s.n.], 2014. p. 2335–2344.

ZHANG, R. et al. Automatic methods to extract new york heart association classification from clinical notes. In: 2017 IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2017. **Anais...** [S.l.: s.n.], 2017. p. 1296–1299.

ZHAO, Y.-S. et al. Leveraging text skeleton for de-identification of electronic medical records. **BMC Medical Informatics and Decision Making**, [S.l.], v. 18, n. 1, p. 18, Mar 2018.

ZHOU, H. Developing natural language processing to extract complementary and integrative health information from electronic health record data. In: IEEE 10TH INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS (ICHI), 2022., 2022. **Anais...** [S.l.: s.n.], 2022. p. 474–475.

ZHOU, P. et al. Attention-based bidirectional long short-term memory networks for relation classification. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (VOLUME 2: SHORT PAPERS), 54., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 207–212.

ZHU, X.; SOBHANI, P.; GUO, H. Dag-structured long short-term memory for semantic compositionality. In: HUMAN LANGUAGE TECHNOLOGIES, 2016., 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 917–926.

ZURYNSKI, Y. et al. Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays. **Orphanet Journal of Rare Diseases**, [S.l.], v. 12, n. 1, p. 1–9, 2017.

APÊNDICE

A Protocolo da Revisão Sistemática

Título: Revisão sistemática sobre Extração de relações, *Deep Learning* e Processamento de Linguagem Natural.

Resumo: O tema desta revisão é sobre ER, DL e PLN, buscando também assuntos de interesse referentes às mesmas, como suas ligações com a saúde e técnicas de extração de dados e uso de ontologias. O protocolo foi desenvolvido com a ajuda de especialistas da área de computação (ligados a saúde e IA) e, caso for necessário, poderá sofrer ajustes posteriormente, sendo todas as modificações serão justificadas.

Objetivo: PICOC

Para o auxílio na criação da textistring de busca, Kitchenham and Charters (2007) recomenda a utilização do PICOC (População, Intervenção, Comparação, Resultados e Contexto). A formulação da pesquisa é apresentada a seguir.

FORMULAÇÃO DE PESQUISA

Foco da questão: Está revisão sistemática busca encontrar os artigos que trabalham com ER, DL e PLN, buscando encontrar técnicas e validações de IA, visando principalmente aplicações à saúde.

Questões de interesse: Técnicas utilizadas, Abordagens, Estudos já realizados, Aplicações recorrentes.

Palavras-chaves: *Deep Learning, Natural Language Processing, Data extraction, Ontology, Health*.

Intervenção: Verificar as tecnologias que são utilizadas nas pesquisas.

Controle: Não será utilizado.

Efeito: Identificar as oportunidades de pesquisa na área de IA, voltando o conhecimento para ER, DL e PLN.

Medida do Resultado: Gerar um embasamento teórico para a tese, assim como a publicação de artigos científicos.

População de interesse: Pesquisadores, professores, desenvolvedores e profissionais da área de IA.

Aplicação: Esta revisão sistemática tem como foco pesquisadores, professores e alunos da área da computação aplicada, pois trará uma visão abrangente de tecnologias e metodologias adotadas até o momento para desenvolvimento de pesquisas.

Desenho do experimento: Não será desenvolvido.

Financiamento: CNPQ-DAI

FORMULAÇÃO DE CRITÉRIOS

Definição dos critérios de seleção das fontes de dados: As fontes de dados foram selecionadas através de trabalhos relacionados e indicações dos orientadores deste projeto. As

mesmas possuem artigos relacionados a computação. Para a área da computação serão utilizadas as bases de dados: Web Of Science, que possui o acesso a periódicos em diversas áreas do conhecimento, tendo conteúdo integral apenas para pesquisadores da CAPES (SCIENCE, 2021); IEEExplore, que fornece acesso a algumas das publicações mais citadas no mundo em Ciência da Computação (XPLORE, 2021).

Idiomas das Fontes de Dados: Somente serão consideradas as publicações que estiverem no idioma em inglês ou português.

***String* de busca:** A *String* de busca foi criada conforme as palavras-chave definidas, tendo como elementos obrigatórios da pesquisa o uso de "Processamento de Linguagem Natural" (*natural language processing*), "Extração de Relações" (*data extraction*) e "*Deep Learning*" ("*deep learning*"). Os termos "Saúde" (*health*) e "Ontologia" (*ontology*) serão definidos como elementos auxiliares. A *String* será aplicada nas bases de dados definidas anteriormente. ("*deep learning*") AND ("*relation extraction*") OR ("*natural language processing*" OR "NLP") OR ("*health*") OR ("*ontology*").

Artigos de controle: Optou-se por não utilizar nenhum artigo de controle para esta revisão sistemática.

