UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília

NATHÁLIA ADRIELE DE LIMA

RECUPERAÇÃO DA INFORMAÇÃO DE PRONTUÁRIOS ELETRÔNICOS:

Um modelo de visualização de informação de medicamentos.

Marília
2023

Nathália Adriele de Lima

RECUPERAÇÃO DA INFORMAÇÃO DE PRONTUÁRIOS ELETRÔNICOS:

Um modelo de visualização de informação de medicamentos.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação como parte das exigências para a obtenção do título de Mestre em Ciência da Informação pela Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (UNESP), Campus de Marília.

Orientador: Prof. Dr. Leonardo Castro Botega

Área de Concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Marília
2023

L732r

Lima, Nathalia Adriele de
Recuperação da Informação de Prontuários Eletrônicos : Um modelo de visualização de informação de medicamentos / Nathalia Adriele de Lima. -- Marília, 2023
145 p.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília
Orientador: Leonardo Castro Botega

1. Recuperação e visualização da informação. 2. Prontuários eletrônicos do paciente. 3. Informações de medicamentos. 4. PLN. 5. Grafos de conhecimento. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Nathália Adriele de Lima

RECUPERAÇÃO DA INFORMAÇÃO DE PRONTUÁRIOS ELETRÔNICOS:
Um modelo de visualização de informação de medicamentos.

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), como requisito parcial para a obtenção do título de Mestre em Ciência da Informação.

Área de Concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Banca Examinadora

Prof. Dr. Leonardo Castro Botega (Orientador)
Universidade Estadual Paulista (UNESP) – Faculdade de Filosofia e Ciências de Marília-SP.

Prof. Dr. Danilo Medeiros Eler
Universidade Estadual Paulista (UNESP) – Faculdade de Ciências e Tecnologia (FCT) – Câmpus de Presidente Prudente.

Prof. Dr. Ivan Luiz Marques Ricarte
Universidade Estadual de Campinas (UNICAMP)

Marília, 3 de julho de 2023.

Dedico esta pesquisa à minha amada família, em especial ao meu querido tio e incentivador número um Rogério, por seu apoio inestimável e inspiração constante.

AGRADECIMENTOS

A Deus, a minha tão preciosa e abençoada família pelo incentivo, apoio e companheirismo em todos os momentos.

Ao meu orientador Prof. Dr. Leonardo Castro Botega pelas orientações e intervenções precisas durante todo o desenvolvimento desta pesquisa.

A todos do Programa de Pós Graduação em Ciência da Informação da Universidade Estadual Paulista “Júlio de Mesquita Filho” - UNESP.

Ao grupo de pesquisa GIHC e ao grupo HAIS do Hospital das Clínicas de Marília pelo direcionamento e parceria.

“La véritable éducation consiste à pousser les gens à penser par eux-même.” A verdadeira educação consiste em levar as pessoas a pensar por si mesmas.
Noam Chomsky

RESUMO

O Prontuário Eletrônico do Paciente (PEP) é um padrão de software que moderniza a coleta, armazenamento e recuperação de informações médicas dos pacientes, incluindo dados de medicamentos. No entanto, a forma como as informações estão estruturadas não favorece a recuperação, pois a presença massiva de textos não estruturados nos prontuários médicos pode dificultar a recuperação de dados, afetando o tempo, a precisão e o entendimento dos dados. Pesquisas são necessárias para aprimorar a qualidade e a rapidez da recuperação de informações médicas, considerando o volume e a complexidade dos dados não estruturados. Este estudo reconhece a importância do entendimento desses dados de medicamentos por profissionais de saúde, e propõe um modelo de recuperação e visualização de informações de medicamentos dos prontuários eletrônicos, utilizando Processamento de Linguagem Natural (PLN) e Grafos de Conhecimento. Entende-se que abordagens que utilizam esses métodos e ferramentas têm mostrado resultados promissores na recuperação e visualização de informações de medicamentos. Esta pesquisa destaca a necessidade de aprimorar a recuperação e integração de informações dos prontuários eletrônicos, ressaltando a importância desses dados, em especial informações de medicamentos. Modelos de PLN baseados em aprendizado profundo e o uso de grafos de conhecimento podem extrair, agrupar e vincular informações relevantes, melhorando a recuperação e visualização dos dados de medicamentos nos prontuários. Conclui-se que é fundamental continuar aprimorando os prontuários eletrônicos, e o uso de PLN e grafos de conhecimento pode contribuir para a recuperação e visualização de variadas informações médicas/sociais de pacientes, melhorando o sistema de saúde pública.

Palavras–chave: Recuperação e visualização da informação; Prontuários eletrônicos do paciente; Informação de medicamentos; Processamento de linguagem natural; Grafos de conhecimento.

ABSTRACT

The Electronic Health Record (EHR) is a software standard that modernizes the collection, storage, and retrieval of patients' medical information, including medication data. However, the way the information is structured does not favor retrieval, as the massive presence of unstructured text in medical records can hinder data retrieval, affecting the time, accuracy, and understanding of the data. Research is needed to improve the quality and speed of medical information retrieval, considering the volume and complexity of unstructured data. This study recognizes the importance of understanding medication data by healthcare professionals and proposes a model for retrieving and visualizing medication information from electronic health records using Natural Language Processing (NLP) and Knowledge Graphs. It is understood that approaches using these methods and tools have shown promising results in the retrieval and visualization of medication information. This research highlights the need to improve the retrieval and integration of information from electronic health records, emphasizing the importance of this data, particularly medication information. NLP models based on deep learning and the use of knowledge graphs can extract, group, and link relevant information, improving the retrieval and visualization of medication data in health records. It is concluded that it is essential to continue improving electronic health records, and the use of NLP and knowledge graphs can contribute to the retrieval and visualization of various medical/social information of patients, thereby improving the public healthcare system.

Keywords: Retrieval and visualization of information; Electronic patient records; Medication information; Natural language processing; Knowledge graphs.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PEP	Prontuário Eletrônico do Paciente
PLN	Processamento de Linguagem Natural
NLP	Natural Language Processing
HCM	Hospital das Clínicas de Marília
BRAPCI	Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação
GIHC	Grupo de Interação Humano-Computador
HAIS	Health Artificial Intelligence Study
ANVISA	Agência Nacional de Vigilância Sanitária
CSV	Comma-Separated Values
CFM	Conselho Federal de Medicina
SOC	Sistema de Organização do Conhecimento
CI	Ciência da Informação
RDF	Resource Description Framework
OWL	Web Ontology Language
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-Training Transformer
NER	Named Entity Recognition
REL	Extração de Relações de Entidades
API	Application Programming Interface
IOB	Inside–Outside–Beginning
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks

SUMÁRIO

1 INTRODUÇÃO

O Prontuário Eletrônico do Paciente, muitas vezes denominado simplesmente como PEP, é, essencialmente, um modelo de padronização de software que foi idealizado e desenvolvido visando modernizar os processos de coleta, armazenamento e recuperação de dados e informações médicas de pacientes. A adoção do prontuário eletrônico busca a otimização dos processos de gestão da informação relacionada a identificação e atendimento clínico/social dos pacientes, por meio da digitalização, integração e interoperabilidade dos sistemas de informação de saúde. Tanto da rede pública de saúde, como da rede privada. Com essa prerrogativa o prontuário eletrônico do paciente, atualmente, constitui-se em um sistema imprescindível para assegurar uma assistência integral e continuada aos pacientes, possibilitando um melhor acesso às informações e ao perfil histórico da saúde destes usuários.

No entanto, embora a digitalização dos prontuários eletrônicos seja uma inovação importante no campo da tecnologia da informação aplicada à saúde, esta pesquisa identificou que a mera existência e utilização dessa tecnologia não melhora automaticamente a qualidade dos processos de recuperação da informação. Existem limitações na recuperação da informação nesses sistemas que dificultam a recuperação de dados específicos e importantes. Como pontua Benício (2020) que, indica que limitações relacionadas à recuperação da informação nesses sistemas emergem por conta dos textos não-estruturados, derivados de anotações de textos livres que, invariavelmente, criam dificuldades para se recuperar esses dados. Assim, tanto humanos quanto sistemas enfrentam desafios nessa tarefa complexa e custosa.

Os dados em formato de textos livres, ou seja, não-estruturados, gerados a partir de anamneses, relatórios de diagnóstico, prescrição médica e de receituário, evidentemente, compõem a maior porção percentual dos dados armazenados em sistemas de prontuários eletrônicos. Estes textos livres expressos em linguagem natural apresentam desafios relacionados à recuperação e utilização da informação, em oposição a fonte de dados já estruturada e pronta para uso ou consulta. Estas anotações não estruturadas são naturalmente de difícil recuperação para profissionais médicos, que obrigatoriamente têm que entender rapidamente a situação geral do atendimento do paciente, para a melhor tomada de decisão.

Esta dificuldade, deve-se ao grande volume de dados e a problemas naturais de construção de textos livres, como: os erros comuns de digitação, ortográficos e gramaticais, e o uso de sinônimos, abreviações e termos específicos, que dificultam a recuperação otimizada dessas informações por parte dos profissionais. Neste contexto, conforme Kormilitzina *et al*. (2020), a tarefa de recuperação de dados não-estruturados também é desafiadora para sistemas

de recuperação de informação, que por conta da falta de estruturação, homogeneidade e da própria natureza constituinte destes dados, encontram dificuldades ao precisar extrair informações inerentes de frases, sentenças e transcrições médicas. Desta forma, infere-se que uma forma eficaz para melhorar a recuperação e visualização da informação em referidos sistemas é usando mecanismos para a extração e estruturação desses dados de textos livres.

Ainda neste contexto de desafios há, especialmente, os problemas ligados a recuperação de informações sobre medicamentos em textos não estruturados e semiestruturados de receituários eletrônicos. Sendo que, um dos principais problemas relacionados à recuperação de informações sobre medicamentos em prontuários eletrônicos é a falta de padronização na forma como as informações são registradas neste tipo de formato. Onde diferentes profissionais de saúde podem usar terminologias ou abreviações distintas ao documentar uma prescrição de medicamentos. Também é comum um profissional mencionar o nome comercial ou de referência de um medicamento, enquanto outro utiliza o nome genérico ou similar ou até mesmo o princípio ativo, o que torna complexa a busca e a recuperação precisa dos dados. Além disso, as informações podem ser inseridas de maneira inconsistente, com variações na grafia, na formatação ou até mesmo na ordem dos dados, bem como incorreções gramaticais e presença de siglas médicas e jargões (Carvalho, 2018). O que torna ainda mais desafiador o processo de recuperação de informações específicas sobre medicamentos.

A falta de estruturação adequada dos dados relacionados a medicamentos nos prontuários eletrônicos dificulta a agilidade e eficácia na extração e no rápido entendimento de informações relevantes. De forma geral, as informações são registradas sem uma estrutura clara ou separação em campos específicos. Informações importantes, como nome do medicamento, princípio ativo, dosagens, horários e via de administração, entre outros, podem estar dispersas em diferentes partes do prontuário. Isso torna a recuperação de informações específicas e relacionadas uma tarefa complexa e demorada.

Para solucionar esses problemas, conforme é indicado na revisão de literatura desta pesquisa, estão sendo desenvolvidas soluções baseadas em tecnologias de aprendizado de máquina. Essas soluções visam melhorar a recuperação de informações sobre medicamentos em textos não estruturados, por meio de técnicas como extração de entidades, normalização de terminologias e detecção de relações entre termos. No entanto, ainda há desafios a serem superados para alcançar uma recuperação automatizada, precisa e confiável das informações sobre medicamentos nos prontuários eletrônicos.

Logo, a recuperação de informações sobre medicamentos em textos não estruturados nos prontuários eletrônicos apresenta uma série de desafios, tais como a identificação correta

de termos médicos, a normalização de dados, a extração de informações relevantes, a resolução de ambiguidades e a integração de diferentes fontes de informações em uma única base de dados centralizada. A falta de padronização, a inadequada estruturação dos dados e a falta de interoperabilidade entre sistemas são obstáculos que dificultam a coleta de informações específicas, exigindo assim melhorias focadas na recuperação desses dados.

Para isso, existem inúmeras técnicas e ferramentas que podem ser utilizadas. Dentre essas técnicas, atualmente destaca-se como principal o Processamento de Linguagem Natural (PLN), especificamente a tarefa de Reconhecimento de Entidades Nomeadas (NER) essencial para transformar os dados de textos livres em informações úteis e acessíveis, qualificadas e estruturadas. De acordo com Jagannatha (2018), tarefas de PNL como NER podem identificar medicamentos e seus atributos (dosagem, método, duração e frequência), indicações e eventos adversos relacionados a medicamentos. A tarefa de identificação de relações entre entidades pode detectar relacionamentos entre as entidades nomeadas: relacionamentos entre medicamento e indicação, e relacionamentos entre atributos.

Em conjunto com as tarefas de PLN destaca-se a criação e a incorporação de Grafos de Conhecimento, que fornece uma maneira mais interpretável e eficiente para desbloquear e armazenar conhecimentos heterogêneos, permitindo uma fácil visualização e interpretação humana e de sistemas de recuperação e de indexação de informação. Conforme Geleta1 et al. (2021) atualmente, os grafos de conhecimento desempenham um papel fundamental na integração e indexação de dados, oferecendo uma estrutura de representação comum que permite um modo mais compreensível e eficaz para consultar diversas fontes de dados.

Os grafos de conhecimento surgiram com o objetivo de abstrair determinado grupo de informação, objetivando organizar o conhecimento de forma estruturada de modo a integrar informações extraídas de variadas fontes de dados, criando visualizações facilmente acessíveis a humanos e máquinas. Assim, os grafos de conhecimento têm sido usados para construção de sistemas voltados à vinculação de entidades promovendo a formação de bases robustas de conhecimento que contribuem para a recuperação e visualização da informação, promovendo uma maior, melhor e mais rápida compreensão do usuário final da informação. Logo, para a recuperação e visualização de informações contidas em textos não estruturados e/ou semiestruturados, os sistemas de informação dependem de métodos específicos para tornar esses dados e informações mais inteligíveis para humanos e formalizados para a construção de sistemas automatizados de recuperação de informação.

Especialmente, esta pesquisa teve como foco principal a construção de um modelo que possibilita a recuperação e visualização de informações de medicamentos contidas em textos

não-estruturados e/ou semiestruturados presentes em prontuários eletrônicos e listas de medicamentos. Especificamente, para isso, esse modelo que faz uso de técnicas de Processamento de Linguagem Natural (PLN) e Grafos de Conhecimento para extrair, organizar e estruturar de forma eficiente informações especificas de medicamentos.

O modelo identificará entidades nomeadas relacionadas a medicamentos e empregará Grafos de Conhecimento para estruturar e estabelecer relações entre elas. Um banco de dados gráfico hospedará uma base de conhecimento estruturada, permitindo uma recuperação mais rápida e precisa das informações dos pacientes, com o objetivo de facilitar o acesso a dados específicos. Dessa forma, este modelo visa fornecer um ponto de partida para extrair outras informações presentes nos registros eletrônicos. Considera-se que, embora a pesquisa esteja concentrada na recuperação de dados sobre medicamentos, neste cenário específico, o desafio principal permanece constante e inalterado, isso tratando de medicamentos em específico ou informações de terapias em geral. A essência do problema reside na habilidade de coletar informações pertinentes a partir de dados não estruturados.

O estudo procura fornecer soluções atualizadas e eficazes para recuperar e exibir informações específicas de medicamentos em textos não estruturados e semiestruturados, como prescrições eletrônicas e listas de medicamentos. O objetivo é auxiliar no acesso rápido a informações relevantes para o tratamento adequado dos pacientes, contribuindo para decisões mais fundamentadas e eficientes.

Da mesma forma, esta pesquisa, sendo dirigida por um estudo de caso, visa por meio da disponibilização de uma base de conhecimento estruturada em grafos, para contribuir com o desenvolvimento de um sistema de recomendação de medicamentos similares que está sendo idealizado e desenvolvido pela equipe de TI do HCM, especificamente pelo grupo de estudos HAIS, pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília.

A ideia principal deste sistema é utilizar o conhecimento e as informações presentes em um banco de dados gráfico para habilitar a construção e aprimoramento de um sistema de recomendação de medicamentos. Tal sistema deve identificar e sugerir automaticamente medicamentos similares, ou seja, aqueles que possuem o mesmo princípio ativo do medicamento referência. Isso permite apresentar ao médico alternativas discretas de medicamentos similares e, assim, mais econômicos, que podem ser prescritos aos pacientes. Dessa forma, a elaboração de um modelo para obter informações de medicamentos e o desenvolvimento de uma base de conhecimento em grafos pode auxiliar na criação de sistemas de informação mais eficientes e precisos na área da saúde.

1.1 Problema de Pesquisa

Adotar tecnologias informacionais emergentes, como os prontuários eletrônicos, é uma forma objetiva de promover a inovação. O prontuário no formato eletrônico digitaliza e automatiza os processos trazendo melhorias na prestação de serviços médicos ao paciente, e, gera novas oportunidades de uso e aplicação da informação. No entanto, a incorporação de novas tecnologias da informação traz consigo novos desafios, principalmente no que se refere à gestão e recuperação do grande volume de dados gerados e armazenados nesses sistemas. Sendo necessário, o desenvolvimento constante de novos métodos e modelos para otimizar a recuperação e a representação da informação e do conhecimento contido nestes.

Essa nova realidade impõe desafios urgentes a pesquisadores dos mais variados campos do conhecimento. Em especial aos interessados na pesquisa e proposição de soluções voltadas para a recuperação, representação e visualização da informação. Posto que, concomitantemente, com a adesão crescente dos prontuários eletrônicos surgiram novas possibilidades para sanar uma antiga demanda para agilizar e melhorar a recuperação de informações clínicas referentes ao real estado de saúde e histórico geral do paciente.

Sabe-se que a qualidade e a rapidez do atendimento médico têm de modo crescente sido afetadas, principalmente, devido ao grande volume de dados não estruturados, como anamneses e prescrições, gerados e armazenados diariamente em sistemas de prontuários eletrônicos. O que, inevitavelmente, tem tornado inviável a recuperação manual dessas informações, causando prejuízos significativos na qualidade e na rapidez do atendimento médico, além de causar estresse desnecessário aos profissionais de saúde.

Considerando as diversas oportunidades de melhorias e, as demandas surgidas com a adoção dos prontuários eletrônicos. O problema desta pesquisa se encontra na necessidade de promover de modo eficiente a recuperação de informações relevantes de medicamentos, contidas em anotações de textos não estruturados de receituários gerados em prontuários eletrônicos dos pacientes. A recuperação de informações específicas de medicamentos se refere à capacidade de encontrar e fornecer informações precisas e relevantes sobre medicamentos, incluindo suas características, propriedades, doses e outras informações relevantes. Essa informação é essencial para profissionais da área da saúde.

Esse é um problema bastante complexo e atual, e que tem se constituído objeto de estudo de diversos pesquisadores na última década. As dificuldades de promover de modo eficiente a recuperação de informações relevantes contidas em anotações de textos não estruturados de receituários gerados em prontuários eletrônicos é atribuída principalmente ao

grande e crescente volume de informações não estruturadas e a falta de um padrão de representação de dados. O que tornou inviável a recuperação puramente manual dessas informações, causando prejuízos significativos na qualidade e na rapidez do atendimento médico, além de modo geral causado estresse desnecessário aos profissionais de saúde.

É consenso que para solucionar esse problema é necessário empregar métodos, técnicas e tecnologias atuais e automatizadas para a recuperação, representação e visualização da informação. Essas devem ser capazes de lidar com o grande volume de dados gerados e armazenados nesses sistemas, além de serem capazes de extrair informações relevantes de medicamentos contidas em anotações de textos não estruturados de receituários.

Logo a pesquisa nesse campo é de extrema importância, pois a qualidade e a rapidez do atendimento médico têm sido afetadas devido ao volume e complexidade dos dados não estruturados gerados e armazenados diariamente em sistemas de prontuários eletrônicos.

Nesse ponto, é fundamental reforçar que, mesmo que este estudo esteja focado na recuperação de informações sobre medicamentos em um contexto específico, o desafio central permanece o mesmo, independentemente de abordar a recuperação de informações de medicamentos em específico ou o tratamento como um todo, ou mesmo, outro tipo de informação. A essência do problema reside na capacidade de extrair informações relevantes de dados não estruturados. Tendo a pesquisa, o propósito de indicar soluções atualizadas e eficientes para recuperar e visualizar informações específicas a partir de textos não estruturados e/ou parcialmente estruturados, como receituários eletrônicos e listas de medicamentos. Para auxiliar os profissionais de saúde a acessar rapidamente informações relevantes no contexto de atendimento dos pacientes. Deste modo o desafio central consiste em extrair informações valiosas de dados não estruturados, independentemente do contexto, e desenvolver ou aplicar métodos e técnicas para mapear e extrair dados de referidos textos, de modo eficiente.