SELEÇÃO DOS ESTUDOS

Critérios para a inclusão/exclusão dos resultados:

- O ano de publicação do artigo deve estar dentro do período de 2016 e 2023;
- Ser um artigo científico publicado;
- Deve estar escrito em inglês ou português;
- A publicação deve estar disponível na íntegra na internet ou disponível através de convênios das instituições de ensino.

Definição dos tipos de estudo: Serão selecionados estudos dos tipos teóricos, qualitativos e quantitativos referentes aos temas.

Procedimentos para seleção dos estudos: Na primeira etapa, será executada a *string* de busca nas bases de dados selecionadas. Todas as publicações encontradas serão exportadas no formato BibTex para serem cadastradas na ferramenta StArt. Após o registro, serão validados os critérios de inclusão já definidos. Avante, serão realizadas as leituras de todos os títulos, palavras-chave e resumos, passando em seguida para uma validação de introdução e conclusão e, na fase final, a leitura integral dos artigos selecionados.

Fases de seleção de artigos:

- Fase 1 - Validar os critérios de inclusão e exclusão;
- Fase 2 - Leitura do título, palavras-chave e resumo;
- Fase 3 - Leitura da introdução e conclusão;

- Fase 4 - Leitura integral dos artigos e validação das respostas para as perguntas.

Critérios de qualidade das fases da Revisão Sistemática: Os critérios de qualidade que serão avaliados nesta revisão sistemática estão descritos a seguir:

Os artigos foram aplicados à área da saúde?

- Qual o objetivo da pesquisa?
- Apresentou uso de ontologias?
- O artigo foi aplicado em algum protótipo?
- Qual o idioma alvo?
- Quais relações foram extraídas?
- Qual a área de atuação?
- Qual a origem do *dataset*?
- Quais são as abordagens, técnicas e métodos usados?

Análises adicionais: Não há.

B Lista de relações da ontologia

Está sessão apresenta alguns tipos de relações presentes na ontologia e exemplos reais de uso, extraídos do EHR Gemed Onco.

B.1 Sintomas e diagnósticos

Paciente relata {sintoma ou lista sintomas}.

Exemplos:

- Paciente relata evolução com distensão abdominal e parada de eliminação de gases, colonoscopia com tumor em sigmoide;
- Paciente relata radioterapia em 10/2015 por Ca de próstata...
- Paciente relata tratamento anterior de endometriose.

Perguntas:

- Quais os sintomas que o paciente {PacienteNome} relata?

{diagnóstico} {características diagnóstico} {órgão}.

Exemplos:

- Carcinoma de células renais cromófobo IVE Pulmão;
- Ca de cólon qm 2º linha paliativa;
- Adenocarcinoma de próstata G 6, PSAi= 7,0ng/ml , T2c / Risco intermediário;
- Ca de cólon EII alto risco.

Perguntas:

- Qual o diagnóstico do paciente {PacienteNome}?;
- Quais pacientes possuem o diagnóstico {Diagnostico}?.

Paciente evoluiu com {sintoma} há {período de tempo} com {diagnóstico}.
Exemplos:

- Paciente evoluiu com obstrução intestinal há 2 anos com adenocarcinoma de reto ressecado.

Perguntas:

- Quais sintomas os pacientes com diagnóstico {Diagnostico} relatam?
- O sintoma {sintoma} está associado com quais diagnósticos?

Paciente com {sintoma1} e {sintoma2}.
Exemplos:

- Paciente om dor e desconforto em HCD. Observação: HCD não é um órgão e sim uma doença (Heavy Chain Desease).

Paciente notou {sintoma} em {órgão} {local órgão}.
Exemplos:

- Paciente notou nódulo em mama E...
- Notou nódulo em mama D.

Perguntas:

- Quais sintomas estão aparecem no órgão {orgao}?

Paciente notou há {período de tempo} {sintoma} {órgão} com {característica sintoma}.
Exemplos:

- Paciente notou há 6 meses nódulo em mama E com crescimento moderado.

Perguntas:

- Há quanto tempo paciente notou {sintoma} no órgão {orgao}?

Há {período de tempo} notou {sintoma} {órgão}.
Exemplos:

- Há 3 meses notou nódulo em mama direita...
- Há 3 meses edema de face.

Paciente e familiares relatam {diagnóstico} diagnosticado há {período de tempo}.
Exemplos:

- Paciente e familiares relatam Ca de endométrio diagnosticado há 6-7 anos sem tratamento...

Perguntas:

- Há quanto tempo paciente {PacienteNome} notou diagnóstico {diagnostico}?

{diagnóstico} {estadio} {características diagnóstico}.
Exemplos:

- Ca de ovário EIV resistente a platina;
- CA de mama EIIa.

{sintoma} {intensidade sintoma}.
Exemplos:

- Dor lombar intensa, sem melhora com tramadol;
- Dor lombar intensa, paresia de membros inferiores.

{sintoma} em região {área do corpo}.
Exemplos:

- Dor em região cervical direita.

B.2 Procedimentos e protocolos

Realizou {procedimento} em {data}.
Exemplos:

- Realizou cirurgia em urgência dia 18/03/2017;
- Realizou ressecção endoscópica;
- Realizou Stent de veia cava com estabilização dos sintomas.

Perguntas:

- Quando o paciente {PacienteNome} realizou o procedimento {procedimento}?

Recebeu {procedimento} dia {data}.

Exemplos:

- Recebeu primeira semana de qt dia 20/04/2017;
- Recebeu 4x AC...

Recebeu {protocolo} em {local} dia {data}.

Exemplos:

- Recebeu primeira linha com FOLFIRINOX em SP dia 27/01/2017...

Perguntas:

- Quando o paciente {PacienteNome} recebeu o protocolo {protocolo} da última vez?
- Quais datas o paciente {PacienteNome} recebeu o protocolo {protocolo}?

Iniciou {protocolo} em {data}.

Exemplos:

- Iniciou Tamoxifeno em 03/2015;
- Iniciou Anastrozol, Eligard, Zometa dia 06/01/2016.

Continuaremos {protocolo} {dose protocolo}

Exemplos:

- Continuaremos Gencitabina 1900mg EC D1,4,15 ciclo 3;
- Continua Tamoxifeno 20mg VO.

Libero {protocolo} {dose protocolo}.

Exemplos:

- Libero Eligard 22,5mg;
- Libero ciclo 27 Fulvestranto + Zoladex + Herceptin + Perjeta;
- Libero ciclo 4 Xeliri + Avastin;
- Libero ciclo 8 Temodal 450mg VO D1 a5.

Libero ciclo {nº ciclo} {protocolo}.

Exemplos:

- Libero ciclo 4 Gemzar;
- Libero ciclo 14 Herceptin e Anastrozol;
- Libero ciclo 5 opdivo.

Perguntas:

- Qual o número do último ciclo do protocolo {protocolo} liberado para o paciente {PacienteNome}?

Solicito inicialmente {protocolo}.

Exemplos:

- Solicito inicialmente Paclitaxel semanal.

Perguntas:

- Quais são os protocolos solicitados para o paciente {PacienteNome}?

Inicio {protocolo} {periodicidade dias} de {intervalo entre ciclos} dias.

Exemplos:

- Inicio Cisplatina e Etoposide D1 a3 de 21/21 dias.

B.3 Medicamentos em uso

Usando: {medicamento} {apresentação medicamento}.

Exemplos:

- Usando: Durogesic 50 - 75mcg.

Perguntas:

- Quais os medicamentos que o paciente {PacienteNome} faz uso?
- Quais pacientes utilizam o medicamento {medicamento}?

Em uso de {medicamento} {apresentação medicamento} {posologia}.

Exemplos:

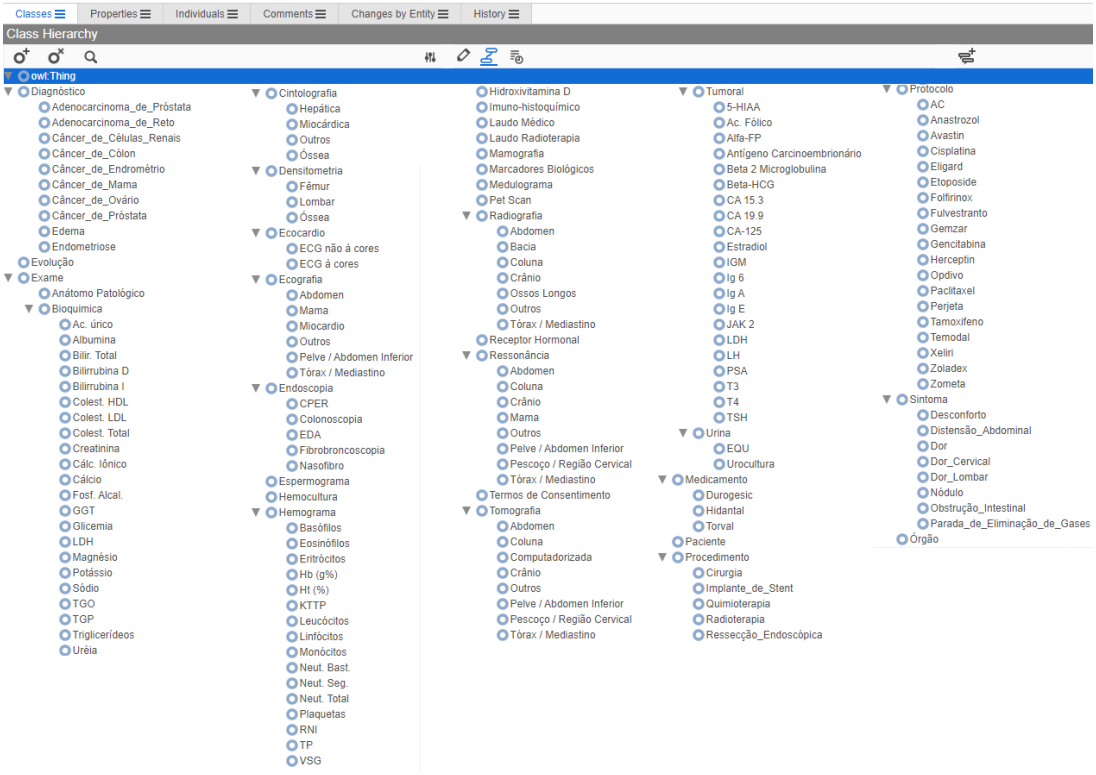
- Em uso de Torval 500mg 12/12h , e hidantal 100mg 2cp de 1212h...