1.2 Objetivos

O objetivo geral desta pesquisa é desenvolver um modelo para recuperação e visualização de informações de medicamentos descritos nos Prontuários Eletrônicos do Pacientes (PEP), com o objetivo de aprimorar a recuperação de dados não estruturados presentes nesses sistemas. Para alcançar isso, serão empregadas técnicas de Processamento de Linguagem Natural (PLN), especialmente a Extração de Entidades Nomeadas (NER), e de Grafos de Conhecimento para estruturar e agrupar as informações.

Deste modo, com o intuito de se alcançar a meta primeva da pesquisa delineou-se alguns objetivos específicos que são destacados a seguir:

- Empreender, por meio de uma revisão de literatura uma investigação do conhecimento atual acerca dos prontuários eletrônicos e, da gestão e organização da informação e do conhecimento contido nesses. Bem como, buscou-se com a pesquisa promover um levantamento de métodos, técnicas e tecnologias para a recuperação e vinculação da informação e do conhecimento.
- Realizar pesquisas, investigações e experimentos relacionadas a recuperação de dados de medicamentos, e apresentar os resultados com o objetivo de avaliar a viabilidade de um modelo (incluindo processos, métodos e técnicas) para recuperação, estruturação e visualização de informações de medicamentos presentes em textos não estruturados, coletados e armazenados em prontuários eletrônicos dos pacientes.
- Apresentar a aplicação de um modelo que extrai, agrupa e vincula informações de medicamentos em receituários médicos por meio do mapeamento de entidades nomeadas e seus relacionamentos. Em complemento, será apresentada uma prova de conceito que ilustrará o emprego de métodos e tarefas de PLN e Grafos de Conhecimento para recuperar e visualizar informações de anotações textuais não estruturadas de medicamentos, melhorando a eficiência na recuperação de informações de prontuários eletrônicos.
- Implementar e promover a disponibilização de uma base de conhecimento (uma coleção organizada e estruturada de informações) de medicamentos em um banco de dados gráficos. E, desta forma, gerar recursos de visualização e navegação para permitir a recuperação e exploração de informações de medicamentos de maneira eficaz, escalável, precisa e coerente.

1.3 Justificativa

À medida que cresce o volume dos atendimentos e da prestação de serviços médicos, se desenvolve, a necessidade para melhorar a agilidade e a qualidade dos serviços de saúde prestados. E, em um cenário de recursos financeiros limitados, faz-se imprescindível pesquisas que busquem meios que efetivamente contribuam com a melhora na prestação dos serviços e do fluxo de trabalho dos profissionais médicos. Nesse cenário, o advento de novas tecnologias da informação como os prontuários eletrônicos, que cada vez mais, estão sendo implementados nas instituições de saúde pública, contribuem para uma significativa melhora na coleta de

dados, digitalizando boa parte do processo de coleta e armazenamento de variadas informações dos pacientes.

Entretanto, a adoção do prontuário eletrônico do paciente fez emergir novas demandas e possibilidades, e, como anteriormente pontuado, por Benício (2020), uma das mais importantes demandas é a da melhoria na recuperação e representação da informação e do conhecimento gerado e contido nestes sistemas. Isso porque, a maioria das informações coletadas não estão estruturadas, posto que estas são coletadas em textos livres. Assim, de acordo com Souza e Almeida (2019), estas informações que são coletadas de forma puramente textuais, ou seja, em um formato de dados não-estruturados, verdadeiramente se constitui em uma barreira que naturalmente gera dificuldades e prejudica a recuperação de informação clínica de pacientes. Evidentemente, isso impacta fortemente na agilidade e na qualidade dos serviços médicos prestados.

Portanto, existe uma crescente e evidente necessidade do desenvolvimento de modelos, métodos, e/ou sistemas que otimizem o processamento e a recuperação desses dados. Nesse sentido, conforme Savova *et al* (2019) deve-se lançar mão ao máximo dos avanços recentes da tecnologia da informação "[...] para usar totalmente a riqueza dos dados armazenados que se acumulam rapidamente em prontuários eletrônicos." Nesse esforço, é possível afirmar que esta pesquisa se justifica, pois contribui com o uso de tarefas de Processamento de Linguagem Natural e de representação por Grafos de Conhecimento, para a recuperação da informação de medicamentos descritos em prontuários, por meio da construção de visualizações.

Posto isso, diante da natureza interdisciplinar e colaborativa que caracteriza a Ciência da Informação entende-se que ainda há infindas contribuições necessárias a serem feitas por pesquisadores desse campo do conhecimento visando promover melhorias na recuperação e visualização da informação e do conhecimento nos sistemas de prontuários eletrônicos. Assim, consequentemente, a CI pode promover pesquisas que colaboraram decisivamente com o aumento da qualidade do serviço de saúde, que é uma área essencial e, requer melhorias constantes no suporte info-tecnológico, para melhoria da força de trabalho de atendimento médico.

1.4 Procedimentos Metodológicos e Ferramentas de Coleta

A presente pesquisa científica se caracteriza como de natureza teórica e aplicada, e seu desenvolvimento está fundamentado em métodos e processos metodológicos que foram essenciais à coleta, análise e interpretação de dados de medicamentos de referências e similares,

bem como, de modelos de anotações textuais advindas de prontuários eletrônicos, que foram fornecidos pelo Hospital das Clínicas de Marília.

Tais métodos, proporcionaram, efetivamente, a correlação entre o embasamento teórico e o desenvolvimento técnico-científico-aplicável da pesquisa, que visou, como pontuado, a recuperação e visualização da informação de medicamentos descritos em receituários médicos eletrônicos.

Posto isso, esta pesquisa de mestrado se destaca fundamentalmente, como: de natureza qualitativa, descritiva e aplicada, e tem uma forte característica bibliográfica exploratória, sendo orientada por um estudo de caso.

Conforme Freitas (2000), existe uma tendência que aponta que, para se chegar a uma solução mais objetiva de um problema de pesquisa, deve-se proceder uma atividade mais subjetiva, qualitativa, possibilitando um melhor delineamento do escopo e do foco do objeto de estudo.

Desta forma, visando alcançar os objetivos propostos, aplicou-se inicialmente uma abordagem qualitativa para a análise dos dados de medicamento e anotações de receituário médico em um esforço para entender como organizar tais dados e, encontrar maneiras eficientes de extrair informações e construir visualizações a partir destas.

Realizou-se também uma revisão bibliográfica visando analisar, identificar e inferir as possíveis contribuições necessárias do campo da Ciência da Informação para o aprimoramento dos prontuários eletrônicos e das técnicas e métodos emergentes para a recuperação, representação e visualização da informação e do conhecimento. Bem como, construir uma base forte e aprofundada de conhecimentos e saberes relacionados aos principais temas tratados neste trabalho de pesquisa científica.

1.4.1 Características da Pesquisa

Esta pesquisa também se caracteriza, como descritiva e bibliográfica, pois descreve inúmeras características técnicas e conceituais dos objetos investigados, conforme afirma Vergara (1998), “a pesquisa descritiva expõe características do objeto de estudo ou de determinado fenômeno” o que “contribui efetivamente para a construção de uma base para tal explicação.” E de característica bibliográfica pois foi fundamentada em referenciais teóricos e análise e revisão bibliográfica, efetivada por meio da coleta de artigos através de pesquisa de termos de busca em bases de dados, seguindo-se critérios de seleção e de análise aprofundada de conteúdo.

De acordo com Lima e Mioto (2007), “a pesquisa bibliográfica implica em um conjunto ordenado de procedimentos de busca por soluções, atento ao objeto de estudo, que não pode ser aleatório.” Logo, a pesquisa bibliográfica teve uma importância significativa no desenvolvimento da pesquisa, tanto para a formulação de hipóteses e definição dos objetivos, como para a problemática da pesquisa, em seus aspectos técnicos relativos ao reconhecimento e extração de entidades nomeadas e seus relacionamentos, e, a vinculação das informações em estruturas de grafos de conhecimento. Sendo essencial para o máximo entendimento dos domínios, e, para o desenvolvimento de um método e/ou modelo eficaz para reconhecer, extrair e visualizar informações relevantes de dados não-estruturados.

A pesquisa se qualifica como de natureza teórica-aplicada devido, essencialmente, a seu propósito duplo de gerar conhecimentos teóricos, e, de desenvolver procedimentos que possam, efetivamente, serem aplicados na prática para a construção de um modelo para a recuperação e visualização das informações de receituários médicos. Silva (2001) destaca que “a pesquisa aplicada, tem o objetivo de gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos, e envolve verdades e interesses locais.”

Ao que se refere ao tipo, esta pesquisa, devido principalmente, a sua natureza e proposta, com características claramente multidisciplinares que se relacionam e, se complementam com estudos variados que abrange os campos de conhecimento da Ciência da Informação, Ciência da Computação, Linguagem Aplicada e Saúde Humana, se constitui em uma pesquisa do tipo exploratória. A pesquisa exploratória conforme Gil (2002) tem o objetivo primevo de “proporcionar familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses, sendo que, a grande maioria das pesquisas exploratórias envolvem levantamento bibliográfico e análise que estimulem a compreensão”.

Posto isso, esta pesquisa além da exploração metódica dos dados de medicamentos, e de anotações de receituários médicos em formato eletrônico. Também, com o objetivo de se obter uma visão geral acerca das pesquisas mais recentes e técnicas de PLN e Grafos de Conhecimento, e, com o intuito de conhecer o máximo das características do domínio em questão, foi fundamentada em pesquisa e análise bibliográfica pontual e ordenada, em livros, artigos, entre outras fontes de literatura.

Como pontuado, esta pesquisa foi balizada por um estudo de caso, onde os objetos de estudo da pesquisa são modelos textuais de receituários medicamentosos dos prontuários eletrônicos do HC de Marília. Neste contexto, fica evidente o caráter particular que objetiva investigar as verdades relacionadas aos prontuários desta instituição (ao que se refere a recuperação e visualização da informação de medicamentos descritos nestes).

Adicionalmente, a pesquisa também sugere um caráter coletivo ou representativo da investigação, sobretudo devido à estrutura padrão que define todos os documentos do mesmo modelo, bem como, a não padronização na formulação de anotações de texto livre, em determinados campos desses prontuários eletrônicos. Para Patton (2002), o propósito básico de um estudo de caso é reunir informações detalhadas e sistemáticas sobre um determinado fenômeno. Nesse sentido, Yin (2001), argumenta que o estudo de caso é muito útil para investigar a teoria de novos conceitos, e, verificar como estes elementos teóricos são aplicados e utilizados na prática. E ainda conforme aponta Yin (2001), "o estudo de caso nem sempre precisa incluir observações diretas e detalhadas como fonte de provas, e, pode mesclar provas quantitativas e qualitativas.

1.4.2 Apontamentos Teóricos e Práticos da Pesquisa

Finalmente, ao que se refere às questões teóricas desta pesquisa, e, quanto aos detalhes dos procedimentos metodológicos aplicados neste trabalho. Inicialmente, e, em momentos específicos da pesquisa foi efetuada a composição do corpus do referencial teórico/básico. Isso, por meio de uma pesquisa bibliográfica empreendida através de várias buscas nas bases de dados de periódicos científicos, a se saber, nas bases de dados: BRAPCI, Web of Science, E-lis, SCIELO, ResearchGate, Google Acadêmico e PubMed.

Os termos de buscas aplicados nas pesquisas foram: "Natural Language Processing" e "Processamento de Linguagem Natural"; "Electronic Health Record" e "Prontuário Eletrônico do Paciente"; "Representation of Information Electronic Patient Records" e "Representação da Informação Prontuário Eletrônico do Paciente"; "Knowledge Graphs" e "Grafos de Conhecimento". A partir da coleta dos artigos se realizou a leitura eletiva (que se deu tendo como base os objetivos e o problema de pesquisa), que resultou na escolha de 31 artigos. Os quais, por meio de leituras e estudos aprofundados foram resumidos com o intuito de se destacar: o problema de pesquisa, os objetivos, a metodologia, resultados e conclusões finais de cada artigo.

Ao que se refere a parte prática/aplicada dessa pesquisa, esta teve início com a aquisição dos dados advindos de duas fontes específicas. A primeira delas foi o HCM, que por meio do grupo de estudos GIHC que compõe a linha de pesquisa em "Informação e Tecnologia" da Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP) Campus de Marília, através do grupo de estudos HAIS pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília, que disponibilizou modelos de anotações textuais de receituários médicos.

A outra fonte de dados utilizada foi a Anvisa (Agência Nacional de Vigilância Sanitária), que disponibiliza em seu site oficial, a “lista de medicamentos similares e seus respectivos medicamentos de referência”, atualizada até 11/05/2020, que conforme o Diário Oficial da União. lista de medicamentos referência e seus respectivos similares, onde estes são classificados por ordem alfabética do medicamento de referência. A lista é disponibilizada em formato PDF, e, além de conter os nomes dos medicamentos, traz o princípio ativo, o detentor do registro, a forma farmacêutica, a concentração e a data da inclusão de cada medicamento.

Com a aquisição dos referidos dados, iniciou-se o processo de análise exploratória e a transformação dos dados. Onde os dados de modelos textuais de anotações medicamentosas disponibilizados pelo HAIS foram analisados e reproduzidos para a construção de um conjunto de dados (dataset) contendo variados nomes de medicamentos, princípio ativo, dosagem e concentração. Este dataset serviu de base para o desenvolvimento das tarefas de PLN, onde em uma fase inicial realizou-se a anotação desses medicamentos para a construção de um modelo automático de reconhecimento e extração de entidades nomeadas e relacionamentos. As anotações foram realizadas de forma semiautomática usando a ferramenta open source denominada “Label Studio” que gerou um novo conjunto rotulado de dados textuais contendo anotações de receituários de medicamentos, que foram exportados em um arquivo CSV.

A partir destes dados foram criados pipelines de código (definições de implantação por meio do código-fonte) com a linguagem Python, que com o uso dos recursos da biblioteca SpaCy. Por fim, o modelo foi avaliado e melhorado com ajustes de hiperparâmetros (parâmetros que devem ser definidos para ajustar e treinar um modelo), e, finalmente foi salvo, e usado para a extração e persistência dos dados no banco de dados gráficos denominado Neo4j, demonstrando na prática a viabilidade da aplicação do referido modelo. Por fim, os dados foram agregados a outro grafo contendo todos os medicamentos da referida lista da Anvisa.

1.5 Estrutura da Pesquisa

Estruturalmente, este trabalho dissertativo está disposto em oito seções que abordam os temas constituintes da pesquisa. Onde a seção 1, “Introdução” além de introduzir e delimitar o escopo da pesquisa, expõe os principais desafios e as oportunidades. Esta também, detalha o problema de pesquisa, a justificativa, objetivos gerais e específicos, bem como, a metodologia aplicada para o desenvolvimento desta pesquisa e, a presente estrutura da pesquisa.

A seção 2, denominada “Prontuários Eletrônicos do Paciente” contém uma breve introdução histórica do advento dos prontuários eletrônicos, sua adoção e implementação enquanto modelo padrão, no Brasil. Expõem-se os conceitos principais desta ferramenta

informacional, evidenciando as oportunidades e os desafios que surgem a partir da gradual adoção deste modelo digital de coleta e armazenamento de dados e informações clínicas e sociais de pacientes. E destaca a importância da pesquisa voltada para a melhoria constante desta ferramenta, visando uma melhor interoperabilidade e a integridade de dados e informações clínica-médica-sociais de pacientes. Discorre também, sobre a representação, recuperação e visualização da informação no domínio específico dos prontuários eletrônicos do paciente, apontando desafios e os esforços necessários para organizar e representar a informação e o conhecimento contido nesses referidos sistemas.

Na seção 3, denominada: "Demandas Informacionais em PEP: Estudos Incluídos na Revisão de Literatura desta Pesquisa" são apresentados conceitos e atributos básicos da Ciência da Informação destacando a organização da informação e do conhecimento, e os sistemas de organização do conhecimento. Ainda nesta seção, visando uma análise profunda acerca das demandas informacionais dos prontuários eletrônicos é apresentado o corpus do referencial teórico básico, apresentando o resumo de cada um dos artigos componentes destacando pontos principais do problema de pesquisa, bem como, objetivos, metodologia, resultados e conclusões de cada artigo. Com isso, pretende dar uma visão geral acerca das demandas informacionais do PEP, e das recentes pesquisas no domínio dos prontuários eletrônicos do paciente, ao que se refere especificamente a recuperação e representação do conhecimento e das informações coletadas e armazenadas em sistemas de prontuários eletrônicos. Bem como, do uso de técnicas de extração de conhecimento de dados não-estruturados, processamento de linguagem natural e grafos de conhecimento, aplicados à recuperação e visualização de informação de prontuários eletrônicos. Evidenciando, por fim, as contribuições necessárias do campo da CI.

A seção 4, “Representação da Informação e do Conhecimento” traz um histórico geral de variados modelos de formalização da informação e do conhecimento, visando traçar o panorama histórico da evolução destes até os Grafos de Conhecimento. Especialmente, delineando detalhadamente os conceitos definidores destes, sua evolução histórica, aplicações e sua atual importância na construção de sistemas inteligentes de indexação, representação, recuperação e visualização semântica e inter-relacionada da informação e do conhecimento. Com esse propósito, em subseções discorre sobre conceitos fundamentais teóricos e estruturais de Ontologias, Redes Semânticas, frames, RDF, OWL e Grafos de Conhecimento. Correlacionando e defendendo o uso de grafos de conhecimento na recuperação e visualização da informação de dados não-estruturados contidos nos receituários dos prontuários eletrônicos.

A seção 5, “Processamento de Linguagem Natural” trata sobre técnicas e ferramentas de PLN voltadas para resolução de tarefas relacionadas ao reconhecimento e extração de

entidades nomeadas e relacionamentos, e como estas entidades (informações) podem ser usadas para a construção e atualização de Grafos de Conhecimento, especificamente a extração de textos não-estruturados de domínio clínico em prontuários eletrônicos.

A seção 6, “Modelo de Recuperação e Visualização de Informações de Medicamentos” apresenta os métodos para o desenvolvimento do modelo proposto para a recuperação e visualização da informação proveniente da extração de dados de anotações textuais não-estruturadas de medicamentos descritos em prontuários eletrônicos do paciente. Bem como, detalhar a incorporação de um modelo estruturado em grafos de conhecimento contendo medicamentos referências e seus respectivos similares, e o princípio ativo de cada medicamento. Também descreve de forma detalhada todos os recursos e procedimentos executados para o desenvolvimento deste referido modelo.

A seção 7, “Apresentação e Interpretação dos Resultados”, neste capítulo são apresentados os resultados alcançados da pesquisa e do modelo de recuperação e visualização de medicamentos descritos em prontuários eletrônicos.

A seção 8, “Discussões e Argumentação" nesse capítulo é discutido os principais problemas abordados e a justificativa de uso de processamento de linguagem natural e principalmente, o uso de grafos de conhecimento para representação dos dados de medicamentos, destacando vantagens e benefícios do uso dessa formalização e tecnologia.

A seção 9, “Considerações Finais”, evidentemente, é onde são apresentadas as considerações finais a respeito do desenvolvimento deste trabalho de pesquisa e do modelo, e afirma as contribuições do projeto de pesquisa.

2 PRONTUÁRIOS ELETRÔNICOS DO PACIENTE

Nesta seção, são apresentados os conceitos e as principais características definidoras dos prontuários eletrônicos do paciente. Esta também traz considerações fundamentais acerca da qualidade de dados coletados em prontuários eletrônicos, e discorre-se sucintamente sobre a indexação dos dados nestes referidos sistemas de informação de domínio médico.

2.1 PEP: Conceitos e Principais Características

O desenvolvimento e adoção de novas tecnologias voltadas à coleta, organização, recuperação e gestão da informação e do conhecimento constitui-se em uma forma muito objetiva de se promover a inovação e por conseguinte, a melhora efetiva no campo da saúde. Neste sentido, uma inovação muito importante a ser destacada é o advento dos Prontuário Eletrônico do Paciente, ou simplesmente (PEP). Os prontuários eletrônicos de modo gradual vêm sendo implementados e aprimorados nos serviços públicos e privados de saúde.

O Prontuário Eletrônico do Paciente genericamente pode ser definido como, um modelo lógico-físico que permite a criação, armazenamento e gestão de coleções heterogêneas de dados relacionados ao histórico geral do paciente. O que possibilita novos padrões de processos de atendimento ao paciente e de pesquisa clínica e social, por meio de análise de dados.

O Prontuário Eletrônico do Paciente é um modelo de solução de software que foi idealizado e desenvolvido de forma padronizada objetivando otimizar os processos de coleta, armazenamento e a recuperação dos dados médicos. Bem como, de todas as informações que estão relacionadas à identificação social e ao processo de atendimento dos pacientes. Idealmente, o PEP tem a prerrogativa de promover a modernização dos serviços de saúde tornando-os mais integrados, interoperáveis e eficientes. Otimizando a qualidade e a rapidez de todo o processo e procedimentos envoltos no atendimento ao paciente.

No Brasil o prontuário eletrônico foi introduzido no ano de 2002, e, de acordo com as orientações e determinações da Resolução CFM Nº 1638/2002, o sistema deve seguir uma estrutura padrão que define todos os documentos do mesmo modelo. Sendo como aponta a referida resolução “[...] um documento valioso para o paciente, corpo médico, instituições de saúde, e, finalmente, para iniciativas voltadas ao ensino, pesquisa e aos serviços públicos de saúde, além de se constituir em um instrumento de defesa legal.” (BRASIL, 2002).

Desde sua implementação, enquanto modelo, o prontuário eletrônico o paciente tem rapidamente se constituindo em um recurso fundamental para a melhora e manutenção do sistema integrado de saúde. Sendo utilizado como objeto de intermédio entre o corpo de

profissionais de saúde, o paciente e outros especialistas. Se constituindo em "[...] um documento multidisciplinar, atemporal e abrangente, isso devido aos diversos tipos de profissionais que registram informações neste documento e acessam seus conteúdos informacionais" (CARVALHO, 2018).

As informações coletadas e armazenadas no prontuário eletrônico conforme Galvão e Ricarte (2012) são heterógenos, abrangendo uma grande "diversidade e especificidades temáticas." Logo a lista de dados que podem ser coletados e armazenados nos prontuários eletrônicos é verdadeiramente extensa e diversa, como destacam os autores:

> "... identificação da instituição de saúde que prestou e/ou está prestando a assistência; identificação do paciente, número de registro na instituição, nome civil, nome social, local e data de nascimento, sexo, estado civil, nomes dos pais, nome do cônjuge, profissão, responsável, endereço residencial, telefones para contato, procedência do paciente; lista de diagnósticos com respectivos códigos de classificação; históricos das doenças; informações sobre sistemas e aparelhos; doenças concomitantes; exame físico, peso, altura, estado geral, mucosas, pele, temperatura, pressão sanguínea; antecedentes pessoais e familiares; hábitos alimentares e aspectos nutricionais; condições de trabalho e moradia; aspectos educacionais, psicológicos, sociais; hipóteses de diagnóstico; exames complementares; pareceres solicitados; condutas adotadas; prescrição (medicamentos a serem empregados pelo paciente, com dose e horário de administração); retorno à assistência (instituição, dia e hora da nova assistência); data e hora dos atendimentos recebidos pelo paciente; gravidez e parto; processo e situação de nascimento; antibiogramas; anticorpos; audiometrias; bacterioscopias; biópsias; citologias; compatibilidade sanguínea; culturas; diálise peritoneal; ecocardiografias; ecografias; eletroencefalografias; endoscopias; exames micológicos; gasometria; hemodiálise; hemograma; ionograma; provas de função hepática e renal; quimioterapia; radiografias; radioterapia; tomografias computadorizadas; internação; autorização para realização de procedimentos assinado pelo paciente ou por seu responsável; identificação da clínica de assistência ao paciente; identificação da enfermaria e do leito de assistência ao paciente; procedimento pré-anestésico, anestesia, e procedimento pós-anestésico; procedimento cirúrgico; hora do início e do fim do procedimento recebido; identificação, assinatura e registro do profissional de saúde que prestou assistência ou procedimento; encaminhamento do paciente; assistência ambulatorial; atendimento de urgência; transferência do paciente entre unidades; óbito, necropsia e declaração de óbito, dentre outras." (GALVÃO e RICARTE, 2011, p. 81-82).

Nesse contexto, Galvão e Ricarte (2012) reafirmam que a implementação dos prontuários eletrônicos do paciente faz emergir inúmeras e importantes oportunidades. Posto que, enquanto rica fonte de informação, este vai muito além de um mero ferramental de assistência médica ao paciente, se constituindo em uma fonte de dados fundamental para a pesquisa, integração e a interoperabilidade de dados clínicos-socio-informacionais.

Deste modo, esta imprescindível ferramenta de saúde tem como objetivo principal a modernização digital de todo o processo informativo envolto no atendimento primário do

paciente, e de forma efetiva contribui para agilizar e otimizar a coleta e o acesso às informações e ao perfil histórico da saúde de um paciente. Em um esforço para otimizar o processo de atendimento clínico e, o atendimento médico ao paciente, e, desta forma, melhorar o nível de qualidade da saúde de toda população. Logo, os prontuários eletrônicos do paciente têm se mostrado uma poderosa ferramenta médica de coleta, armazenamento e recuperação da informação e do conhecimento médico e clínico-social, contribuindo efetivamente para a inovação e otimização de todo o processo de atendimento do paciente, e se constituindo em uma inesgotável fonte de dados e informações que viabilizam um sem número de possibilidades de pesquisas científicas de variados domínios do conhecimento.

Logo, fica evidente que a adoção e a gradual implementação dos prontuários eletrônicos do paciente traz inúmeros benefícios, porém, também traz consigo alguns desafios sendo que os mais importantes se relacionam com a qualidade e recuperação dos dados. A qualidade dos dados coletados no PEP costuma ser afetada pelo preenchimento de anamneses de forma inadequada, por dados confusos, uso de siglas, abreviações, uso de jargões, etc. Ou seja, falta de padrão textual de preenchimento. Desta forma, a qualidade de dados é afetada por uma série de possibilidades, impactando no ciclo de vida dos dados desde a coleta, armazenamento, recuperação e a representação da informação. Fazendo, inevitavelmente, emergir problemas de ordem informacional que impactam a qualidade da prestação do serviço de atendimento.

2.2 A Qualidade dos Dados de Prontuários Eletrônicos

Nos prontuários eletrônicos, a qualidade dos dados é um aspecto crítico, posto que, pode ter um grande impacto na recuperação de informações. A qualidade dos dados coletados se imprecisos, incompletos ou inconsistentes podem afetar a capacidade dos profissionais de saúde de acessar informações relevantes e importantes sobre os pacientes, de modo rápido e preciso, o que pode levar a atendimentos demorados, inadequados e a conclusões e diagnósticos errôneos.

Assim, nesses sistemas a qualidade dos dados é afetada já desde o momento em que um profissional da saúde realiza o procedimento de coleta dos dados durante o processo de anamnese. Nesse contexto, a coleta, organização e armazenamento da informação, nem sempre, é feita de modo a facilitar a recuperação e visualização desses dados. O que em um cenário dinâmico, e, por muitas vezes caótico como o é, o da saúde pública, a tarefa de recuperação da informação torna-se uma barreira para que o profissional da saúde faça o atendimento, ou realize um diagnóstico, ou mesmo tenha uma melhor tomada de decisão associada ao paciente.

Conforme destacam Oliveira e Favaretto (2021), a eficácia dos prontuários eletrônicos está intimamente ligada à qualidade dos dados coletados e armazenados, sendo que, a baixa qualidade dos dados, especialmente no setor da saúde, terá efeitos ruins de longo prazo e alcance. Desta forma, para os autores "[...], quanto maior a qualidade dos dados, melhores tenderão a ser os resultados tanto para pacientes, como para a instituição."

Neste contexto, como pontuado, um fator real que contribui consideravelmente para a diminuição da qualidade dos dados coletados e armazenados em sistemas de prontuários eletrônicos, e, por conseguinte, afetando sua recuperação, são os dados coletados e armazenados em formato não-estruturados, que compõem boa parte dos dados médicos coletados.

Os prontuários eletrônicos do paciente, de modo geral possui informações que foram registradas em campos puramente textuais, onde a coleta das informações é realizada em um formato de dados não-estruturados e sem nenhuma padronização sintática ou semântica. Deste modo se constituindo, verdadeiramente, em uma situação que compromete a qualidade e a integridade dos dados e informações coletadas. Este aspecto particular dos prontuários eletrônicos impacta decisivamente no processo de recuperação da informação clínica-social dos pacientes, posto que, entre outras, faz com que os profissionais de atendimento médico despendam considerável tempo pesquisando nos dados históricos do paciente. Ou mesmo fazendo com que informações de suma importância se percam no processo, ou, passem despercebidas inviabilizando a objetividade do entendimento real, da condição do paciente.

Portanto, a capacidade de lidar com dados não estruturados é uma questão importante pois afeta a eficácia da recuperação da informação. E, a qualidade dos dados não estruturados em prontuários eletrônicos é crucial para a ótima recuperação de informações relevantes e, pode ser aprimorada, entre outras, por meio de técnicas avançadas de recuperação da informação. Como a utilização de técnicas atuais de mineração de texto e processamento de linguagem natural útil na extração de informações relevantes a partir desses dados não estruturados.

2.3 Recuperação e Visualização da Informação em PEP

Como em qualquer campo do conhecimento, na abrangente área da saúde as emergentes tecnologias informacionais se fazem imprescindíveis nos mais diversos setores que compõe o sistema de saúde, estando presente na área administrativa e de gestão clínica-hospitalar. Bem como, na área de pesquisa científica, voltada à descoberta e produção de novos medicamentos e de modelos de tratamento médicos, no ensino e na pesquisa científica. Nesse cenário, fica evidente que o desenvolvimento de sistemas de informação dedicados ao melhoramento progressivo dos prontuários eletrônicos, especialmente visando otimizar a recuperação e estruturação de variadas informações de pacientes, se mostra um fator prioritário.

Neste sentido, a construção de soluções para otimizar a recuperação e representação das informações coletadas e armazenadas em prontuários eletrônicos, constitui-se em um desafio muito complexo frente à recuperação da informação de dados de pacientes. E, essa complexidade surge, sobretudo porque os prontuários eletrônicos, além de campos de coleta de dados estruturados, que são dados que possuem campos com valores já pré-definidos, também é composto por faixas de formulários que coletam dados e/ou informações puramente compostas de textos livres, ou seja, textos não-estruturados e, portanto, não padronizados.

De acordo com Souza e Almeida (2019), as informações coletadas por meio de campos de formulários puramente textuais em um formato de dados não-estruturado, se constitui em uma barreira que efetivamente dificulta e prejudica a recuperação e uso da informação clínica de pacientes, especialmente para pesquisas. Em parte, posto que, conforme destaca Martha *et al* (2004) em textos livres, e, portanto, não-estruturados "[...] a capacidade de narrativa empregada para representar a realidade das condições do paciente é limitada apenas pela criatividade do autor, fazendo surgir incoerência entre um documento e outro." Neste contexto, verifica-se que as anamneses, contém, invariavelmente, problemas relacionados a qualidade e rigor textual e, como já pontuado, contém: erros de digitação, ortográficos, gramaticais, sinônimos, concordâncias verbais, siglas, abreviações, termos médicos, entre outros, criando dificuldades consideráveis para o processo de recuperação da informação clínica.

Portanto, este aspecto dos prontuários impacta de modo decisivo na recuperação dos dados e informação clínica dos pacientes, o que, por fim contribui negativamente, para que o corpo médico recupere os dados desejados. Fazendo com que se desperdice um tempo considerável filtrando no histórico do paciente informações verdadeiramente importantes, para que, de forma objetiva este entenda o contexto geral de atendimento e condição do paciente.

Logo, fica evidente que o considerável volume de dados do tipo textual livre inseridos diariamente em prontuários dos pacientes faz com que a tarefa de recuperação dessas informações fique cada vez mais inviável de ser realizada de forma puramente humana.

Assim, em meio a esse imensurável montante de dados de anotações não-estruturadas, as técnicas e modelos computacionais inteligentes, voltados tanto para estruturação e representação de dados, como para o processamento de linguagem natural, se apresentam como ferramentas eficientes e imprescindíveis para recuperar, estruturar e, desta forma, desbloquear muito do conhecimento médico e social oculto nos grandes conjuntos de dados clínico-sociais não-estruturados, contidos nos prontuários eletrônicos de paciente (LIU *et al*, 2012).

A estruturação eficiente de textos contribui para otimização do acesso e recuperação da informação, posto que promove uma compreensão e inteligibilidade rápida e natural. Assim, de acordo com Valentim *et al* (2010) os dados e informações coletados e armazenados, "[...] quando adequadamente estruturados, tendem a fornecer informação relevante para quem o está acessando e, por conseguinte, auxiliam na construção do conhecimento por parte do usuário."

O objetivo primevo dos prontuários eletrônicos é o de otimizar e facilitar a recuperação das informações do paciente, agilizando os processos e melhorando a produtividade e os serviços de atendimento ao paciente. Logo, a importância do uso de técnicas para facilitar a recuperação da informação. Sendo um fator crucial no esforço para melhorar a recuperação da informação dos pacientes, especialmente as derivadas de fontes não-estruturadas de dados. Nesse esforço, os denominados Grafos de Conhecimento em conjunto com técnicas modernas de Processamento de Linguagem Natural ou (PLN) baseadas em Aprendizado de Máquina, despontam como ferramentas eficientes, capazes de promover de modo otimizado a recuperação e a representação clara e intuitiva da informação e do conhecimento clínico.

Nesse contexto, um aspecto muito importante ao que se relaciona a informação médica contida nos Prontuários Eletrônicos dos Pacientes, é como a informação e/ou conhecimento de ordem clínica é organizado e representado nesse tipo de sistema de registros médicos. De acordo com Bräscher e Café (2010) "[...] a organização do conhecimento é um processo de modelagem que visa construir representações do conhecimento". No âmbito da Ciência da Informação, a representação da informação e do conhecimento basicamente está relacionada com as formas de simbolizar a informação e o conhecimento.

Nos Prontuários Eletrônicos dos Pacientes, a representação da informação e do conhecimento essencialmente objetiva a otimização da interoperabilidade semântica. Nesse sentido, um aspecto impotente é a semântica de dados, ou modelo de dados semânticos, que, devido seu alto potencial de integração, tem, cada vez mais, se constituído como uma

característica padrão no desenvolvimento de tecnologias e sistemas de software informacional, como os PEPs. Isso porque, esses denominados sistemas semânticos objetivam atender aos anseios constituintes da denominada: Web Semântica. E, portanto, visam disponibilizar dados em um formato padrão, acessível e gerenciável, onde os dados ou conjuntos de dados sejam sempre disponibilizados de forma inter-relacionada. Ou seja, visam disponibilizar dados altamente vinculados, em um modelo onde dados essencialmente incluam informações de base semântica que dê significado aos dados, e seus relacionamentos (W3C, 2021).

No cenário atual, com os PEPs revolucionando o armazenamento e gerenciamento de informações de pacientes, a recuperação de informações valiosas (como informações de medicamentos) a partir de textos não estruturados e semiestruturados nesses sistemas eletrônicos continua sendo um desafio significativo. Posto que, hoje, de modo geral, métodos manuais são predominantemente utilizados para recuperação, os quais são demorados, propensos a erros e dificultam a tomada de decisões eficientes. Para superar essas limitações, entende-se com essa pesquisa que a integração de técnicas de processamento de linguagem natural (PLN) e grafos de conhecimento ao processo de recuperação oferece uma solução promissora, especialmente, para a recuperação de dados e informações de medicamentos.

Deste modo, entende-se que as limitações atuais da recuperação não automatizada implicam na demora e intensidade de trabalho já que extrair informações de textos não estruturados nos PEPs frequentemente requer que profissionais de saúde revisem e analisem manualmente cada registro, o que exige um investimento significativo de tempo e esforço. Há também a falta de escalabilidade, posto que, a medida que o volume de dados médicos eletrônicos continua a crescer, os métodos de recuperação manuais se tornam cada vez mais inadequados para lidar com a análise de dados em grande escala de forma eficiente. E por fim, há a propensão a erros e inconsistências, já que a recuperação baseada em seres humanos é suscetível a erros, inconsistências e variabilidade entre anotadores, afetando a confiabilidade e a precisão das informações extraídas.

Logo, a recuperação de informações de textos não estruturados de PEPs, é atualmente não automatizada e altamente propensa a erros. Para melhorar esse processo, podem ser usadas técnicas de processamento de linguagem natural (especialmente, a tarefa de reconhecimento e extração automatizada de entidades e relacionamentos) e a estruturação em grafos de conhecimento para desbloquear todo o potencial dos PEPs e melhorar os resultados.

2.4 Receituários de Medicamentos em Prontuários Eletrônicos

Dentre as diversas informações geradas e contidas em um prontuário de paciente, a prescrição ou receituário de medicamentos é um dos elementos informacionais mais fundamentais contendo informações detalhadas, como dosagem, posologia, duração do tratamento e informações adicionais relevantes. Desta forma, os receituários eletrônicos de medicamentos constituem-se em uma parte essencial desses sistemas, registrando as prescrições médicas e desempenhando um papel crucial na troca de informações entre os profissionais de saúde envolvidos no cuidado ao paciente. A digitalização e armazenamento desses receituários trazem vantagens significativas, posto que permitem uma fácil modificação dos dados, o que possibilita manter as informações sobre medicações sempre atualizadas.

Além disso, a prescrição eletrônica, de acordo com Neumamm *et al*. (2023), tem como objetivo principal a segurança do paciente e o uso racional de medicamentos, prevenindo erros de prescrição e uso, e prevenindo danos. Oliveira *et al*. (2022) destacam que as tecnologias e sistemas informatizados, como o prontuário eletrônico, promovem a assistência farmacêutica no Brasil, melhorando a qualidade e segurança dos serviços de saúde e reduzindo danos, ao permitir o uso consciente de medicamentos. Bem como, possibilita a integração desses dados com outros sistemas de saúde permitindo uma visão mais completa do histórico do paciente.

O receituário de medicamentos em prontuários eletrônicos do paciente abriga uma variedade de informações relevantes, e dados fundamentais, incluindo:

- **Nome do paciente:** O nome completo do paciente é sempre incluído na prescrição para garantir que o medicamento seja prescrito para a pessoa correta.
- **Nome do medicamento:** O nome completo do medicamento, incluindo a dosagem e a forma (comprimido, cápsula, líquido etc.), é importante para garantir que o medicamento correto seja fornecido.
- **Posologia:** A posologia indica a frequência com que o medicamento deve ser tomado, bem como a quantidade a ser tomada.
- **Duração do tratamento:** A duração do tratamento também é importante para garantir que o paciente continue a tomar o medicamento pelo tempo necessário para que seja eficaz.
- **Indicação:** A indicação é a razão pela qual o medicamento está sendo prescrito. Isso ajuda a garantir que o medicamento correto seja prescrito para tratar a condição médica adequada.

- **Contraindicações:** As contraindicações indicam quais pacientes não devem tomar determinado medicamento e os efeitos colaterais que podem ocorrer em pacientes que tomam o medicamento.
- **Interações medicamentosas:** As interações medicamentosas ocorrem quando dois ou mais medicamentos são tomados juntos e podem interferir uns com os outros. Essa informação é importante para evitar interações prejudiciais.
- **Observações:** As observações podem incluir informações adicionais sobre o medicamento prescrito, como instruções especiais de administração ou informações sobre efeitos colaterais específicos que o paciente deve estar ciente.

Como listado as prescrições em prontuários eletrônicos contém informações relevantes para a segurança e eficácia do atendimento e do tratamento dos pacientes. De modo geral, todas essas informações são elementos básicos para garantir que o médico monitore o progresso do tratamento ao longo do tempo. Sendo, da mesma forma, como pontuado, cruciais para promover uma comunicação eficaz entre os profissionais de saúde envolvidos no cuidado ao paciente.

Diante disso, entende-se que a análise e recuperação dos dados contidos nos receituários eletrônicos têm o potencial de melhorar as práticas clínicas em diversos aspectos. Por exemplo, a identificação de padrões na prescrição de medicamentos pode fornecer insights valiosos sobre o tratamento de pacientes individuais e em larga escala. Isso permite a identificação de abordagens terapêuticas eficazes, provendo melhores resultados para os pacientes.

A recuperação de dados e informações desses receituários de medicamentos pode ajudar a melhorar as práticas clínicas, prevenir erros medicamentosos e fornecer dados valiosos sobre um, ou um conjunto de pacientes. Além disso, a análise e recuperação de dados de receituários eletrônicos pode auxiliar na prevenção de erros medicamentosos, pois, ao identificar possíveis interações medicamentosas, alergias do paciente, entre outros, os profissionais de saúde podem tomar medidas preventivas adequadas. A detecção precoce de erros de prescrição e a promoção de práticas seguras de medicação são de vital importância para evitar danos aos pacientes.

Desta forma, a recuperação de informações sobre medicação do paciente são cruciais para os cuidados de saúde. No entanto, cerca de 80% dessas informações estão contidas em texto não estruturado gerados em prontuários eletrônicos, tornando a extração manual difícil e demorada. Portanto, é de suma importância extrair e estruturar menções de medicamentos e suas informações relevantes, como dosagem, concentração, frequência, duração e via de administração, entre outras (Jouffroy *et al*., p. 1, 2021). Assim, é essencial destacar, e, de acordo com Neumamm *et al*. (2023), que a automação na coleta e registro estruturado de dados

relacionados à prescrição e uso de medicamentos possibilita análises mais completas, confiáveis e ágeis. Isso facilita a tomada de decisão tanto no contexto clínico (uso racional de medicamentos, segurança do paciente, etc.), e no contexto de gestão.