C Ontologia Completa

Essa sessão apresenta a estrutura completa da ontologia. A figura 50 apresenta todas as classes. A figura 51 ilustra as propriedades de objetos. Já a figura 52 exibe as propriedades de dados. Por fim, a figura 53 mostra as instâncias presentes na ontologia.

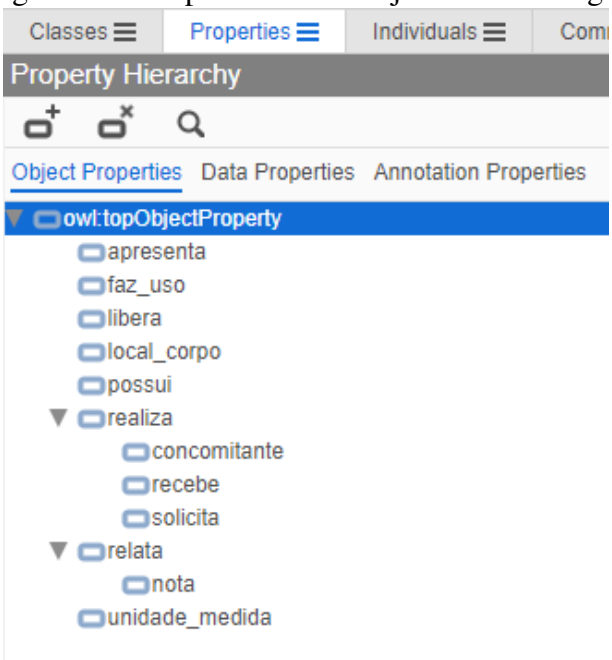
D Evoluções avaliadas

Figura 50 – Classes da Ontologia



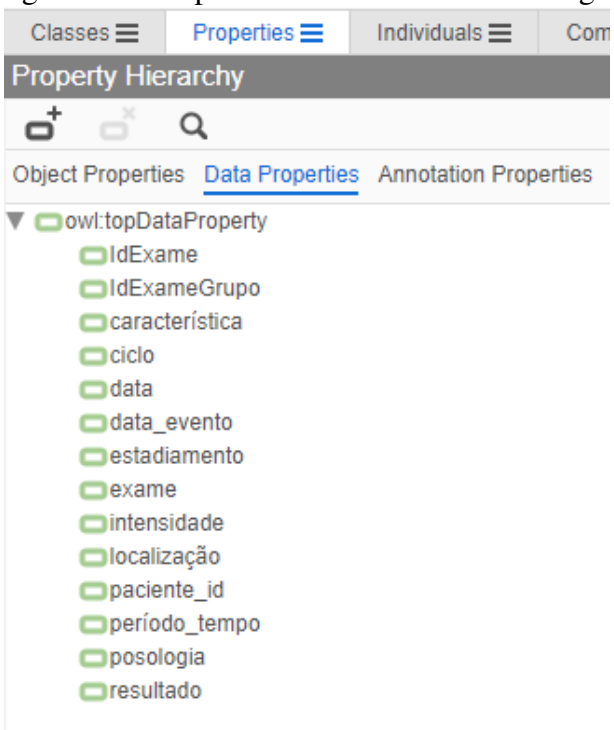
Fonte: Elaborado pelo autor

Figura 51 – Propriedades de Objetos da Ontologia



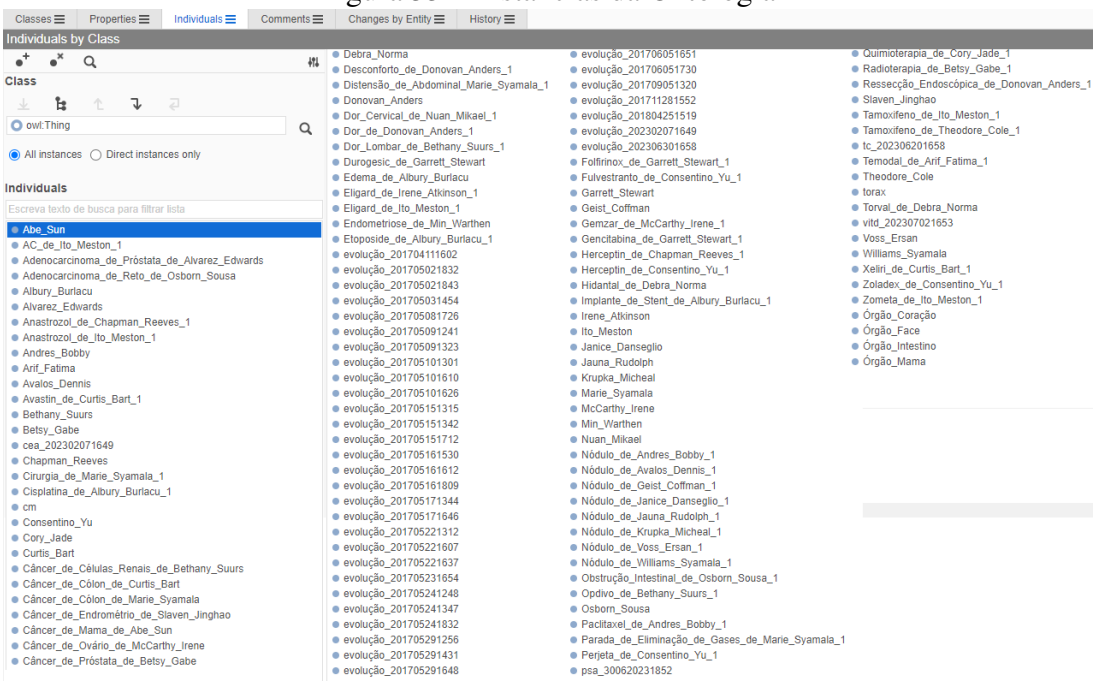
Fonte: Elaborado pelo autor

Figura 52 – Propriedades de Dados da Ontologia



Fonte: Elaborado pelo autor

Figura 53 – Instâncias da Ontologia



Fonte: Elaborado pelo autor

Quadro 8 – Evolução 1

procedente de XXXXXXX 17/03/2015 PSAT 9,52 06/06/2015 prostatectomia radical (Dr XXXXXXX): adenocarcinoma gleason 8, margens livres, sem invasão de vesicual seminal e sem invasão de capsula. pT2pN0M0 RT de 66,6 Gy em leito prostático 06/07 a 02/09/2017 Dr XXXXXX. sem BH. 16/04/2020 GASTRECTOMIA PARCIAL: GIST medindo 2,8CM. Margens livres - IHQ Ki67 5% CD117 positivo. 02/2020 COLONOSCOPIA: adenocarcinoma IN SITU de cólon. 06/10/2017 30/08/2018 12/2018 07/2019 02/2020 10/12/2020 22/04/2021 13/09/2021 24/01/2022 04/2022 30/06/2022 12/2022 PSAT 1,26 » RT » 1,00 > 0,21 0,13 0,05 0,05 0,08 0,12 0,23 0,34 0,47 0,45 0,66 TESTO 362 404,37 Paciente sem queixas no momento

Quadro 9 – Evolução 2

10/03/2017 COLT 204 LDL 140 GLICEMIA 101 CREA 0,92 HMG OK 11/03/2017 CO: baixa probabilidade de lesões ósseas 17/03/2017 us de abdome total: esteatose difusa moderada. rxtx sem alteração 03/2017 rm de pelve: sem sinal de recidiva de doença. 21/02/2019 rxtx sem alteração US de abdome total: esteatose leve. TC de pelve: sem alteração patologica, sem lesão suspeita. 01/2020 COLONO: lesão pólipos de reto e cólon descente, adenoma com displaisa de alto grau, associado a ADENOCARCINOMA IN SITU. EDA: lesão submucosa de 1,5 cm. 01/2020 rxtx sem alteração. US de abdome total com esteatose hepática moderada. 05/08/2020 COLONO: 3 pólipos de sigmoide e reto. 1 pólipo com displasia de alto grau. EDA cicatriz em corpo gástrico e gastrite leve. 02/2021 CO: sem alteração. tctx com calcificações coronarianas e aortica. TC de abdome total: gastrectomia parcial. Prostatectomia. Sem linfonodos e sem lesão suspeita. 16/09/2021 US de abdome total: esteatose hepática leve/moderada. US de tireoide sem alteração. RXTX sem alteração. 16/06/2021 EDA sem alteração. COLONO diverticulos. 15/02/2022 CO sem alteração. 17/02/2022 RM de pelve: sem lesão residual ou recidivante. 16/11/2022 US de abdome com esteatose hepática leve. EDA gastrite leve. Esofagite erosiva grau B.