Logo, a análise e recuperação dos dados e informações contidos nesses receituários de medicamentos podem trazer benefícios significativos para a área da saúde, ajudando a prevenir erros de administração de medicações, aprimorar recomendações de medicamentos e, as práticas clínicas e fornecer insights valiosos sobre o tratamento de pacientes de forma personalizada e/ou de modo geral, em larga escala. Esta pesquisa se concentra na construção de um modelo para a recuperação e estruturação de informações de medicamentos especificamente, informações relevantes, como: nome de medicamentos, dosagem, concentração, frequência de administração. Mas, enquanto modelo, este pode ser ampliado e usado para recuperar outras informações importantes, como as descritas na lista anterior.

Posto isso, e, como pontuado no capítulo 1.1, a recuperação e visualização de dados de medicamentos nos prontuários são aspectos cruciais para a solução do problema abordado nesta pesquisa. Compreende-se que a identificação, extração e estruturação de informações sobre medicamentos presentes nos prontuários possibilitam um acesso mais efetivo às informações relevantes contidas em textos não estruturados ou semiestruturados. Ao extrair informações como nomes de medicamentos, concentração, forma e posologia dos prontuários, é possível desenvolver e aprimorar visualizações detalhadas do perfil médico de cada paciente. A visualização dos dados de medicamentos desempenha um papel crucial na compreensão das relações entre diferentes medicamentos e suas características. Assim, ao mapear essas informações em gráficos ou outras representações visuais, é possível identificar e recuperar informações de medicamentos, e principalmente, informações de medicamentos similares.

Considerando o problema em questão e com o intuito de alcançar os objetivos estabelecidos, o escopo desta pesquisa concentra-se no problema de recuperação e visualização dos seguintes dados de medicamentos: nomes de medicamentos de referência e de seus similares, princípio ativo, concentração, forma e posologia. Pois, entende-se que a identificação e estruturação desses dados são elementos essenciais para a construção de visualizações e de um sistema de recomendação de medicamentos similares. Além disso, a identificação e visualização dos dados de medicamentos podem ser expandidos de forma indefinida, a partir dessas, e de maneira evolutiva, vindo a oferecer várias ramificações importantes para a consulta automática de informações relacionadas a medicamentos. É fundamental ressaltar que a capacidade eficiente de identificar e visualizar os dados de medicamentos nos prontuários é de extrema importância, pois afeta diretamente a qualidade e o tempo de atendimento.

3 DEMANDAS INFORMACIONAIS EM PEP: ESTUDOS INCLUÍDOS NA REVISÃO DE LITERATURA DESTA PESQUISA

Em cada artigo científico selecionado e analisado para compor o corpus documental desta revisão de literatura, além de pesquisas que contemplam a organização, representação e a recuperação da informação em sistemas de prontuários eletrônicos do paciente. Foram obrigatoriamente, como critério de eletividade, selecionadas pesquisas acadêmicas que investigaram o uso de técnicas atuais de Processamento de Linguagem Natural e/ou do uso de Grafos de Conhecimento no contexto dos prontuários eletrônicos dos pacientes.

Desta forma, para uma seleção sistematizada de literatura as análises integrais das produções científicas foram orientadas por categorias que apontam para as possíveis contribuições da Ciência da Informação. Isso, baseando-se no trabalho de Galvão e Ricarte (2012), que em suas pesquisas fizeram um levantamento profundo de demandas informacionais necessárias, relacionadas ao prontuário eletrônico do paciente. Desta forma**,** conforme as demandas identificadas pelos referidos autores, foram selecionados estudos relacionados a: aquisição da informação, preservação da informação, identificação da informação, seleção da informação, organização da informação, análise da informação e comunicação da informação.

Onde conforme Galvão e Ricarte (2011):

A Aquisição da Informação se relaciona ao ato de projetar, validar e implementar metodologias inerentes aos processos, fluxos, compartilhamento e troca de informações. Que de acordo com Galvão e Ricarte (2011) constitui-se, atualmente, em uma demanda, por parte dos profissionais da informação, para a recepção de dados e informações oriundos de contextos externos aos prontuários eletrônicos e, portanto, externo à própria instituição de saúde.

A Preservação da Informação refere-se aos meios e formas de armazenamento e preservação da informação, especialmente da informação digital, diante da fragilidade inerentes à preservação de documentos digitais, bem como, a problemas relacionados à integridade da informação e a veloz obsolescência das tecnologias digitais. Assim, conforme Galvão e Ricarte (2011) há uma genuína preocupação quanto ao controle de situações ambientais ou tecnológicas inadequadas que pode levar a perda de dados contidos, tanto no prontuário eletrônico do paciente, como, em outro objeto informacional pertencente ao campo da saúde.

A Identificação da Informação relaciona-se às possíveis contribuições para a aplicação da técnica computacional (como de processamento de linguagem natural e de estruturação da informação em grafos de conhecimento) em prontuários eletrônicos. Logo, consiste no processo de identificação para organização e recuperação da informação, sendo necessária a correta classificação e/ou descriminação de informações inerentes a pacientes, comorbidades,

medicamentos, procedimentos, entre outros. Bem como, procedimentos, relacionados à validação e implementação de metodologias para registros e associações de instrumentos de apoio terminológico ao prontuário. (Galvão e Ricarte, 2012).

A Seleção da Informação, por sua vez, está relacionada com a constante necessidade do estabelecimento de padrões, políticas e/ou critérios de seleção e atualização das informações a serem registradas e recuperadas dos prontuários eletrônicos. Conforme Galvão e Ricarte (2011), essa necessidade de seleção de dados e informações que compõem o prontuário, se justifica diante das constantes mudanças tecnológicas, científicas e políticas no campo da saúde.

A Organização da Informação, conforme Galvão e Ricarte (2011) faz, evidentemente, referência à organização da informação, mas, também, no contexto dos prontuários eletrônicos se refere a recuperação da informação. Desta forma, no referido contexto se trata de uma atividade, essencialmente, voltada para a localização de dados e informações variadas e necessárias ao ótimo atendimento do paciente. Bem como, para o uso desses dados e informações para a pesquisa e ensino em saúde e em ciências político-sociais.

A Análise da Informação, por sua vez, faz referência aos processos ligados à análise e a síntese da informação contida em prontuários eletrônicos dos pacientes. Dos quais, de acordo com Galvão e Ricarte (2011) podem ser coletados e/ou reunidos dados e informações específicas de um paciente no momento da assistência ou atendimento médico, de modo específico visando um melhor e otimizado atendimento ao paciente. E, de modo geral, promovendo o efetivo aprimoramento dos processos de gestão, bem como, balizando a criação de políticas públicas de saúde, baseadas, essencialmente, em dados concretos e factuais.

A Comunicação da Informação, finalmente, se relaciona com o planejamento, tradução, harmonização e a manutenção de nomenclaturas clínicas. Que conforme Galvão e Ricarte (2011) padroniza aspectos essenciais para o contexto da saúde de modo geral. Entre outras, possibilitando a observação de terminologias e entidades de anotações de prontuários eletrônicos que podem ser compreendidas por equipes multiprofissionais a qualquer tempo.

Posto isso, segue o resumo dos 31 artigos componentes do corpus documental que cientificamente fundamenta esta revisão de literatura, que relaciona, especialmente, a recuperação da informação no contexto dos prontuários eletrônicos, por meio de técnicas avançadas de processamento de linguagem natural e dos denominados grafos de conhecimento.

Galvão e Ricarte (2012) em sua pesquisa fizeram um levantamento de demandas informacionais e tecnológicas para otimização da formulação da informação contida em prontuários eletrônicos do paciente, buscando identificar como dadas demandas podem conter a resolução de problemas por meio de conhecimentos provenientes da área da ciência da

informação. A metodologia empregada pelos autores, constitui-se em uma revisão de literatura relacionada aos prontuários eletrônicos, e conexões com a ciência da informação. Com esse propósito, foram pontuadas observações relacionadas ao uso de prontuários eletrônicos em diferentes suportes e instituições, sendo que a pesquisa teve um olhar multidisciplinar contando com a participação de profissionais da saúde, ciência da informação e tecnologia da informação (TI), bem como gestores e pesquisadores que fazem uso de documentos de prontuários. Nesse contexto, os autores observam que há demandas de ordem informacional e tecnológica, relacionadas ao processo de geração, comunicação, identificação, seleção, aquisição, organização, recuperação, entre outras. Conclui-se assim que o prontuário eletrônico apresenta demandas no campo de atuação de profissionais da informação, constituindo-se como uma área de pesquisa onde a ciência da informação pode dar uma contribuição relevante, contribuindo efetivamente para que o campo da saúde amplie e aperfeiçoe teorias e metodologias relacionadas a recuperação da informação.

Souza e Almeida (2019) investigaram formas de descrever conexões com terminologias médicas padronizadas advindas de dados clínicos de textos livres de prontuários eletrônicos, e assim promover a recuperação de referidos dados por meio de ontologias e vocabulários controlados. Os autores apontam que boa parte dos dados disponibilizados em prontuários eletrônicos compõe-se de textos não-estruturados, o que de acordo com os mesmos, impõe fortes barreiras a serem transportas para a utilização otimizada dos dados clínicos, tanto para tratamento como para pesquisa. Conclui-se que, os dados de texto livre, ou seja, não-estruturados em anotações médicas contém grande variedade de sinônimos, acrônimos, e peculiaridades, dissociadas de terminologias médicas padronizadas, resultando em complicações e tornando extremamente complexa a recuperação das informações.

Schrodt *et al* (2020) buscaram com o seu trabalho analisar o uso de grafos de conhecimento para representar e recuperar informações clínicas de pacientes advindos de prontuários eletrônicos, nesse sentido, investigaram as principais inovações em dado campo, com o intuito de apresentar um panorama geral. Foi empregada uma metodologia de revisão sistemática de literatura e meta-análises. Os resultados demonstram que os grafos de conhecimento atualmente têm sido amplamente empregados para representar o conhecimento e a informação contida em prontuários eletrônicos, tanto informação de dados temporal e casual, bem como, dados biológicos e heterogêneos, como informações sobre doenças e tratamentos. Em conclusão, a pesquisa mostrou que o uso de grafos de conhecimento para representar informações de dados clínicos de pacientes, é uma técnica muito promissora,

possibilitando diferentes formas de uso para construção do conhecimento a partir dos prontuários eletrônicos.

Wang *et al* (2020) desenvolveram um grafo de conhecimento para representação da informação clínica direcionada à prevenção de diabetes tipo 2. A construção do grafo foi fundamentada em uma abordagem baseada na medicina orientada em evidência. O grafo construído representa através de entidades e relacionamentos, sintomas associados a doenças, onde palavras chaves extraídas de dados de prontuários relacionam diabetes do tipo 2, extraindo características e fatores de risco. A metodologia consistiu na coleta de dados que englobam pesquisas clínicas relacionadas a inúmeras doenças associadas ao diabetes tipo 2. Procedeu-se à extração de características e fatores de risco, bem como tratamentos e testes, por fim, deu-se à construção do grafo de conhecimento. Concluiu-se, que os denominados grafos de conhecimento, fundamentados em base de dados baseada em domínio podem fornecer informações qualificadas e assim, servir como uma ferramenta de apoio a decisões de tratamento e prevenção de doenças, como diabetes do tipo 2.

Jackson *et al* (2016) em seu trabalho fizeram uso de processamento de linguagem natural para extrair dados de sintomas de doenças relacionadas a problemas mentais graves, visando utilizar os dados obtidos para a realização de pesquisas na área de saúde mental. A metodologia utilizada se baseou no desenvolvimento, bem como na validação dos dados e informações extraídas dos textos médicos contidos em prontuários eletrônicos, visando a identificação de possíveis sintomas. Os resultados mostraram uma considerável eficácia, posto que sintomas de doenças mentais graves foram identificados por uma equipe de psiquiatras com base em características linguísticas em registros. Conclui-se assim, que a aplicação de técnicas de processamento de linguagem natural em textos clínicos relacionados a doenças mentais graves, podem contribuir eficazmente para a extração de uma extensa gama de sintomas e doenças mentais.

Chase *et al* (2017) analisaram em sua pesquisa, a possibilidade de aplicação de técnicas de processamento de linguagem natural em dados disponíveis em prontuários eletrônicos do paciente para identificar sintomas relacionados à doença de Esclerose Múltipla, a partir de sinais e sintomas identificados em dados de pacientes. A metodologia adotada, consistiu na extração de dados e informações advindas de anotações clínicas de pacientes, bem como no mapeamento de classificação de termos clínicos. Os resultados obtidos indicam que a aplicação de técnicas e algoritmos de processamento de linguagem natural, podem alcançar uma considerável precisão na previsão baseada em sinais e sintomas relacionados à Esclerose Múltipla. Concluiu-se, que a aplicação de processamento de linguagem natural melhora consideravelmente a

precisão de diagnóstico prévio de referida doença, posto que um modelo de previsão baseado em dada técnica pode prever precocemente os primeiros sinais e sintomas dos pacientes.

Li *et al* (2020) analisaram os procedimentos sistemáticos e eficientes para a criação do grafo de conhecimento a partir de dados denominados como dados do mundo real provenientes de prontuários eletrônicos do paciente em grande escala, considerando como método de representação e aprimoramento do conhecimento fundamental para representar as várias entidades e relacionamentos ao que tange o aprimoramento dos sistemas médicos apoiados por inteligência artificial (IA), sistemas de apoio à decisão clínica para diagnóstico e tratamento. Os autores destacaram a tentativa de trabalhos anteriores pela busca do avanço e desenvolvimento de um grafo de conhecimento automático, com o uso de técnicas de processamento de linguagem natural (PNL) para extrair o conhecimento automático de prontuários eletrônicos e reduzir esforços manuais. A metodologia aplicada na pesquisa abrangeu um conjunto de dados originais incluídos por visitas clínicas não identificadas e dados de diversos pacientes, e foram consideradas as etapas para o procedimento de construção do grafo de conhecimento, desde o reconhecimento de entidade, normalização de entidade, extração de relação, cálculo de propriedade, limpeza de gráficos, classificação de entidades relacionadas, até a incorporação de gráficos. E também, uma nova estrutura quádrupla para representar o conhecimento médico e o algoritmo de tradução probabilística em hiperplanos usado para aprender a incorporação do grafo de conhecimento gerado. Os resultados evidenciaram que o grafo de conhecimento utilizado para a representação de todas as entidades e relações, mostrou-se eficaz usando o agrupamento de doenças que permite aos profissionais da área da saúde localizarem e visualizarem rapidamente as principais informações dos registros do prontuário do paciente com alta qualidade. Concluiu-se, que os grafos de conhecimento possuem grande potencial em fornecer efetivamente informações qualificadas e servir como uma ferramenta de apoio a decisões voltadas a representação do conhecimento de domínio específicos.

Jagannatha *et al* (2019) em seu trabalho buscaram descrever a medicação, bem como, os eventos adversos a medicamentos relatados em dados provenientes de prontuários eletrônicos do paciente, visando dar uma visão geral de medicação, indicação medicamentosa e eventos adversos. A metodologia consistiu no uso de técnicas de reconhecimento de entidades nomeadas que identificam medicamentos específicos, bem como, algumas características, como indicação, dosagem, duração da medicação, frequência de uso e gravidade. O resultado indica que o uso de técnicas e algoritmos de processamento de linguagem natural, mostram-se como uma solução possível para resolver problemas relacionados à detecção automática e de

alta precisão de eventos adversos provocados por medicamentos, sendo relevantes para a segurança do medicamento aplicado. Conclui-se, que embora haja muito espaço para avanços na área, o uso de técnicas de processamento de linguagem natural apresenta notavelmente melhorias na resolução de problemas relacionados aos dados e informações de ordem medicamentosa.

Kormilitzin *et al* (2021) em seu trabalho buscaram treinar o modelo baseado em processamento de linguagem natural para recuperar dados de textos não-estruturados de pacientes que foram a princípio anotados à mão e posteriormente digitalizados. Este modelo foi treinado visando o reconhecimento de sete categorias, a se saber: nomes de medicamentos, vias de administração, frequência, dosagem, força, forma e duração. Basicamente, a metodologia usada consistiu em treinar o modelo de processamento de linguagem natural usando dados clínicos de anotações de textos livres de uma coleção de 2 milhões de registros de pacientes. Os resultados apontam que mesmo com a considerável semelhança entre os conjuntos de dados utilizados, é fundamental um ajuste rigoroso de domínio para se obter resultados e previsões mais precisas. Conclui-se, que o processamento de linguagem natural baseado em aprendizagem de transferência tem um papel fundamental na criação de modelos de previsão aplicável em domínios heterogêneos, dentro do contexto de recuperação de dados clínicos de prontuários de pacientes.

Turki *et al* (2022) empreenderam um estudo sistemático e aprofundado em relação aos desafios e oportunidades no aspecto de coordenar esforços para estruturar e formalizar o conhecimento relacionado aos dados da pandemia de COVID-19, por meio da rede interconectada, multidisciplinar, e a natureza internacional da pandemia e uso de grafos de conhecimento de origem coletiva. A metodologia aplicada no estudo, consistiu no uso da plataforma Wikidata, como base de conhecimento aberto, colaborativo e indisciplinar e disponível no formato RDF, que pode ser consultado de forma eficiente usando SPARQL, uma linguagem de consulta semântica para extrair dinamicamente informações triplas de gráficos de conhecimento em grande escala. Os resultados apontam a possibilidade de potencializar e sistematizar o conhecimento de forma computável com o propósito de identificar e acelerar a resposta ao patógeno e futuras epidemias através da integralização e representação de informações multidisciplinares disponível no formato RDF (Resource Description Framework) padronizado, do qual os dados são organizados em entidades e os relacionamentos associadas a doença infecciosa, patógeno subjacente, a pandemia resultante e tópicos relacionados. Conclui-se, que desta forma, o grafo de conhecimento criado para o COVID-19 no Wikidata, pode ser visualizado e explorado para propósitos como o apoio à tomada de decisão, bem como,

o mesmo possui potencial para desenvolver-se com qualidade e abrangência, suportando outros tipos de informações tanto no domínio de pesquisa educacional quanto acadêmica.

Zeng *et al* (2018) em sua pesquisa investigaram a efetividade do uso de um algoritmo de processamento de linguagem natural em prontuários eletrônicos. A metodologia empregada constituiu-se na extração de informações de dados clínicos contidos em prontuários eletrônicos do paciente, através do uso de técnicas computacionais de PLN, e foi fundamentada em uma estrutura sintática e semântica. Os resultados evidenciam que a aplicação de algoritmos e técnicas baseadas em processamento de linguagem natural, podem contribuir de modo significativo para extração de entidades nomeadas, bem como, suas relações com categorias clínicas, desta forma, otimizando a caracterização das relações de recursos em narrativas clínicas. Conclui-se, que o ajuntamento de fontes heterogêneas de dados para construção de um dataset abrangente, traz em si o potencial para melhorar de modo significativo o desempenho do modelo de extração de relações.

Benício (2020) buscou identificar dificuldades existentes no processamento e recuperação de dados de textos clínicos não-estruturados, e entender a relevância do uso de novas técnicas computacionais baseadas em processamento de linguagem natural para a recuperação de informações. Para isso, propôs uma ferramenta de recuperação de termos médicos das anamneses de prontuários eletrônicos, visando a estruturação de forma a relacionar com padrões de diagnósticos patológicos. Os resultados apontam que o uso de referidas técnicas de PLN baseadas em estruturas sintáticas e semânticas para extração de dados, promove a geração de informações mais compreensíveis e padronizadas. Conclui-se, que métodos de processamento de linguagem natural e mineração de texto, podem sanar problemas relacionados a recuperação de dados advindos de textos não-estruturados de anamneses clínicas de prontuários eletrônicos, levando a extração de dados mais compreensíveis e padronizados.

Sun *et al* (2021) em sua pesquisa investigaram o uso de openEHR, ou seja, de especificação de padrão informatizado em saúde para o gerenciamento, armazenamento e recuperação de dados clínicos em prontuários eletrônicos de pacientes, visando aperfeiçoar a interoperabilidade semântica de informações e resolver problemas de ambiguidade da recuperação da informação. A metodologia consistiu na aplicação de uma tecnologia de expansão de consulta no modelo Word2Vec (modelo de algoritmo de rede neural) envolvendo processamento de linguagem natural propondo a localização de sinônimos como substitutos dos termos de pesquisa originais na recuperação dos arquétipos. Conclui-se, que a abordagem que usa a tecnologia e corpus de processamento de linguagem natural para encontrar sinônimos como substitutos para termos de pesquisas, podem promover uma melhora significativa na

precisão e resolver a ambiguidade nas tarefas de recuperação da informação oriundas de dados clínicos.

Juhn *et al* (2020) investigaram em sua pesquisa como a adoção de sistemas de prontuários eletrônicos e a consequente geração de um grande volume de dados clínicos de pacientes, geram uma gama de novas possibilidades ao que se refere a realização de pesquisas clínicas. Nesse sentido, os autores focaram seus esforços no estudo de técnicas computacionais de processamento de linguagem natural, para obter insights a partir de dados ocultos em meio ao grande volume de dados. Os resultados evidenciam que técnicas de linguagem de processamento natural permitem de forma significativa o uso de gráficos automatizados para identificar pacientes com características clínicas específicas. Conclui-se, que técnicas de PLN são eficientes para a extração e descoberta de novas informações clínicas, e a utilização para aplicação em pesquisas clínicas.

Berman *et al* (2021) em seu trabalho buscaram identificar e avaliar a presença de doenças cardiovasculares, por meio do processamento e análise de grandes conjuntos de dados provenientes de prontuários eletrônicos do paciente. Para isso desenvolveram módulos baseados em modelos de processamento de linguagem natural para avaliar a presença de cinco comorbidades de origem cardiovasculares (a saber: hipertensão, dislipidemia, diabetes, doença arterial coronariana e acidente vascular cerebral/ataque isquêmico transitório.) A metodologia consistiu na análise de anotações clínicas de estudos cardiovasculares selecionados de forma aleatória, por profissionais da saúde especializados. Os resultados obtidos indicam uma alta eficácia do uso de módulos de processamento de linguagem natural, sendo que a percentagem de acertos de predição dos modelos de cada uma das cinco especificidades foi sempre acima de 85%. Os autores destacam que o nível de precisão dos modelos tende a ser mais elevada para condições de menor complexidade de diagnóstico (como, diabetes e hipertensão), e de precisão inferior para previsões de comorbidades com uma maior complexidade (como, infarto do miocárdio e acidente vascular cerebral embólico). Conclui-se que o uso de modelos de processamento de linguagem natural são eficazes e podem ser usados para avaliar a presença de doenças cardiovasculares em dados não-estruturados gerados em prontuários eletrônicos. E, desta forma, podem ser uma ferramenta importante no desenvolvimento de sistemas eletrônicos de registro médicos cardiovasculares e de armazenamento de referidas informações clínicas.

Koleck *et al* (2019) investigaram a aplicação de técnicas e modelos de processamento de linguagem natural para extrair e analisar informações de dados de anotações clínicas de texto não-estruturados advindos de prontuários eletrônicos do paciente. Os trabalhos analisados descrevem o uso de PLN em uma ampla gama de sintomas em diversas especialidades clínicas.

A metodologia foi estruturada por meio de uma revisão sistemática de literatura com abordagem qualitativa e objetivo exploratório, analisou-se as abordagens do processamento de linguagem natural, e incluíram ferramentas, métodos de classificação e processamento baseado em regras com curadoria manual. Os resultados obtidos mostram que especialmente há um foco atual no desenvolvimento de métodos de extração de informações relacionadas a sintomas, objetivando a descoberta, organização e classificação de doenças. Os autores concluem que para o desenvolvimento de algoritmos e modelos de processamento de linguagem natural eficazes para relacionar sintomas e comorbidades, é imprescindível que ocorra uma boa coleta de dados, com diagnósticos precisos do quadro de saúde dos pacientes. De forma, a efetivamente possibilitar descobertas informacionais contidas em dados oriundos de anotações clínicas de texto livres, e o uso destas para balizar tomadas de decisão. Os autores ainda observam que para que haja avanços é de suma importância que tanto os profissionais da área da computação, como profissionais da área da saúde, de modo alinhado disponibilizarem (respeitando o sigilo das informações sensíveis desses pacientes) dados em um formato aberto.

Wu *et al* (2022) investigaram a relevância e perspectiva da aplicação da inteligência artificial (IA) no domínio do COVID-19 combinados a análise bibliométrica e grafos de conhecimento para analisar a significância dos dados provenientes de variadas fontes de publicações, instituições e países que declararam que os artigos contribuíram para diagnosticar, rastrear, classificar e prever o COVID-19 por IA. A metodologia aplicada consistiu na análise interna no banco de dados cientifico Web of Science (WoS) para realizar a análise estatística, juntamente com a função de análise do software Citespace para criar o diagrama de sequência de citações da literatura no campo da IA aplicada na COVID-19, que reproduz o histórico e comparação dos resultados da análise estatística de WoS e Citespace, combinado com o software VOSviewer, as palavras-chave de amostra são examinadas por análise de co-palavras e análise de co-citação para obter a atual e a tendência de desenvolvimento futuro. Os resultados demonstraram que há cooperação entre instituições e países, e recomenda-se o fortalecimento e cooperações inter-regional. Quanto aos resultados da análise de palavras-chave, revelaram-se muito similares e confirmam a confiabilidade um do outro, contudo, a falta de dados abertos implica negativamente na avaliação do desempenho do modelo de IA. Conclui-se, que se torna evidente a importância no estudo relacionado a IA aplicada no COVID-19, pois, contribui no melhor entendimento sobre as tendências de IA aplicada na COVID-19 por meio do grafo de conhecimento. Assim como, a contribuição para o avanço da pesquisa e melhora na compreensão dos pesquisadores sobre o foco da pesquisa nesta área, e o progresso em processamento de linguagem natural (NLP), aprendizado de máquina (ML), aprendizado

profundo (DP), dados análise e outros campos, que demonstram o potencial de IA no suporte ao gerenciamento do sistema referido.

Silva *et al* (2019) apresentaram uma síntese dos mais recentes métodos de desenvolvimento no campo do aprendizado de máquina para inferência em grafos de conhecimento, e discutiram o valor e a importância de referidos ativos na representação do conhecimento e do raciocínio nos mais diversos domínios. Para este fim, investigaram métodos e técnicas de aprendizado de máquina empregadas em métodos de construção de grafos de conhecimento, e apresentaram os principais desafios e oportunidades tecno-científicas. O estudo traz um levantamento do estado da arte, onde é apresentado sucintamente a contextualização e o uso de técnicas e modelos de aprendizado de máquina aplicados às tarefas relacionadas a construção de grafos de conhecimento, bem como, algumas possíveis aplicações destes. Destacam-se das técnicas que foram apresentadas, um conjunto de modelos destinados à tarefa que complementam os grafos de conhecimento, principalmente as baseadas em aprendizado automatizado de representações para grafos. Dentre o referido conjunto de modelos apresentados estão: Modelos RDF e SPARQL; modelos de tarefas em grafos de conhecimento; de construção automatizada de bases e grafos de conhecimento; de extração de entidades e relacionamentos, e destaca-se também o emprego de ontologias e a avaliação e treinamento de modelos. Conclui-se que o considerável interesse nesse campo é devido a fatores, como: o modo natural como o conhecimento e a informação são representadas na forma de grafos, bem como, a atual imensurável geração e disponibilidade de dados digitais heterogêneos e multi-relacionais. E que, portanto, há em aberto uma vasta gama de oportunidades e desafios de pesquisa nesse campo, nos mais diversos domínios do conhecimento.

Lopes (2020), em sua pesquisa investigou os grafos de conhecimento com o objetivo de inferir as perspectivas atuais do desenvolvimento e emprego destes, para a organização e representação do conhecimento em diversos domínios e organizações. A pesquisa pretendeu identificar, além dos principais desafios para a representação do raciocínio e do conhecimento, inferir e expor as contribuições necessárias do campo da Ciência da Informação para a concepção, desenvolvimento e uso dos grafos de conhecimento. Para tal, o autor investigou métodos complementares de construção de grafos, como: Ontologias, Tesauros e Redes Semânticas, bem como, a convergência de variadas tecnologias para a construção de modelos de grafos semânticos altamente integrados. Segundo o autor, os resultados obtidos evidenciaram que os grafos de conhecimento essencialmente constituídos por meio do ajuntamento de estruturas de representação de conhecimento em domínios específicos, podem

promover a representação intuitiva do raciocínio e do conhecimento. E, sua construção decorrente da combinação de dados vinculados, constitui-se em uma arquitetura de solução otimizada e altamente interoperável. Conclui-se que a concepção, desenvolvimento e uso de grafos de conhecimento por conta de sua crescente importância e disseminação, se constitui em uma fonte de pesquisas atual e pertinente na área de Ciência da Informação, pois devido sua característica interdisciplinar, esta pode efetivamente favorecer a padronização e o desenvolvimento de instrumentos, voltados a representação do conhecimento de domínio específicos fundamentados em grafos. O autor aponta, que ainda existem inúmeros desafios ao que se refere a pesquisas de aspectos técnicos e científicos no campo de desenvolvimento de representação do conhecimento baseado em grafos. E que é prerrogativa da Ciência da Informação aliada, obviamente, a outras áreas de conhecimento como a Ciência da Computação, por meio de pesquisas, disponibilizar as orientações tecno-científicas necessárias para aqueles, que necessitam de aportes teóricos para compreender e consumir tais grafos, em diferentes domínios.

Kurbatova e Swiers (2021), desenvolveram uma solução de software de grafo de conhecimento (base de conhecimento Grakn) que utiliza integração de dados com referências cruzadas de ontologias visando facilitar a alternância entre hierarquias de ontologias para integração de dados hierárquicos de doenças. Para isso, utilizaram Processamento de Linguagem Natural, e buscaram soluções específicas para identificar correspondências entre variadas ontologias de doenças e, construíram um sistema baseado em grafos capaz de realizar consultas de dados hierarquizados de doenças, isso, com o intuito de mapear doenças de uma ontologia para outra. Os resultados demonstraram que a inclusão de dados presentes em ontologias de doenças melhorou significativamente, no entanto, ainda existem referências a termos obsoletos e ausência de correspondência. Conclui-se assim, que o referido sistema de integração de dados com referências cruzadas de ontologias de doenças facilita a construção de grafos de conhecimento biomédico, disponibilizando uma solução eficiente e eficaz para problemas relacionados a ontologias de múltiplas doenças. Entende-se também que referências cruzadas de doenças quando disponibilizadas em um arquivo simples e com acesso facilitado e editável tende a melhorar a correspondência ontológica em áreas de doenças específicas.

Ji *et al* (2021) empreenderam uma pesquisa aprofundada sobre grafos de conhecimento visando construir uma revisão abrangente do estado da arte desse campo. A pesquisa tem foco no âmbito da representação do conhecimento, métodos e principais aplicações, bem como, em avanços emergentes e perspectivas sobre pesquisas futuras. Para isso, analisaram as técnicas mais recentes aplicadas no desenvolvimento e geração de grafos de conhecimento, como:

aprendizagem de representação de conhecimento (Knowledge Representation Learning ou KRL), incorporação de conhecimento de grafos (Knowledge Graph Embedding ou KGE), mapeamento de entidades e relações de classificação tripla, reconhecimento de entidade nomeadas e extração de relações de entidades. Conclui-se que técnicas de aprendizagem automática ou aprendizado de máquina, e modelos de conhecimento fundamentados em grafos de conhecimento e baseados em ricas ontologias e modelos semânticos, beneficiam a integração de informações heterogêneas, e, por consequência, a representação do conhecimento em variados domínios.

Geleta *et al* (2021), desenvolveram um grafo de conhecimento para apoiar a criação de novos medicamentos, o modelo de grafo de nome Biological Insights Knowledge Graph (BIKG), conforme os autores o BIKG combina dados relevantes para o desenvolvimento de medicamentos. Esses referidos dados advêm de fontes públicas e internas, e são essenciais no fornecimento de Insights para tarefas como a identificação de novas medicações e redirecionamento de medicamentos existentes. Nesse artigo os autores esclarecem diferentes aspectos do ciclo de vida da construção do referido grafo, desde a obtenção de dados até a exploração destes. Entre outras, evidencia-se que com os mais recentes avanços em aprendizado de máquina, os grafos de conhecimento (além da própria representação do conhecimento) têm mais um propósito destacável, que é o de efetivamente servir como dados de treinamento em modelos de aprendizado de máquina. Uma conclusão importante é que dados de treinamento devem ser muito ponderados ao se construir grafos de conhecimento, posto que estes efetivamente influenciam nas escolhas de design. Para exemplificar, um modelo muito expressivo pode ser eficaz na captura de aspectos refinados de dados de domínio, mas pode concomitantemente criar dificuldades para a aplicação de aprendizado de máquina devido a problemas de escalabilidade.

Nicholson e Greene (2020) descrevem em seu trabalho abordagens para o desenvolvimento e aplicação de grafos de conhecimento na área da biomedicina. E com esse propósito apontam prós e contras de se construir grafos por meio de bancos de dados de forma não automática e por meio de sistemas de mineração de texto e, por fim, também apontam a eficácia prática da construção e uso dos grafos de conhecimento e futuras aplicações a serem exploradas. Os autores afirmam que a construção de grafos de conhecimento através de banco de dados preenchidos por especialistas por meio de curadoria manual, é, por si só, uma forma atualmente inviável (mesmo que estes contenham dados relativamente preciso), posto que, a quantidade de dados a serem analisados é extremadamente grande. Mas, esse processo pode ser usado para gerar conjuntos de dados padrão ouro, de forma que o processamento moderno de

linguagem natural por meio de métodos de aprendizado máquina possam aprender desses dados. Deste modo, identificam a importância do uso de técnicas de aprendizado de máquina, aplicados à mineração de texto e extração de relacionamento com base em regras, para que de forma automatizada possam obter rapidamente sentenças, padrões gramaticais e relacionamentos de um grande volume de dados textuais. Conclui-se que os grafos de conhecimento são, e cada vez mais serão amplamente utilizados para solução de problemas no campo da biomedicina, e que a utilização de algoritmos de aprendizado de máquina aplicados à mineração de textos tem um papel preponderante na obtenção automática e rápida de novas descobertas biomédicas, a partir de grafos de conhecimento. Por fim, os autores destacam que: "é um momento promissor para pesquisas sobre a construção e aplicação de grafos de conhecimento. Mesmo porque, a literatura revisada por pares está crescendo em uma taxa crescente e manter uma compreensão completa é cada vez mais desafiador para os cientistas".

Zimmermann *et al* (2015) o trabalho teve como objetivo estimar os fatores associados à qualidade de evidência e sua relação com as recomendações de incorporação de medicamentos da Comissão Nacional de Incorporação de Tecnologias em Saúde (Conitec). O estudo investigou os potenciais preditores de qualidade quanto às evidências apresentadas nos relatórios, juntamente com a relação com as recomendações emitidas pela Conitec no decorrer do período de 2012 a 2015. A metodologia utilizada incluiu a seleção de relatórios completos de avaliação de medicamentos, a identificação de variáveis relacionadas à qualidade de evidências e recomendações e a análise de dados por meio de software estatístico. Os resultados mostraram que foram avaliados 67 relatórios de recomendação de medicamentos, que representam 72% dos relatórios de medicamentos disponíveis na Conitec, e destacaram a importância da avaliação consistente da qualidade das evidências nas recomendações para a incorporação de medicamentos no SUS. Conclui-se que avaliar a qualidade das evidências se torna um fator essencial da Avaliação de Tecnologias em Saúde (ATS) sendo crucial para os processos de tomada de decisão no Sistema Único de Saúde (SUS). O estudo também identificou potenciais preditores de baixa qualidade de evidências, como doenças raras, fonte de demanda externa e ano do relatório.

Ávila *et al* (2019) apresentaram em seu trabalho o MediBot, um chatbot desenvolvido para fornecer informações referentes a medicamentos e seus riscos, ressaltando a importância de fornecer acesso a informações sobre medicamentos para o público em geral. Também discutiram os riscos da automedicação sem orientação médica e os possíveis problemas de saúde que dela podem advir. A metodologia utilizada para o desenvolvimento do MediBot foi baseada nas tecnologias Web Semântica e Linked Data, que permitiram a integração de dados

de variadas fontes e a representação de dados em um vocabulário unificado através do uso de ontologias. Os resultados indicaram a eficiência no uso dessa ferramenta que possibilita consultas por meio de linguagem natural, transformadas em consultas SPARQL sobre um Linked Data Mashup sobre dados de medicamentos fornecidos pelas fontes ANVISA e Sider. Dessa forma, os autores concluem que o MediBot, um chatbot baseado em ontologia, pode ser utilizado como ferramenta para acessar dados sobre medicamentos e seus riscos. E sugerem que trabalhos futuros para expandir o conjunto de consultas de resposta rápida e disponibilizar o MediBot em outras plataformas de mensagens apresenta-se como algo promissor para promover o acesso à informação pelo público em geral e reduzir os riscos de automedicação sem orientação médica.

Lin *et al* (2020) em sua pesquisa apresentaram um novo método chamado Knowledge Graph Neural Network (KGNN) que melhora a previsão de interações medicamentosas. Ele faz isso aproveitando os recursos de medicamentos e entidades relacionadas em um gráfico de conhecimento, o que permite a captura efetiva de medicamentos e de seus potenciais vizinhos. A metodologia consistiu na aplicação do KGNN que aprende com as vizinhanças de cada entidade no grafo como seu receptivo local e integra essas informações de vizinhança com o viés da representação da entidade atual, e assim, permite expandir naturalmente o campo receptivo a vários saltos para modelar informações topológicas de alta ordem e obter potenciais correlações de longa distância entre medicamentos. Os resultados da comparação com outros métodos mostraram que o KGNN supera significativamente as linhas de base nos dois conjuntos de dados apresentando um desempenho superior ao de outros métodos. Os autores concluem que a pesquisa apresentada o modelo KGNN supera os outros modelos de previsão DDI clássicos e de última geração em dois conjuntos de dados amplamente utilizados. O KGNN explora os recursos de medicamentos e entidades relacionadas no gráfico de conhecimento, enquanto os outros apenas aprendem com recursos de medicamentos semelhantes.

Chen *et al* (2021) abordaram em sua pesquisa o método de fusão de recursos MUFFIN baseado em aprendizado profundo que visa prever interações medicamentosas adversas (DDIs) usando a estrutura química do medicamento e o gráfico de conhecimento biomédico (KG). Ele integra recursos extraídos da estrutura molecular da droga e KG para aprimorar a capacidade de previsão de DDI. A metodologia adotada consistiu em um modelo que usa uma estratégia cruzada de dois níveis, com componentes de nível cruzado e escalar, para fundir recursos multimodais de forma eficiente. O MUFFIN pode aprender a representação do medicamento com base nas informações da sua estrutura e no KG com informações biomédicas ricas, o que ajuda a aliviar a restrição de dados rotulados limitados em modelos de aprendizado profundo.

Os resultados mostraram que o MUFFIN teve o melhor desempenho em tarefas de predição DDI de classe binária, multiclasse e multirrótulo, e que a arquitetura de dois níveis foi efetiva na combinação de recursos multimodais por meio do processo de fusão de recursos multi granularidade, melhorando assim a capacidade de previsão de DDIs. Desse modo, conclui-se que a abordagem proposta pode melhorar a capacidade preditiva de DDIs, considerando a sinergia entre a estrutura química da droga e o conhecimento biomédico.

Jouffroy *et al* (2021) destacaram em seu trabalho objetivaram o desenvolvimento de um sistema para extrair informações relacionadas a medicamentos de textos clínicos escritos em francês usando técnicas de processamento de linguagem natural (PLN). A metodologia delineou-se no estudo e construção de um sistema híbrido que combina um sistema especializado baseado em regras, incorporação de palavras contextuais treinadas em notas clínicas e uma rede neural recorrente profunda. Como resultado, o sistema foi avaliado usando recall, precisão e medida F em nível de token. A medida F geral foi de 89,9%, e as medidas F para cada categoria foram 95,3% para nome do medicamento, 64,4% para menções de classes de drogas, 95,3% para dosagem, 92,2% para frequência, 78,8% para duração e 62,2% para condição de admissão. Conclui-se que a associação de regras especializadas, incorporação contextualizada profunda e redes neurais profundas melhorou a extração de informações sobre medicamentos.

Zhang *et al* (2021) discutiram sobre uma nova abordagem para identificar candidatos a medicamentos para reaproveitamento no tratamento de COVID-19 usando conhecimento derivado da literatura e métodos de preenchimento de grafos de conhecimento. Esta abordagem envolve a extração de triplos semânticos usando o SemRep e a construção de um grafo de conhecimento usado para prever candidatos de reaproveitamento de medicamentos usando cinco algoritmos de conclusão de gráfico de conhecimento neural de última geração. A metodologia consistiu em modelos treinados e avaliados usando uma abordagem de fatiamento de tempo e os medicamentos previstos são comparados com uma lista de medicamentos relatados na literatura e avaliados em ensaios clínicos. Também foi realizada uma análise sobre o uso de padrões de descoberta para identificar medicamentos candidatos adicionais e gerar hipóteses plausíveis sobre as ligações entre os medicamentos candidatos e o COVID-19. Os resultados evidenciaram que a abordagem mostra-se viável não apenas para descobrir candidatos a medicamentos para o COVID-19, mas também para gerar explicações mecanicistas. Desse modo, os autores concluem que a abordagem pode ser viável não apenas para descobrir candidatos a medicamentos para COVID-19, mas também generalizada para outras doenças e questões clínicas.