Quadro 10 – Evolução 3

30/03/2016 PAAF com carcinoma metastático 14/04/2016 IHQ: inconclusivo (pouco material) CO: captação apenas em sacroiliacas bilateral (baixa probabilidade de meta) 04/2016 PET TC: captações de baixa intensidade de SUV em linfonodo supraclavicular esqeurdo, submandibular direito, tonsila, parede abdominal e 1 linfonodo iliaco. (linfoma? apesar do PAAF) 02/08/2017 PSAT 0,04 TGC 232 HEMOGLOBINA GLICADA 5,4% HMG, TSH, T4 LIVRE NORMAIS 08/2017 US DE ABDOME TOTAL: esteatose de grau acentuado. rxtx sem alteração pulmonar, sinais de HAS. 11/2017 CO: sem alteração TGC 300 osteopenia dp -2,3 % 25/05/2018 US de abdome total: cistos hepático e renal, colelitoase com 3 calculos até 0,7 cm, esteatose moderada. US cervical: nódulo benigno em tireoide, linfonodo de 7 mm infraclavicular esquerdo (residual??). 12/06/2019 US de abdome total: esteatose hepática moderada, colelitiase e polipo de vesicula. RXTX normal.

Quadro 11 – Evolução 4

08/2022 RM de abdome total: houve desaparecimento das lesões hepáticas, redução de 80% na lesão em cabeça de pâncreas 1,6 x 1,0 (era 3,6x2,3cm). avascular e sem restrição a difusão. TCTX sem alteração

Quadro 12 – Evolução 5

enc pelo Dr XXXXXXXX 07/2015 ressecção de lesão em MIE: melanoma sem linfonodos positivos. 11/2017 DHL normal 17/10/2017 tctx: graanulomas calcificados em ambos os pulmões. 19/09/2017 US de abdome total: normal. 14/09/2019 RM de abdome com esplenomegalia homogenea. 02/05/2020 TCTX com nódulos residuais calcificados. EF: sem alterações no joelho esquerdo e sem linfonodos palpáveis. Leve edema. cd: solicito TCS + lab

Quadro 13 – Evolução 6

enc pela Dra XXXXXXXX Hepatologista. Hepatite C Hepatocarcinoma por critérios de Barcelona RM com 4 nódulos de 1,2 a 4,5 cm. Alfa feto > 2.000 (03/2022) Paciente nega sangramento 14/05/2022 20/06/2022 08/2022 15/09/2022 07/10/2022 31/10/2022 12/2022 alfa feto 2.430 1.079 986 823 675 PLT 60MIL 80mil 79 MIL 75MIL TSH 2 5 75mil

Quadro 14 – Evolução 7

Câncer de próstata há 15 anos - cirurgia Recidiva local volumosa atual - PSAi: 21,8 Adenocarcinoma de pulmão esquerdo - 2° primário de pulmão cT2NxM0 Cirurgia em 01/11/22 - pT3N0 - EC IIB Cis-Alimta adjuvante desde 12/12/22 Paciente vem para segundo ciclo. Teve algumas intercorrências após os primeiro ciclo. Conta que ficou assintomatico por 4 dias, e no 5° começou constipação importante. Após alguns dias, passou a ter náusea grau 2, sem vômitos. Tomou Bisacodil repetidamente e teve piora do enjoo, como eventual melhora da constipação. Desde então, ficou o resto do intervalo com náusea e pobre aceitação alimentar. Perdeu 7kg. Não teve vômitos em nenhum momento. Na ultima semana, teve diagnóstico de ITU com febre e EAS infeccioso, já em uso de Cipro, com melhoras. Hoje, fraco e abatido, negando febre, náusea ou dores. Hemograma normal. Função renal OK .

Quadro 15 – Evolução 8

enc pela Dra XXXXXX (XXXXXX) 22/10/2018 CDI de mama direita. BRCA 1 e 2 sem mutação 01/11/2018 IHQ RE 90% RP 1% her2 0/3+ KI67 60% 28/05/2019 Mastectomia direita e linfadenectomia: ?? T2N1M0 AC-T semanal 29/11/2018 a 09/05/2019 02/09/2019 RT de mama direita 50GY Dr XXXXXX. TMX com início em 19/06/2019 a 03/10/2022 Taxol + avastin inicio em 10/10/2022 Refere que tolerou bem; 12/03/2020 10/2022 16/11/2022 12/2022 CA 15.3 3,7 116 49 29,7