Serafim *et al* (2022) em sua pesquisa buscaram descrever os produtos infocomunicacionais sobre medicamentos emitidos pelo Centro de Informações sobre Medicamentos da Universidade Federal de Sergipe Campus Lagarto (CIMUFS-LAG) como parte do serviço de informação proativa durante o período entre março de 2020 e fevereiro de 2022, no âmbito da pandemia de COVID-19. A metodologia abordada foi um estudo descritivo, e dirigido por métodos mistos, de caráter transversal e retrospectivo, realizado através de análise dos documentos referentes às informações proativas internas pelo CIMUFS-LAG. Os resultados demonstraram que durante o período compreendido entre março de 2020 a março de 2022, os CIM/SIMs relataram-se para os assuntos pandêmicos, acolhendo a população em geral, e trabalhadores de saúde através de atendimentos e divulgação de conteúdo informativo em grupos no Whatsapp, Facebook, Instagram, podcasts e nos endereços eletrônicos de cada um deles. Conclui-se a relevância do acesso à informação de qualidade, compreensível e amplamente divulgada como a melhor estratégia para combater a infodemia e o risco de iatrogenia decorrente da desinformação. No entanto, o estudo também aponta as limitações da investigação e a necessidade de mais pesquisas sobre o processo e desenvolvimento de produtos infocomunicacionais sobre medicamentos para promover seu uso seguro e racional.

Silva *et al* (2023) conduziram um estudo de caso-controle em um hospital terciário no sul do Brasil para comparar o desempenho de modelos de aprendizado de máquina com o Medication Fall Risk Score (MFRS) na previsão do risco de queda relacionado a medicamentos e classes de medicamentos. As características selecionadas para o modelo de predição foram medicamentos pertencentes às classes dos analgésicos, antipsicóticos, anticonvulsivantes, benzodiazepínicos, anti-hipertensivos, medicamentos cardíacos, antiarrítmicos, antidepressivos e diuréticos, as mesmas classes de medicamentos incluídas no MFRS. Os resultados indicaram que o modelo Naive Bayes teve um desempenho significativo ao se comparar com os outros algoritmos e MFRS na previsão do risco de queda. Portanto, os autores concluem que o modelo desenvolvido a partir desse conjunto de dados apresentou melhores resultados e sugerem que modelos preditivos construídos por algoritmos de aprendizado de máquina podem ser fortes aliados para identificar riscos para aprimorar o atendimento ao paciente.

3.1 Análise e Considerações Acerca das Demandas Informacionais

A análise do corpus documental deste estudo foi realizada buscando apontar possíveis contribuições da Ciência da Informação frente às demandas e resolução de problemas relacionados à informação no contexto do prontuário eletrônico do paciente. Estas demandas, como apontado anteriormente, foram identificadas por Galvão e Ricarte (2012) que conduziram

uma pesquisa que realizou um levantamento das demandas informacionais e tecnológicas, para otimização da organização e recuperação da informação contida em prontuários eletrônicos. Desta forma, foi possível delinear, investigar e apontar algumas contribuições necessárias da Ciência da Informação no âmbito informacional dos sistemas de prontuários eletrônicos. Como já citado, as referidas demandas fazem referência a dificuldades frente a: aquisição da informação, preservação da informação, identificação da informação, seleção da informação, organização da informação, análise da informação e comunicação da informação.

A partir da análise das literaturas selecionadas identificou-se que a Ciência da Informação pode gerar contribuições importantes para a organização, representação e recuperação da informação em sistemas de prontuários eletrônicos. Isso, por meio do uso de sistemas de organização da informação (taxonomia, tesauros, ontologias, redes semânticas e/ou grafos de conhecimento), que podem ser apoiados e/ou corroborar com métodos e técnicas atuais de processamento de linguagem natural. Assim, devido a característica multidisciplinar e colaborativa da Ciência da Informação, existem contribuições necessárias a serem feita pela área, que conforme Sant'Ana (2016) "[...] pode contribuir em ambientes que contam com a presença de acesso e uso intensivo de dados, buscando elementos que possibilitam a construção de estruturas de referências que permitam identificar características em contextos específicos."

Na tabela 1, com o intuito de delinear o escopo e o referencial dos trabalhos constituintes do corpus documental desta revisão. Bem como, com o objetivo de destacar as demandas de ordem informacional (representação e recuperação da informação e do conhecimento aplicáveis em sistemas de informação de prontuário eletrônico do paciente) e, que são abordadas em cada um dos trabalhos selecionados, segue uma sistematização contendo os atributos: ID (identificador), título, autoria e ano de publicação, campo de pesquisa, palavras-chave e demandas informacionais abordadas em cada um destes.

Tabela 1: Categorização referencial dos trabalhos selecionados e apontamento de demandas informacionais abordadas em cada trabalho.

ID	**DADOS REFERENCIAIS DOS TRABALHOS SELECIONADOS**	
1	**TÍTULO: 1**, Autor(es e/ou as) e Ano de publicação	**O prontuário eletrônico do paciente no século XXI: as contribuições necessárias da Ciência da Informação** - (GALVAO, Maria Cristiane Barbosa; RICARTE, Ivan Luiz Marques) **- 2011.**
	Campo (s) de pesquisa	Ciências da Informação
	Palavras-chaves	Prontuário Eletrônico do Paciente; Ciência da Informação; Processos Informacionais; Informação Clínica

	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização e Recuperação, Análise e Comunicação
2	**TÍTULO: 2**, Autor(es e/ou as) e Ano de publicação	**Análise de Dados Clínicos Textuais de Prontuários Eletrônicos do Paciente para Integração com Terminologias Médicas Padronizadas –** (SOUZA, Amanda D; ALMEIDA, Maurício B. de.) **– 2019.**
	Campo (s) de pesquisa	Gestão e Organização do Conhecimento
	Palavras-chaves	Prontuários Eletrônicos do Paciente, Terminologias Médicas, PLN, Extração de Dados, Sistemas de Informação em Saúde
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização e Recuperação, e Análise
3	**TÍTULO: 3**, Autor(es e/ou as) e Ano de publicação	**Graph-Representation of Patient Data: a Systematic Literature Review –** (SCHRODT, Jens) **- 2020**
	Campo (s) de pesquisa	Tecnologias da Informação e Biometria Médica
	Palavras-chaves	Graph Theory, Systematic Literature Review, Electronic Health Record, Temporal Patient Graph
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização e Recuperação, e Análise
4	**TÍTULO: 4**, Autor(es e/ou as) e Ano de publicação	**Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidences -** (WANG, Lei; XIE, Huimin; HAN, Wentao) **- 2020.**
	Campo (s) de pesquisa	Informática Médica
	Palavras-chaves	Knowledge Graph; Diabetes; Evidence-Based Medicine; Risk Prediction
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização e Recuperação, Análise e Comunicação
5	**TÍTULO: 5**, Autor(es e/ou as) e Ano de publicação	**Natural language processing to extract symptoms of severe mental illness from clinical text -** (JACKSON, Richard G; PATEL, Rashmi) **- 2016.**
	Campo (s) de pesquisa	Biomedicina e Tecnologia da Informação
	Palavras-chaves	Natural Language Processing, Data Extraction, Severe Mental Illness
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção e Análise
6	**TÍTULO: 6**, Autor(es e/ou as) e Ano de publicação	**Early Recognition of Multiple Sclerosis Using Natural Language Processing of the Electronic Health Record -** (CHASE, Herbert S.; MITRANI, Lindsey) **– 2017.**
	Campo (s) de pesquisa	Biomedicina e Tecnologia da Informação
	Palavras-chaves	Early Diagnosis, Electronic Health Records, Natural Language Processing, Multiple Sclerosis

	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação e Análise
7	**TÍTULO: 7**, Autor(es e/ou as) e Ano de publicação	**Real-world data medical knowledge graph: construction and applications -** (LI, Linfeng; WANG, Peng; YANB, Jun) **– 2020.**
	Campo (s) de pesquisa	Inteligência Artificial, Medicina e Informação
	Palavras-chaves	Real-World Data Medical, Knowledge Graph
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização e Recuperação, e Análise
8	**TÍTULO: 8,** Autor(es e/ou as) e Ano de publicação	**Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes -** (JAGANNATHA, Abhyuday; LIU, Feifan; LIU, Weisong) **- 2019.**
	Campo (s) de pesquisa	Ciências da Informação, Ciências da Computação, Saúde
	Palavras-chaves	Electronic Health Record Notes, Natural Language Processing, Named Entity Recognition
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
9	**TÍTULO: 9**, Autor(es e/ou as) e Ano de publicação	**Med7: a transferable clinical natural language processing model for electronic health records -** (KORMILITZINA, Andrey; VACIA, Nemanja; LIUA, Qiang) **- 2020.**
	Campo (s) de pesquisa	Inteligência Artificial, Ciência da Computação, Processamento de Linguagem Natural
	Palavras-chaves	Clinical Natural Language Processing, Neural Networks, Self-Supervised Learning, Active Learning
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção e Análise
10	**TÍTULO: 10**, Autor(es e/ou as) e Ano de publicação	**Representing COVID-19 information in collaborative knowledge graphs: The case of Wikidata** - (TURKI, Houcemeddine; TAIEB, Mohamed Ali Hadj; SHAFEE, Thomas) **- 2022.**
	Campo (s) de pesquisa	Informação, Grafos de Conhecimento e Biomedicina
	Palavras-chaves	Wikidata, Knowledge Graph, Covid-19, SPARQL, Linked Open Data
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização **e** Recuperação, Análise
11	**TÍTULO: 11**, Autor(es e/ou as) e Ano de publicação	**Natural Language Processing for EHR-Based Computational Phenotyping -** (ZENG, Zexian; DENG, Yu; LI, Xiaoyu; NAUMANN, Tristan; LUO, Yuan) **- 2018.**
	Campo (s) de pesquisa	Processamento de Linguagem Natural, Informação e Prontuários Eletrônicos, Aprendizado de Máquina, Ciência da Computação

	Palavras-chaves	Electronic Health Records, Natural Language Processing, Computational Phenotyping, Machine Learning
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização **e** Recuperação, e Análise
12	**TÍTULO: 12**, Autor(es e/ou as) e Ano de publicação	**Aplicação de mineração de texto e processamento de linguagem natural em prontuários eletrônicos de pacientes para extração e transformação de texto em dados-estruturados** (BENÍCIO, Diego Henrique Pegado) **- 2020.**
	Campo (s) de pesquisa	Processamento de Linguagem Natural, Prontuários Eletrônicos de Pacientes e Tecnologia da Informação
	Palavras-chaves	Mineração de Texto; Processamento de Linguagem Natural; Anamnese; Prontuário Eletrônico.
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização **e** Recuperação, e Análise
13	**TÍTULO: 13**, Autor(es e/ou as) e Ano de publicação	**Using NLP in openEHR archetypes retrieval to promote interoperability: a feasibility study in China** (SUN, Bo; ZHANG, Fei; LI, Jing) **- 2021.**
	Campo (s) de pesquisa	Prontuário Eletrônico, Processamento de Linguagem Natural, Rede de Representação Semântica, Recuperação da Informação
	Palavras-chaves	Openehr, Nature Language Processing, Information Retrieval, Interoperability
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
14	**TÍTULO: 14**, Autor(es e/ou as) e Ano de publicação	**Hongfang. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. Journal of Allergy and Clinical Immunology** - (JUHN, Young; LIU, Hongfang) **- 2020.**
	Campo (s) de pesquisa	Prontuários Eletrônicos, Informação, Mineração de Dados, Aprendizado de Máquina, Processamento de Linguagem Natural
	Palavras-chaves	Ehrs, Asthma, Allergy, Informatics, Data Mining, Machine Learning, Natural Language Processing, Algorithms, Artificial Intelligence
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Análise e Comunicação
15	**TÍTULO: 15**, Autor(es e/ou as) e Ano de publicação	**Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project** - (BERMAN, Adam N.; BIERY, David W.; GINDER, Curtis) **- 2021.**
	Campo (s) de pesquisa	Saúde, Informação e Processamento de Linguagem Natural
	Palavras-chaves	Cardiovascular Comorbidities, Natural Language Processing

	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção e Análise
16	**TÍTULO: 16**, Autor(es e/ou as) e Ano de publicação	**Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review -** (KOLECK, Theresa A.; DREISBACH, Caitlin; BOURNE, Philip E.; BAKKEN, Suzanne) - **2019.**
	Campo (s) de pesquisa	Prontuário Eletrônico, Informação e Processamento de Linguagem Natural
	Palavras-chaves	Natural Language Processing, Signs and Symptoms, Electronic Health Records, Review
	Contextos informacionais abordados direta ou indiretamente	Preservação, Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
17	**TÍTULO: 17**, Autor(es e/ou as) e Ano de publicação	**Knowledge graph analysis and visualization of AI technology applied in COVID-19** - (WU, Zongsheng; XUE, Ru; SHAO, Meiyun) **- 2022.**
	Campo (s) de pesquisa	Informação, Grafos de Conhecimento e Visualização de Dados
	Palavras-chaves	COVID-19, Coronavirus Disease, Knowledge Graph, AI, Data Visualization, Visual Analysis
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
18	**TÍTULO: 18**, Autor(es e/ou as) e Ano de publicação	**Aprendizado de máquina e inferência em Grafos de Conhecimento -** (SILVA, Daniel N. R.; ZIVIANI, Artur; PORTO, Fábio) **- 2019.**
	Campo (s) de pesquisa	Informação, Representação Computacional, Aprendizado de máquina e Grafos de Conhecimento
	Palavras-chaves	Aprendizado de máquina, Grafos de Conhecimento. Inferência
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
19	**TÍTULO: 19**, Autor(es e/ou as) e Ano de publicação	**Grafos de Conhecimento: perspectivas e desafios para a organização e representação do conhecimento -** (LOPES, Dener Cesar Ferreira) **- 2020.**
	Campo (s) de pesquisa	Informação e Grafos de Conhecimento
	Palavras-chaves	Grafos de Conhecimento. Representação do Conhecimento. Organização da Informação. Tecnologias Semânticas
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, e Comunicação
20	**TÍTULO: 20**, Autor(es e/ou as) e Ano de publicação	**Disease ontologies for knowledge graphs** - (KURBATOVA, Natalja; SWIERS, Rowan) **– 2021.**
	Campo (s) de pesquisa	Ontologia, Informação Médica, Grafos de Conhecimento e Processamento de Linguagem Natural
	Palavras-chaves	Ontologies, Knowledge graph, Data integration

	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização **e** Recuperação, e Comunicação
21	**TÍTULO: 21**, Autor(es e/ou as) e Ano de publicação	**A Survey on Knowledge Graphs: Representation, Acquisition and Applications** - (JI, Shaoxiong; PAN, Shirui; CAMBRIA, Erik; MARTTINEN, Pekka; YU, Philip, S.) - **2021.**
	Campo (s) de pesquisa	Representação da informação e Grafos de Conhecimento
	Palavras-chaves	Knowledge Graph, Representation Learning, Knowledge Graph Completion, Relation Extraction, Reasoning, Deep Learning
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, e Comunicação
22	**TÍTULO: 22**, Autor(es e/ou as) e Ano de publicação	**Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development** - (GELETA, David; NIKOLOV, Andriy; EDWARDS, Gavin) - **2021.**
	Campo (s) de pesquisa	Biomedicina, Grafos de Conhecimento, Processamento de Linguagem Natural, Aprendizado de Máquina
	Palavras-chaves	Biological Insights, Drug Development, Knowledge Graph, Machine Learning, NLP
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
23	**TÍTULO: 23**, Autor(es e/ou as) e Ano de publicação	**Constructing knowledge graphs and their biomedical applications** - (NICHOLSON, David N; GREENE, Casey S) - **2020.**
	Campo (s) de pesquisa	Biomedicina, Grafos de Conhecimento, Processamento de Linguagem Natural, Aprendizado de Máquina
	Palavras-chaves	Knowledge Graphs, Network Embeddings, Text Mining, Natural Language Processing, Machine Learning, Literature Review
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização **e** Recuperação, Análise e Comunicação
24	**TÍTULO: 24**, Autor(es e/ou as) e Ano de publicação	**A qualidade das evidências e as recomendações sobre a incorporação de medicamentos no Sistema Único de Saúde: uma análise retrospectiva** - (ZIMMERMANN, I. R., de Oliveira, E. F., Vidal, Á. T., Santos, V. C. C., & Petramale, C. A) - **2015**
	Campo (s) de pesquisa	Avaliação da qualidade, Medicamentos no Sistema Único de Saúde, Tomada de decisão, Qualidade das evidências
	Palavras-chaves	Avaliação de Medicamentos, Avaliação da Tecnologia Biomédica, Modelos Logísticos, Tomada de Decisões, Sistema Único de Saúde
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização, Recuperação e Análise

25	**TÍTULO: 25**, Autor(es e/ou as) e Ano de publicação	**MediBot: Um chatbot para consulta de riscos e informações sobre medicamentos** - (AVILA, C. V. S., Rolim, T. V., da Silva, J. W. F., & Vidal, V. M. P) - **2019**
	Campo (s) de pesquisa	Ontologia, Informação Médica, Medicamento, Recuperação da Informação
	Palavras-chaves	Information Retrieval, Medicines, Ontologies, Risks of self-medication, MediBot
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização, Recuperação, Análise e Comunicação
26	**TÍTULO: 26**, Autor(es e/ou as) e Ano de publicação	**KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction** - (LIN, X., Quan, Z., Wang, Z. J., Ma, T., & Zeng, X.) - **2020**
	Campo (s) de pesquisa	Grafos de Conhecimento, Representação do Conhecimento. Organização da Informação, Tecnologias Semânticas
	Palavras-chaves	Knowledge Graph Neural Network, Medicines, Information Retrieval, Prediction
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Identificação, Seleção, Organização e Recuperação, e Análise
27	**TÍTULO: 27**, Autor(es e/ou as) e Ano de publicação	**MUFFIN: multi-scale feature fusion for drug–drug interaction prediction** - (CHEN, Y., Ma, T., Yang, X., Wang, J., Song, B., & Zeng, X) - **2021**
	Campo (s) de pesquisa	Aprendizado de máquina, Inteligência artificial, Saúde, Medicamento, Interações medicamentosas adversas, Farmacologia, Bioinformática
	Palavras-chaves	Artificial intelligence, Machine learning, Health, Medicine, Adverse drug interactions, Pharmacology
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização, Recuperação, e Comunicação
28	**TÍTULO: 28**, Autor(es e/ou as) e Ano de publicação	**Hybrid Deep Learning for Medication-Related Information Extraction From Clinical Texts in French: MedExt Algorithm Development Study** - (JOUFFOY, J., Feldman, S. F., Lerner, I., Rance, B., Burgun, A., & Neuraz, A) - **2021**
	Campo (s) de pesquisa	Processamento de linguagem natural, Medicina, Medicamentos, Textos clínicos, Aprendizado profundo, Redes neurais, Sistema híbrido
	Palavras-chaves	Medication information, Natural language processing, Electronic health records, Deep learning, Rule-based system, Recurrent neural network, Hybrid system
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização, Recuperação, e Análise

29	**TÍTULO: 29**, Autor(es e/ou as) e Ano de publicação	**Drug repurposing for COVID-19 via knowledge graph completion** - (ZHANG, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., & Kilicoglu, H) - **2021**
	Campo (s) de pesquisa	COVID-19, Reaproveitamento de medicamentos, Grafo de conhecimento, Inteligência artificial, Descoberta baseada em literatura, Mineração de texto
	Palavras-chaves	COVID-19, Drug repurposing, Knowledge graph, Literature-based discovery, Text mining
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização, Recuperação, Análise e Comunicação
30	**TÍTULO: 30**, Autor(es e/ou as) e Ano de publicação	**Informações proativas emitidas por um centro de informações sobre medicamentos na pandemia de COVID-19 no período de 2020 a 2022** - (SERAFIM, J. E. F., Matos, L. E. O., & Unfer, T. C) - **2022**
	Campo (s) de pesquisa	Medicamentos, Saúde, Comunicação, Informação, Acesso à informação, Medicina baseada em evidências
	Palavras-chaves	Uso de medicamentos, Comunicação em saúde, Informação em saúde, Acesso à informação, Medicina baseada em evidências
	Contextos informacionais abordados direta ou indiretamente	Aquisição, Preservação, Identificação, Seleção, Organização, Recuperação, Análise e Comunicação
31	**TÍTULO: 31**, Autor(es e/ou as) e Ano de publicação	**Risco de queda relacionado a medicamentos em hospitais: abordagem de aprendizado de máquina** - (silva, A. P. D., Santos, H. D. P. D., Rotta, A. L. O., Baiocco, G. G., Vieira, R., & Urbanetto, J. D. S) - **2023**
	Campo (s) de pesquisa	Medicina, Farmacologia, Aprendizado de máquina, Análise de dados
	Palavras-chaves	Accidental falls, Drug utilization, Supervised machine learning, Patient safety
	Contextos informacionais abordados direta ou indiretamente	Identificação, Seleção, Organização, Recuperação, Análise e Comunicação

Fonte: Elaborado pela autora (2022).

A análise do corpus documental básico deste estudo foi realizada buscando analisar as contribuições e os esforços frente às demandas e resolução de problemas relacionados à informação e o conhecimento no contexto do prontuário eletrônico do paciente e/ou sistemas informacionais análogos. Evidentemente, o uso do PEP trouxe novas demandas como apontam Galvão e Ricarte (2011), que conduziram uma pesquisa que realizou um levantamento das principais demandas informacionais e tecnológicas, para otimização da organização e

recuperação da informação contida em prontuários eletrônicos. Bem como, contribuições necessárias da Ciência da Informação, para esse esforço multidisciplinar.

Dentre as referidas demandas apontadas pelos autores, considerou-se nesta pesquisa as que fazem referência a dificuldades frente a: aquisição da informação, preservação da informação, identificação da informação, seleção da informação, organização e recuperação da informação, análise da informação e comunicação da informação.

Conforme a análise do corpus documental e perante as demandas apontadas nas pesquisas de Galvão e Ricarte (2011). Ou seja: a aquisição da informação, preservação, a identificação, seleção, organização, a recuperação, a análise e a comunicação da informação são ou se constituem pontos essenciais a serem tratados no campo da saúde e, em especial na gestão da informação em prontuários eletrônicos. Nesse ponto, é possível afirmar que importância de referidos processos se dá, posto que:

Aquisição da Informação: Tem por prerrogativa garantir que profissionais da informação recebam dados e informações relevantes de fontes externas aos prontuários eletrônicos e às instituições de saúde. Isso permite que tenham acesso a informações atualizadas, as quais em um outro momento são fundamentais para a tomada de decisões e aprimoramento dos serviços prestados em uma instituição de Saúde.

Preservação da Informação: Tem sua importância, posto que, assegura a integridade e a durabilidade das informações, especialmente no ambiente digital, onde os documentos podem estar sujeitos a fragilidades e a obsolescência tecnológica. Desta forma, a preservação quando feita de forma adequada evita a perda de dados cruciais, mantendo o histórico médico dos pacientes acessível e protegido ao longo do tempo.

Identificação da Informação: Por sua vez, permite a correta organização e recuperação da informação nos prontuários eletrônicos. Assim, a classificação adequada de dados como pacientes, comorbidades, medicamentos e procedimentos facilita a localização rápida e precisa de informações específicas quando necessárias. Além disso, a aplicação de técnicas computacionais atuais auxilia na estruturação e na compreensão dos dados, melhorando a interoperabilidade e a qualidade da informação.

Seleção da Informação: Contribui para que se estabeleça critérios e políticas para selecionar e atualizar as informações registradas nos prontuários eletrônicos. Em um ambiente de saúde em constante evolução, é fundamental garantir a relevância e a atualização dos dados, considerando avanços científicos, tecnológicos e políticas de saúde. Isso contribui para uma tomada de decisão informada, baseada em dados.

Organização e Recuperação da Informação: Promove a eficiência e a eficácia na gestão da informação em prontuários eletrônicos. Posto que, uma organização adequada facilita a localização de dados e informações necessários para o atendimento ao paciente, pesquisa e ensino em saúde. Além disso, uma recuperação eficiente permite o acesso rápido e preciso a informações relevantes, promovendo um atendimento de qualidade e uma tomada de decisão embasada. A organização e recuperação eficazes também contribuem para a pesquisa científica, permitindo o uso desses dados para estudos e análises diversas, além de facilitar a colaboração entre profissionais de saúde.

Análise da Informação: Por sua vez, a análise dos dados contidos ou gerados nos prontuários eletrônicos podem oferecer insights valiosos para o contexto do atendimento ao paciente, bem como, para o aprimoramento dos processos. Ao coletar e sintetizar informações específicas de um paciente, os profissionais podem identificar padrões, tendências e correlações relevantes. Isso possibilita uma abordagem personalizada no tratamento, otimizando os resultados e melhorando a eficiência dos recursos de saúde. Além disso, a análise da informação apoia a criação de políticas públicas embasadas em dados concretos, fortalecendo o planejamento estratégico no campo da saúde.

Comunicação da Informação: Finalmente, a padronização e a harmonização das nomenclaturas clínicas facilitam a comunicação efetiva entre equipes multiprofissionais e a interoperabilidade entre diferentes sistemas de informação em saúde. Essa comunicação fluida e compreensível permite uma troca de informações eficiente, promovendo a segurança do paciente, a continuidade do cuidado e a colaboração interdisciplinar. Além disso, a manutenção adequada das terminologias clínicas contribui para a observação de dados consistentes ao longo do tempo, facilitando a análise e a pesquisa clínica e acadêmica.

Posto isso, a pesquisa e aprofundamento do corpus documental evidenciou a importância atual e urgente da melhoria dos processos de aquisição, preservação, identificação, seleção, organização, recuperação, análise e comunicação da informação no contexto dos prontuários eletrônicos, E identifica que estas estão relacionada a uma melhor gestão dos cuidados de saúde, a uma tomada de decisão informada, à pesquisa científica e ao desenvolvimento de políticas públicas embasadas em evidências. A otimização desses processos permitirá um atendimento eficiente, personalizado e seguro, contribuindo para a melhoria contínua dos sistemas de saúde e o bem-estar dos pacientes.

Assim, a partir da análise das literaturas selecionada identificou-se contribuições importantes principalmente para a organização, representação e recuperação da informação em sistemas de prontuários eletrônicos e/ou sistemas da informação relacionados a saúde de modo

geral. Isso, por meio do uso de variadas tecnologias e de sistemas de organização da informação (taxonomia, tesauros, ontologias, redes semânticas e/ou grafos de conhecimento), usados, por vezes, em conjunto com métodos e técnicas atuais de processamento de linguagem natural.

Assim, devido a característica multidisciplinar e colaborativa da Ciência da Informação, ainda existem contribuições necessárias a serem feita pela área, que conforme Sant'Ana (2016) "[...] pode contribuir em ambientes que contam com a presença de acesso e uso intensivo de dados, buscando elementos que possibilitam a construção de estruturas de referências que permitam identificar características em contextos específicos."

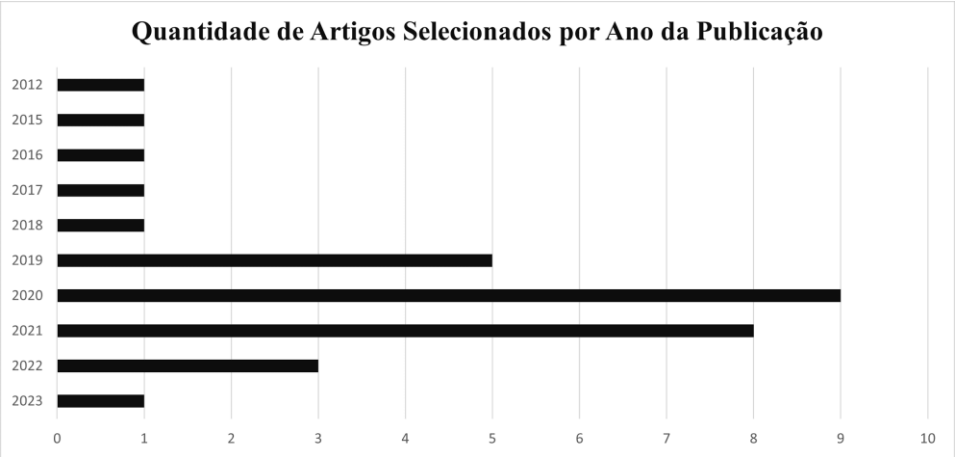
A pesquisa e aprofundamento bibliográfico foram realizados como fundamentação e aprofundamento teórico da pesquisa objetivando identificar e analisar aspectos importantes da recuperação e visualização da informação e do conhecimento no contexto médico. Especificamente, propôs-se analisar o uso de abordagens atuais de processamento de linguagem natural e grafos de conhecimento, e, entender como estas podem e, têm contribuído para atender ou contribuir com a liquidação das demandas informacionais ligadas a: aquisição, preservação, identificação, seleção, organização, análise e comunicação da informação e do conhecimento contido em sistemas de prontuários eletrônicos.

Um primeiro aspecto importante a se destacar é um evidente e crescente esforço de pesquisa no sentido de investigar a aplicação de métodos e tarefas de processamento de linguagem natural e de estruturação em grafos de conhecimento para a otimização da recuperação das informações contidas nos prontuários eletrônicos, e, em variados sistemas de informação de fonte médica. Especialmente, a recuperação e visualização de informações imersas em dados não estruturados. Esses esforços acontecem de forma generalizada em todas as áreas da medicina e saúde, e propõem, entre muitas outras, gerar insights para novos medicamentos, propor melhores tratamentos e/ou abordagem terapêuticas, detectar doenças precocemente ou simplesmente possibilitar e promover acesso as informações importantes de histórico de pacientes. Enfim, as possibilidades de abordagem e aplicações são infindas.

Nesse sentido, conforme Nicholson e Greene (2020), os grafos de conhecimento estão se tornando amplamente utilizados na medicina e biomedicina, e, juntamente com o processamento de linguagem natural pode interpretar perguntas simples e usar informações relacionais para fornecer respostas relevantes para doenças, reaproveitamento de drogas e interações de medicamentos. Já conforme Chase *et al*. (2017), erros médicos e precisão de diagnósticos podem ser melhorados com sistemas que reconhecem doenças, extraindo informações a partir de notas clínicas dos pacientes em busca de sinais e sintomas usando técnicas de processamento de linguagem natural. O que em última análise pode encurtar o

tempo de atendimento, bem como, agilizar o reconhecimento formal de uma dada doença, e, por conseguinte, otimizar o tempo de resposta e/ou o início do tratamento.

Gráfico 1: Número de artigos selecionados por ano da publicação.



Fonte: Elaborado pela autora (2022).

Também é válido destacar que conforme apontava a pesquisa de Galvão e Ricarte (2011), que como já visto empreenderam um levantamento profundo das principais demandas de ordem informacional relacionadas ao necessário melhoramento dos sistemas de prontuários eletrônicos. Onde os resultados de referida pesquisa demostraram que as demandas informacionais mais importantes identificadas estão relacionadas (como visto) as dificuldades de: aquisição, preservação, identificação, seleção, organização, análise e comunicação da informação, infere-se que isso em parte, se deve ao alto volume, heterogeneidade e formato dos dados, que em sua maioria estão armazenadas em formato não estruturado.

Posto isso, e diante da análise do corpus documental desta revisão bibliográfica é possível afirmar que referidas demandas informacionais podem e estão começando a serem sanadas com o uso de variadas tarefas de PLN, Aprendizado de Máquina e Grafos de Conhecimento. Observa-se que há várias pesquisas que mesclam tais técnicas e métodos para solucionar problemas informacionais e de construção do conhecimento médico/paciente.

Nesse sentido, identificou-se que dos 31 trabalhos selecionados e analisados nesta pesquisa, 84% em algum momento mesclam aplicações de PLN, Aprendizado de Máquina e/ou Grafos de Conhecimento para a recuperação e/ou visualização de informações médicas variadas. Sendo que 55% do total de trabalhos analisados na pesquisa faz uso grafos de conhecimento associado ao aprendizado de máquina, ou inversamente, uso de aprendizado de máquina associados a grafos de conhecimento para construir modelos para recuperação e visualização da informação médica. Ou seja, é perfeitamente possível usar estruturas de grafos

de conhecimento para construir modelos de aprendizado de máquina para recuperar informação médica especificas. Ou, usar modelos de aprendizado de máquina para construir grafos de conhecimento para visualizar referidas informações.

Gráfico 2: Número de publicações em relação as demandas informacionais abordadas.

Fonte: Elaborado pela autora (2022).

Como indica o gráfico 2, com relação a abordagem de demandas informacionais identificadas em sistemas de prontuários eletrônicos. Destaca-se que dentre as pesquisas componentes do corpus documental deste trabalho, as questões ligadas a identificação e seleção da informação é de especial interesse. Isso, parece muito natural, posto que, para promover a recuperação da informação contida em tais sistemas é imprescindível usar procedimentos para identificar e selecionar as informações realmente importantes, tanto no contexto médico/paciente, como na pesquisa médica de modo geral.

A análise da literatura também evidenciou que existe uma preocupação genuína e atual no sentido de se buscar a otimização constante dos sistemas eletrônicos de registros clínico-socias, especificamente, busca-se a melhoria constante da representação da informação nos prontuários eletrônicos. Fica evidente que este esforço conjunto é um movimento internacional, e, de forma geral objetiva a representação do conhecimento médico visando especialmente a melhor interoperabilidade possível. Nesse sentido, através da análise do corpus documental selecionado, foi possível identificar e compreender a importância crescente do uso de modelos de processamento de linguagem natural e de grafos de conhecimento.

Os estudos analisados demonstram claramente a aplicabilidade e eficiência das técnicas avançadas de aprendizado de máquina e processamento de linguagem natural na área de medicamentos, mais especificamente para a recuperação da informação. Eles ilustram como essas tecnologias podem ser utilizadas para extrair informações relevantes de textos clínicos,

melhorar a previsão de interações medicamentosas, identificar candidatos a medicamentos para o tratamento de doenças específicas.

O trabalho de Ávila et al (2019) destaca a utilização da Web Semântica e Linked Data no desenvolvimento de um sistema capaz de fornecer informações sobre medicamentos e seus riscos. Esse tipo de abordagem permite que dados de diferentes fontes sejam integrados e utilizados para melhorar a precisão e a qualidade das informações fornecidas aos usuários.

O método proposto por Lin et al (2020), chamado de Knowledge Graph Neural Network (KGNN), utiliza recursos de medicamentos e entidades relacionadas em um gráfico de conhecimento para aprimorar a previsão de interações medicamentosas. Isso demonstra a importância de se considerar as relações entre os medicamentos e outras entidades do conhecimento médico para obter resultados mais precisos. Chen et al (2021) apresentaram um método que utiliza aprendizado profundo juntamente com grafos de conhecimento biomédico, para prever interações medicamentosas adversas. Essa abordagem mostra como a combinação de diferentes tipos de dados e o uso de técnicas avançadas de aprendizado de máquina podem contribuir para a identificação e prevenção de riscos relacionados ao uso de medicamentos.

Jouffroy et al (2021) desenvolveram um sistema que extrai informações relacionadas a medicamentos de textos clínicos usando técnicas de processamento de linguagem natural. Essa aplicação mostra como a análise automatizada de grandes volumes de texto pode ser útil na obtenção de informações relevantes e na organização de dados para apoiar a tomada de decisões relacionadas ao uso de medicamentos. O trabalho de Zhang et al (2021) propõe uma abordagem para identificar candidatos a medicamentos para o reaproveitamento no tratamento de COVID-19, utilizando conhecimentos derivados da literatura e métodos de preenchimento de grafos de conhecimento. Essa abordagem destaca como a utilização de bases de conhecimento e a integração de dados de diferentes fontes podem ajudar na busca de soluções eficazes para o tratamento de doenças específicas.

Além disso, os estudos de Serafim et al (2022) e Silva et al (2023) demonstram como as técnicas de aprendizado de máquina e PLN podem ser aplicadas em contextos específicos, como a emissão de produtos infocomunicacionais sobre medicamentos resultando em melhores resultados e decisões em relação ao uso de medicamentos. Ainda, a abordagem de grafos proposta por Lin et al (2020) que aproveita os recursos de medicamentos e entidades relacionadas em um gráfico de conhecimento, melhorando a previsão de interações medicamentosas. Sugere que esse tipo de abordagem pode lidar com a complexidade de informações interrelacionadas de medicamentos, e pode acelerar o acesso fácil a dados valiosos

como: interações medicamentosas, efeitos colaterais e contraindicações, permitindo que os profissionais de saúde tomem decisões embasadas com base em evidências.

Por fim, a integração de dados de diferentes fontes e o uso de técnicas de processamento e análise de dados textuais permitem uma compreensão mais abrangente dos medicamentos, seus efeitos e riscos associados. Essas abordagens baseadas em dados fornece uma base sólida para a tomada de decisões informadas pelos profissionais de saúde. Além disso, esses estudos também destacam a importância da colaboração interdisciplinar na área da saúde. A integração de conhecimentos da informática, ciência da computação, farmacologia e medicina resulta em soluções mais completas e eficazes. Ao unir especialistas nessas áreas, é possível desenvolver abordagens inovadoras que superam as limitações dos métodos tradicionais e abrem novas possibilidades para a melhoria da saúde e do cuidado com os pacientes.

Os estudos mencionados evidenciam que através do uso de grafos de conhecimento e do processamento de linguagem natural, é possível obter informações mais precisas, prever interações medicamentosas adversas, identificar novos usos para medicamentos existentes e fornecer suporte na tomada de decisões relacionadas ao uso de medicamentos. Essas abordagens promissoras têm o potencial de melhorar a segurança dos pacientes, e contribuir para a medicina personalizada, destacando a importância da constante inovação na saúde.

Assim, os resultados demonstraram que a análise, síntese, e, seleção e extração da informação estão fortemente apoiadas nas recentes pesquisas do campo do processamento de linguagem natural, fundamentada em algoritmos de aprendizado de máquina adequados para processar uma grande quantidade de dados, detectar interações e extrair informações significativas. Oferecendo uma infinidade de meios e novas oportunidades para processamento de textos clínicos de registros médicos não estruturados, promovendo a análise e identificação de conceitos de interesse nestes textos livres, e extraindo conhecimentos importantes.

Conclui-se também, que a Ciência da Informação pode atender às demandas informacionais de aquisição, preservação, identificação, seleção, organização, análise e comunicação. E, por meio de seleção e sínteses de informações contidas em prontuários eletrônicos, fazer com que seja possível que dados importantes sejam associados, medidos e comparados por meio do uso de modelos de processamento de linguagem natural. Aprimorando a representação do conhecimento médico em um nível de entidades, palavras, sentenças e de contexto. Portanto, ainda, existem infindas contribuições necessárias a serem feitas por pesquisadores do campo da Ciência da Informação, para expandir a pesquisa e promover a melhoria da representação da informação e do conhecimento nos prontuários eletrônicos do paciente, e desta forma possibilitar um melhor atendimento e cuidados de saúde.

4. REPRESENTAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO

Nesta seção é apresentado o conteúdo teórico relacionado à recuperação e representação da informação e do conhecimento com o intuito de traçar um panorama histórico-evolutivo das bases formais de representação do conhecimento e, desta forma posicionar e evidenciar a importância atual do modelo de estruturação formal de dados e informação, denominado: Grafos de Conhecimento. Para isso, destaca-se nesta, a adoção e o aprimoramento dos modelos formais de representação usados ao longo das últimas décadas para representar o conhecimento e estabelecer a convergência ótima humano-computador.

É de senso comum que a natureza do entendimento e do conhecimento humano, efetivamente está fundamentada na capacidade de raciocinar, o que, invariavelmente, leva a uma compreensão formal do mundo real e de todas as coisas que o compõem.

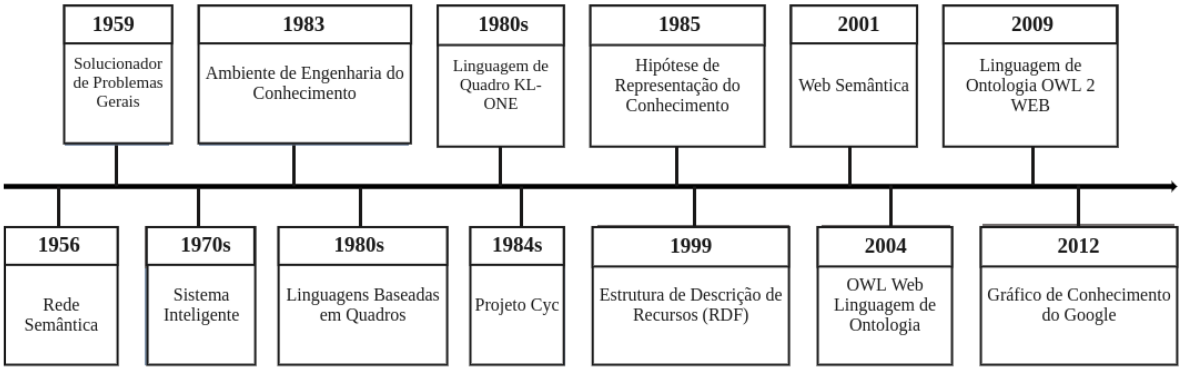
Nesse sentido, de acordo com Jakus (2013), o conhecimento pode ser definido como: "[...] todo o conjunto de dados e informações que os humanos trazem para uso prático, visando realizar tarefas ou gerar novas informações." Já o termo "representação do conhecimento", ainda segundo o mesmo autor, está relacionado com o uso símbolos formais criados, exclusivamente, para representar uma coleção de proposições, tendo como objetivo primevo o de representar conceitualmente uma determinada realidade.