Quadro 16 – Evolução 9

T2N1M0 - MULTICENTRICO 04/2019 RM de mamas: redução de quase 80 do nódulo biopsiado (11-12) e desaparecimento do linfonodo. 29/08/2019 rxtx sem alteração. US de abdome total: cisto em anexo direito, litiase renal. US de mamas com estabilidade desde RM de 2018. 11/11/2019 TC de tórax: nódulo em mama esquerda 3 cm, micronódulos pulmonares de 3 mm. TC de abdome total: sem alteração em relação ao exame anterior. 20/12/2019 MMG cat 2. 07/01/2020 US de mamas: cat 3. 3,1 retroareolar. 1,3 e 1,0 cm. O nódulo retroareolar não teve alteração. 16/03/2020 CO: baixa probabilidade de meta óssea. rxtx sem alteração. 09/04/2020 US de abdome total normal. 17/03/2020 US de mamas: cat 3, porém sem alteração em relação ao exame anterior e sem alteração em relação a RM de 2018. 25/09/2020 US de abdome sem alteração. USTV com endometrio de 4 mm, mioma de 1,6 cm, foco de endometriose em ovário. US de mamas cat 2. 16/09/2020 RXTX sem alteração. 01/12/2020 TCTX com micronodulo pulmonar de 3mm. TC de abdome total com litiase renal de 2 mm. MMG cat 2. 04/05/2021 us de mamas cat 3 estável. USTV com cisto de 0,5 cm em ovário direito sem vascularização ao doppler. 21/10/2021 tctx sem lesão suspeita. TC de abdome sem lesão suspeita. USTV com cisto ovariano 4,6 ml (endometriose). 28/02/2022 MMG cat 2 US de mamas cat 2 12/09/2022 TCTX com multiplas lesões pulmonares sugestivas de metas randomicas. TC de abdome sem alteração. 20/12/2022 TCTX houve redução em número e tamanho das lesões pulmonares 1,2»0,6cm.

Quadro 17 – Evolução 10

Ca de Próstata localmente avançado PSA 63,2ng/ml Captação em costelas A/E 12/07/2016 CO: sem evidência de lesão secundária. RT até agosto/12 Eligard desde 12/2011 A 12/2014 Eligard 02/08/2016 (recidiva bioquimica) ÚLTIMA EM ABRIL/2017 Estava em tratamento no ITC com quimioterapia. Abiraterona + xgeva iniciado em 02/01/2023 a Paciente sem queixas no momento.

Quadro 18 – Evolução 11

Câncer de reto alto EC IVc - lesões hepáticas + peritoneo Ausência de MSI. CEAi: 66,9 TVP de jugular int dir associato a Port mFolfox + Avastin iniciado em Agosto/21 -> mudado de FU para Capecitabina pela trombose em Nov/21 Mar/22 - suspenso Oxaliplatina por neuropatia - segue Capecitabina + Bevacizumabe -> Progressão (positiva para mutação do KRAS exon 2) FOLFIRI iniciado em Out/22 - Adaptado para Irino-FU-LV semanal (D1,8,15,22 a cada 42 dias) em Dez/22 por perda de cateter recorrente Paciente vem hoje para continuidade de tratamento, mantendo fraqueza e hiporexia como únicas queixas. Mantem anemia estável. Iniciou Noripurum semanal na ultima semana. Hb 8,5 Leuco e plaquetas normais.

Quadro 19 – Evolução 12

Pcte da Dra XXXXXX Veio devido a US com nódulo em mama esquerda medindo 6,3 cm. Ao exame físico não consigo definir o nódulo. 16/12/2022 RM de mamas: cat 3, pequenos cistos e nodulos infracentimétricos com curva normal, 1. cd: novo US DE MAMA S EM 3 MESES antes se necessário

Quadro 20 – Evolução 13

ENC PELO DR XXXXXX 05/05/2022 BIOPSIA DE RETO: ADENOCARCINOMA. COLONO LESÃO DE 2 A 10 CM DA BA. 5FU / LV D1A D5 + RT INICIADO EM 23/05/2022 A 08/07/2022 PACIENTE REFERE ÓTIMA TOLERÂNCIA COM A QUIMIOTERAPIA

Quadro 21 – Evolução 14

Enc pelo Dr XXXXXXX 16/09/2021 Adenocarcinoma de reto operada em urgência (obstrução??) pT3pNxMx FOLFOX ADJUVANTE INICIADO EM 26/10/2021 A 05/04/2022 AUSÊNCIA DE INSTABILIDADE DE MICROSSATELITE KRAS/NRAS MUTADO BRAF SEM MUTAÇÕES LAPAROTOMIA COM IMPLANTES PRÓXIMO A COLOSCOMIA 5FU infusional 11/07/2022 a 19/09/2022 FOLFIRI + AVASTIN INICIADO 25/10/2022 A Paciente refere estar bem. 07/03/2022 07/2022 09/2022 30/12/2022 CEA 0,7 0,8 2,39 1,41

Quadro 22 – Evolução 15

18/08/2014 tc de abdome total: sem alteração 05/11/2014 US de abdome total: sem alteração 05/11/2014 RX de TX: sem alteração 19/05/2015 Colonoscopia: LST granular homogênea no ângulo hepático (mucosectomia - polipo hiperplasico), lesões planas elevadas em mucosa retal - polipectomia (adenoma tubular com displaisa de alto grau) 16/06/2015 rx de tx e us de abdome: sem lesão 29/06/2017 CEA 0,91 psat 0,44 19/06/2017 colono: sem alteração. 04/07/2017 tx de tx e abdome: nódulos inespecificos pulmonares e enfisema, baço acessório. 06/03/2018 rxtx sem alteração US de abdome total: esplenectomia prévia, sem mais alteração. CEA 1,2 22/07/2019 EDA: gastrite erosiva plana leve. COLONO com diverticulos e sem pólipos. Anastomose a 18 cm da BA. 09/2021 tctx com massa perihilar direita medindo 6,1 cm com linfonodos hilares e paratraqueais, nódulo subpleural em LSD. 12/2021 RM de coxa direita com lesão intramedular (secundária). RM de crânio : sem lesão. 01/02/2022 EDA: ESOFAGITE LEVE E GASTRITE LEVE, SEM LESÃO. 04/08/2022 US com doppler de MMII : TVP em em poplitea e gastrocnemia mediais esquerda. 04/08/2022 TCTX com nódulos pulmonares estável 1 cm e houve redução de lesão hilar 5 > 4,1 cm e dos linfonodos mediastinais.

Quadro 23 – Evolução 16

enc pelo Dr XXXXXX. 06/08/2021 COLONOSCOPIA: 01/09/2021 COLECTOMIA (DR XXXXXXX): SEM INSTABILIDADE DE MICROSSATELITE FOLFOX EM 1ª LINHA 10 CICLOS HEPATECTOMIA (METASTASECTOMIA): FOLFOX PÓS CIRURGIA 11ª CICLO (DIA 12/04/2022). DIAGNÓSTICO COM META EM FÍGADO. FOLFIRI + AVASTIN INICIADO EM 26/04/2022 A 19/12/2022 FOLFOX + CETUXIMABE iniciado em Paciente refere boa tolerância. CEA: 1,8 » 1,99 (09/2022) > 2,3 (10/2022) 9,99 (16/12/2022).

Quadro 24 – Evolução 17

30/06/2022 tctx e abdome: houve redução das lesões hepáticas e nódulos pulmonares. 10/10/2022 tctx micronodulos pulmonares TC de abdome comm lesão hipodensa pós cirurgica de 5x3,0 e 1,7 cm. Doença estável. 02/01/2023 TCTX com aumento de nódulos pulmonares ate 1,1cm. TC de abdome com aumento de lesões hepáticas 3,1cm).

Quadro 25 – Evolução 18

enc pelo Dr XXXXXX Mãe da paciente XXXXX (ca de colo uterino) Ca de mama direita TRIPLO NEGATIVO com KI67 40%. biopsia de mama contralateral sem malignidade. T1N0M0 medindo 2 cm> qt neo> QDT de mama D com RPC. Neoadjuvância com AC-carbo/taxol semanal ATÉ 28/03/2021 18/03/2021 QDT + LS sem doença residual. RT de mama direita 20/07 a 16/08/2021 Dra XXXXXX Paciente 08/2022 tsh 1,2 t4 1 tireo 20 (vr até 30)

Quadro 26 – Evolução 19

04/2019 us de abdome total com esteatose leve e cisto renal de 3,96cm RXTX sem alteração (visto apenas imagem).

Quadro 27 – Evolução 20

Padre em XXXXXX 18/06/2018 Prostatectomia + linfadenectomia : adenocarcinoma gleason 8, 1/9 LFN, margens livres e sem extravazamento de capsula. ZOLADEX INCIADO EM 10/2018 A 04/11/2020 RT 76 GY COM DR XXXXXXX PSAT INICIAL 18 PSAT 24/08/2018 0,14 » PSAT 0,01 (04/2019)» PSAT 0,06 (03/11/2020). TSH 1,4 (03/11/2020). Paciente sem queixas no momento

Quadro 28 – Evolução 21

14/03/2019 densitometria óssea: - 2,8 DP em coluna. 12/04/2019 Colono: sem alteração EDA: esofagite com duas lesões menores que 0,5 cm. 04/2019 rxtx sem alteração US de abdome total: polipo de colesterol em vesicula com 4,4mm. VLC normal (Dergan). 08/2020 VLC com refluxo faringo laringeo. US cervical sem linfonodos suspeitos. US de próstata com 52 g. US de abdome com pólipo de 7mm. RXTX sem alteração. 15/02/2022 COLONO: PÓLIPO RESSECADO EM CÓLON ASCENDENTE. EDA : ESOFAGITE REGENERATIVA. RXTX SEM ALTERAÇÃO. US DE TIREOIDE COM NÓDULO DE 0,5 CM. US DE ABDOME COM PRTOSTA DE 46 G 10/05/2022 DENSITOMETRIA ÓSSEA DP -2,7 L1-L2 OSTEOPOROSE

Quadro 29 – Evolução 22

adenocarcinoma de reto medio tratado com cirurgia 26/12/2018 ap adenocarcinoma tubular g2 ial epn - linf 0/11 xelox 7x ate 01/07/2019 dpoc d alzheimer- donila duo 10/20 sem sintomas cea 252 vit d 249 seguimento oncologico ad til.