Na perspectiva da ciência da informação, a representação do conhecimento conforme Alvarenga (2003) se relaciona aos estudos sobre: "codificação e uso racional da informação", e implica em processos fundamentais, tais como: "[...] análise de assunto, interpretação, classificação, descrição, e recuperação de documentos compostos de textos, imagens ou sons."

Há historicamente, especialmente com o advento dos sistemas digitais uma busca constante no sentido de otimizar a representação do conhecimento visando a melhor compreensão de um dado domínio ou assunto, e essa busca ou esforço levou ao uso dos princípios do raciocínio lógico para formalizar a representação do conhecimento.

Nesse sentido, criou-se diversos modelos esquemáticos baseados em regras formais, e dentre os diversos formalismos criados visando a otimização da representação de conhecimento destacam-se as ontologias de domínios e os grafos de conhecimento. Isso porque tanto as ontologias, quanto os grafos de conhecimento, são atualmente, parte fundamental na construção de sistemas semânticos de indexação da informação baseados em modelos de aprendizado profundo voltados para processamento de linguagem natural.

Figura 1: Evolução das bases de representação do conhecimento.



Fonte: Adaptado de JI *et al*. (2021).

Como indica o gráfico de linha do tempo da figura 1, há décadas tem se aprimorado os modelos formais de representação e visualização do conhecimento. Buscando principalmente (como, já pontuado) otimizar a convergência humano-computador.

Nesse contexto, conceitualmente, destaca-se a seguir os modelos de formalização em: Redes Semânticas, Frames, RDFs, OWL e finalmente os Grafos de Conhecimento.

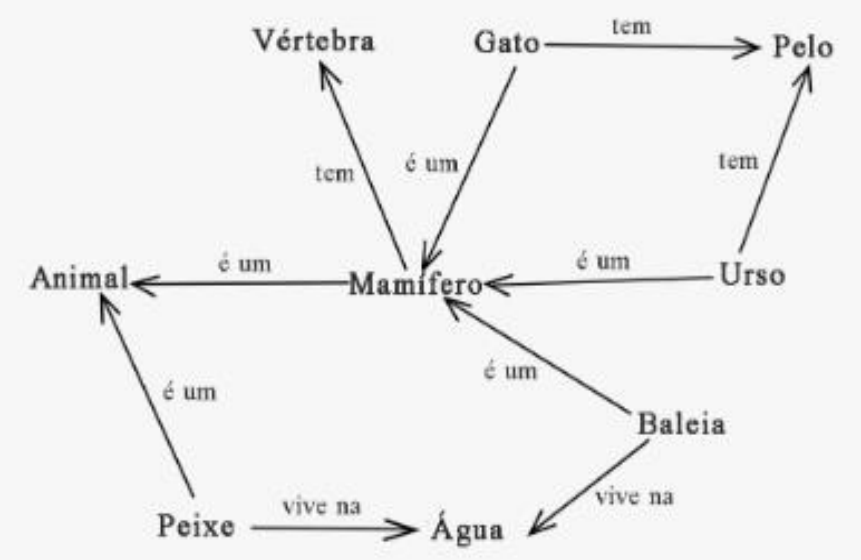
4.1 Redes Semânticas: Representando o conhecimento por meio de nós e arestas

Entende-se que qualquer objeto existente, existe em relação a outros objetos, da mesma forma, cada conceito existente, existe em relação a outros conceitos. Nesse sentido, a representação da informação e do conhecimento sempre teve como principal objetivo e desafio, o de constituir e empregar linguagens artificiais para criar estruturas visando melhor representar conceitos abstratos que se relacionam entre si, formando complexas redes de conhecimento. Com esse propósito, um desses modelos estruturais criados foram as redes semânticas.

A ideia ou uso de redes semânticas, como apontam alguns autores, remontam a centenas de anos, mas foi nas décadas de 60 e 70 do século XX que o interesse acadêmico se volta para este tipo de formalismo visando solucionar problemas relacionados à representação do conhecimento em sistemas automatizados. Nesse contexto, as primeiras implementações efetivas de redes semânticas foram empregadas para definir tipos de conceitos e padrões de relações para sistemas de tradução automática e, tornando-se posteriormente, predominante em estudos e aplicações de inteligência artificial. (LEHMANN, 1992; SOWA, 2015).

As Redes Semânticas por definição são conforme Sowa (1987), notações gráficas constituídas para representar o conhecimento por meio de padrões de nós e arestas, sendo um meio formal de representar as relações entre objetos e conceitos, constituindo-se em uma base de conhecimento que representa relações semânticas entre conceitos em uma rede.

Figura 2: Representação de uma rede semântica.



Fonte: Adaptado de Aiunplugged (2022).

Desta forma, uma rede torna-se efetivamente semântica no momento em que se atribui um significado para cada nó e aresta, e independente do domínio, esta pode representar o conhecimento de qualquer objeto ou coisa ou conjunto destes. Como no exemplo dado, uma rede semântica pode informar a uma máquina (computador) a relação entre diferentes animais, bem como, características específicas.

O emprego de frames, antologias, e redes semânticas como formas de representação do conhecimento, passaram por uma ressignificação com o aprofundamento dos estudos acadêmicos em aprendizado de máquina e inteligência artificial. E seu uso massivo em sistemas inteligentes de recuperação e visualização de informação está fortemente fundamentado nos princípios da Web semântica.

4.2 Representação do Conhecimento Baseada em Frames

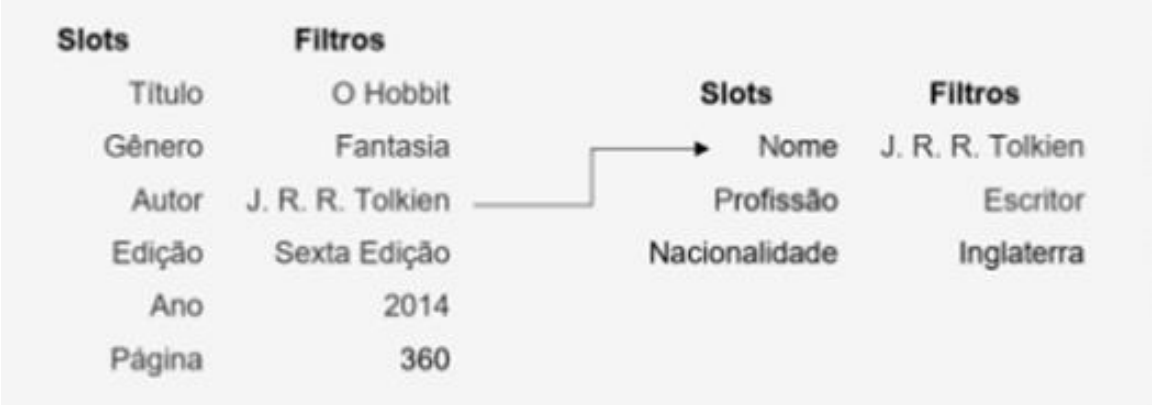
A representação do conhecimento baseada em frames ou frameworks tem sido um dispositivo importante dentro dos modelos esquemáticos de representações de conhecimento fundamentados em estrutura. Constituindo-se em um modelo de formalismo muito difundido e

empregado para representar o conhecimento baseando-se no conceito de herança de propriedades.

Essencialmente um frame é uma estrutura de dados usado para representar uma determinada situação ou conceito, em um ou vários domínios do conhecimento. Um frame conforme Rashid (2015) é um conjunto de propriedades que identificam a condição de um objeto, que está relacionado com outros frames ou objetos. Os frames têm sua origem seminal com a publicação em 1974, do artigo intitulado “A Framework for Representing Knowledge" de autoria de Marvin Minsky, e de acordo com Àvila (1992), Minsky propunha com a teoria dos frames uma forma de “construir uma base de dados, contendo quantidades enciclopédicas de conhecimento necessário à um sistema de raciocínio, de forma estruturada e flexível.”

Hoje em métodos construtores de inteligência artificial, a formalização de/ou em frames tem sido usada como um método primário de representação de conhecimento de preenchimento de slot e filtros. Onde os frames compõem-se em variados grupos interconectados, de modo a serem armazenados juntos, objetivando transmitir uma informação detalhada sobre uma determinada entidade ou conceito de um domínio específico.

Figura 3: Exemplo de frames de dados de livro e pessoa.



Fonte: Elaborado pela autora (2022).

Portanto, frames são simplesmente coleções de atributos (slots) e valores (filtros) associados, que descrevem alguma entidade do mundo real, sendo um tipo de tecnologia amplamente utilizado em várias aplicações, incluindo processamento de linguagem natural e outras aplicações de aprendizado de máquina.

4.3 Web Semântica: RDF e OWL

Ainda na esteira evolutiva das bases e dos conceitos formais de representação da informação e do conhecimento visando otimizar a iteração entre pessoas e computadores, surge e evolui os conceitos da Web Semântica. Onde se preconiza uma estrutura de representação de informação e do conhecimento que é otimizada para sistemas indexadores, para que estes possam prontamente interpretar, encontrar, compartilhar e combinar informações textuais.

Desta forma, o aumento da capacidade e velocidade de processamento possibilitado com as denominadas: Tecnologias Semânticas, contribuiu de forma decisiva com os sistemas de recuperação de grandes volumes de informações variadas de textos, isso de forma rápida e mais assertiva, tornando, assim, a recuperação de texto muito mais específica. (BERNERS-LEE, 2005) e (JAKUS, 2013). As tecnologias que constituem a Web Semântica são denominadas de "pilha da web semântica" e são constituídas por especificações, tecnologias e conceitos. Destacando-se: os modelos formais de representação: RDF e OWL.

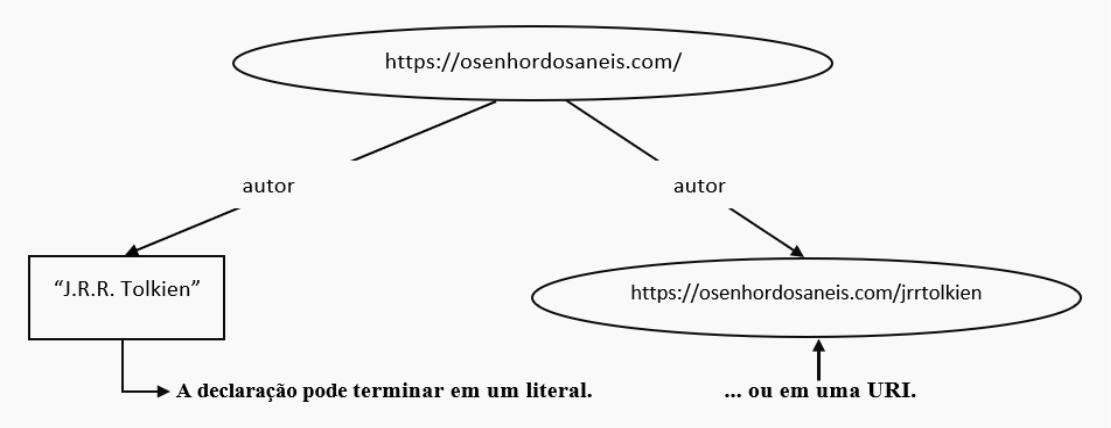
Figura 4: Tecnologias chaves da Web Semântica.

Fonte: Adaptado de Jakus *et al*. (2013).

O Resource Description Framework (RDF), conforme definição de World Wide Web Consortium (2022), é um conjunto de especificações criadas para definir o modelo padrão de intercâmbio de dados e recursos na Web, e que visa facilitar a mesclagem de dados, estendendo a estrutura de links da Web (URIs), nomeando o relacionamento entre as "coisas", onde as extremidades do link, geralmente denominado de "triplo", permite que dados estruturados e semiestruturados sejam: "mesclados, expostos e compartilhados em diferentes aplicativos."

A especificação atual do RDF, conforme W3C (2022) consiste em um conjunto de 161 especificações. As quais são utilizadas em todo o contexto da web e em variados sistemas semânticos para melhor descrever dados e seus relacionamentos, possibilitando desta forma, o compartilhamento de informação de domínios variados, de uma forma mais interpretável para humanos e máquinas. Entre outras, as recomendações RDF possibilitam que mecanismos e sistemas de pesquisa sejam otimizados por meio de metadados. Nesse contexto, os sistemas semânticos possibilitam que as buscas em indexadores disponibilizem um maior controle para o usuário final, gerando resultados mais assertivos. Por fim, o RDF não possui um vocabulário restritivo, permitindo assim, conforme Lima (2005) a definição de um vocabulário (RDF Schema), para ser usado em declarações. Isso, possibilita a interoperabilidade de conjuntos de informação e conhecimento, pois suporta uma gama heterogênea de ontologias de domínio.

Figura 5: Declaração de um objeto RDF simples.



Fonte: Adaptado de Wood *et al*. (2013).

Como pontuado o RDF relaciona ou vincula os links com os denominados triplos semânticos que compreendem o: (sujeito, predicado, objeto), como ilustrado no esquema da figura 5, onde é identificado declaração (https//osenhordosaneis.com/) que é denominada de (sujeito), e, onde o autor é denominado como (predicado), e, cujo valor da propriedade ou predicado é chamada de (objeto), nesse exemplo: https//osenhordosaneis.com/jrrtolkien.

Assim:

1. O sujeito é a URL https//osenhordosaneis.com/

2. O predicado é a palavra autor

3. O objeto é J.R.R Tolkien ou https//osenhordosaneis.com/jrrtolkien

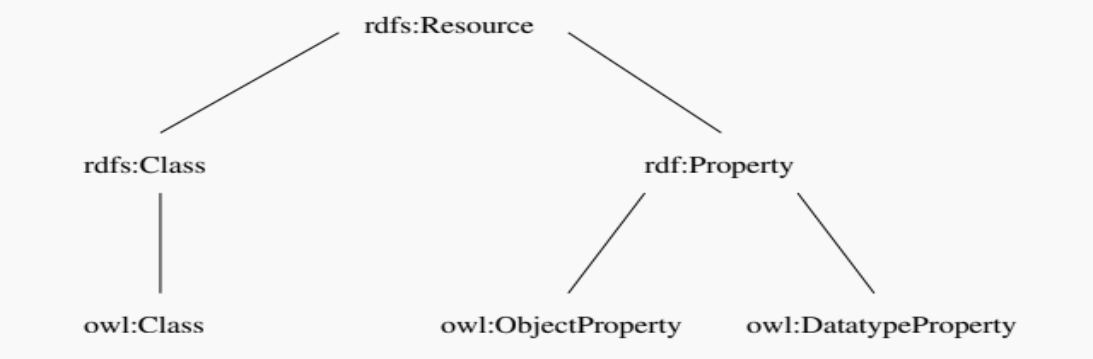
Portanto, o esquema RDF é um vocabulário de linguagem aplicado na Web Semântica visando descrever as propriedades de dados usadas no modelo de recomendações RDF.

Outro elemento importante dentro da "pilha de construção" da Web Semântica. É a OWL (Web Ontology Language), que dentro de sistemas semânticos é utilizada como linguagem padrão para definir ontologias web, adicionando, identificando e descrevendo relacionamentos entre dados. Logo, a linguagem OWL é usada para modelar os dados de forma mais expressiva e flexível, e, para possibilitar um raciocínio lógico automatizado e eficiente.

Conforme definição da W3C (2022), a OWL é uma linguagem computacional de ontologia, baseada em lógica, que é um dos principais padrões da web semântica, sendo um vocabulário que define os conceitos e relacionamentos, denominados de “termos” que são usados para descrever e representar uma determinada área de domínio, sendo projetada conforme a W3C (2022 “[...] para representar conhecimento complexo sobre coisas, grupos de coisas e relações entre coisas.”

E, enquanto linguagem de ontologia computacional, o OWL tem a prerrogativa principal de definir uma terminologia que, efetivamente, pode ser usada em documentos em formato RDF, ou seja, documentos contendo classes e propriedades de um dado objeto.

Figura 6: Relações de subclasse entre OWL e RDF/RDFS.



Fonte: Adaptado de Antoniou e Harmelen (2012).

O OWL, permite especificar taxonomias para classes e propriedades. E conforme ilustrado na Figura 6, a raiz de um documento OWL é o elemento rdf:RDF. Logo, um documento OWL é, efetivamente, um documento RDF, fornecendo compatibilidade entre ambos os padrões de linguagem.

Figura 7: Um documento de ontologia OWL, simples.

```
<!DOCTYPE rdf:RDF [
   <!ENTITY owl "http://www.w3.org/2002/07/owl#">]>
<rdf:RDF xmlns:owl ="http://www.w3.org/2002/07/owl#"
      xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<owl:Ontology rdf:about="">
   <rdfs:label>My Ontology</rdfs:label>
   <rdfs:comment>An example ontology</rdfs:comment>
</owl:Ontology>
<owl:Class rdf:ID="Person" />
<owl:Class rdf:ID="Man">
   <rdfs:subClassOf rdf:resource="#Person" />
</owl:Class>
<owl:ObjectProperty rdf:ID="hasChild" />
<owl:ObjectProperty rdf:ID="hasDaughter">
   <rdfs:subPropertyOf rdf:resource="#hasChild" />
</owl:ObjectProperty>
<owl:DatatypeProperty rdf:ID="age" />
<owl:ObjectProperty rdf:ID="isParentOf">
   <owl:inverseOf rdf:resource="#isChildOf" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="isTallerThan">
   <rdf:type rdf:resource="&owl;TransitiveProperty" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="isFriendOf">
   <rdf:type rdf:resource="&owl;SymmetricProperty" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasSSN">
   <rdf:type rdf:resource="&owl;FunctionalProperty" />
   <rdf:type rdf:resource="&owl;InverseFunctionalProperty" />
</owl:ObjectProperty>
</rdf:RDF>
```

Fonte: Adaptado de Heflin (2022).

A figura 7 é um exemplo de sintaxe RDF/OWL/XML Básica. O primeiro bloco de elemento especifica que a URI base para o conteúdo dentro do elemento RDF. O segundo, que é o elemento owl:Ontology específica a ontologia. Na sequência esta sintaxe declara uma classe (owl:Class) Person (Pessoa) e uma classe Man (Homens) e as identificam com o atributo rdf:ID. Assim, a classe Homens é uma subclasse da classe de Pessoa.

Isso significa que qualquer elemento do tipo Homens, obrigatoriamente, também é do tipo Pessoa. Desta Forma, a sintaxe descreve a propriedade rdfs:subClassOf objetivando especificar as relações entre classes. Na sequência os construtores OWL, owl:DatatypeProperty e owl:ObjectProperty traz algumas especializações das classes, como dados da propriedade e especifica os relacionamentos das subclasses. Como em DatatypeProperty: age (Idade), E em ObjectProperty: isParentOf (é_pai_de), isChildOF (é_filho_de) (HEFLIN, 2022). Conforme W3C (2022) o OWL tem três sub-linguagens cada vez mais expressivas: OWL Lite, OWL DL e OWL Full:

- OWL Lite, “menos expressivo, usado para hierarquia de classes simples e restrições simples e úteis para o caminho de migração rápida para tesauros e outras taxonomias”;

- OWL DL, “mais expressivo, mantém a completude computacional, isto é, todas as conclusões são garantidamente computáveis e é baseado na lógica de descrição”;

- OWL Full, “ainda mais expressivo, liberdade sintática de RDF, e permite que uma ontologia aumente o significado da predefinição”.

Ainda de acordo com W3C (2022), na versão OWL 2, conceitualmente, a diferença entre OWL 2 DL (Semântica Direta) e OWL 2 Full (Semântica Baseada em RDF), e que a OWL 2 DL é uma versão sintaticamente restrita do OWL 2 Full, onde as restrições são projetadas para facilitar a implementação. E o OWL 2 Full é uma extensão mais direta do RDFS.

Ainda na esteira dessa evolução gradual dos modelos de representação e formalização da informação e do conhecimento surgem ou, são aprimorados para o uso em representação do conhecimento em sistemas Web, os denominados Grafos de Conhecimento que na verdade, essencialmente, são o melhoramento ou os novos conceitos e aplicações das redes semânticas.

4.4 Grafos de Conhecimento: Novos Conceitos e Representação

Os grafos de conhecimento ou redes semânticas, como visto, basicamente são esquemas e “desenhos” rotulados que tem a prerrogativa de identificar entidades e seus relacionamentos, constituindo-se em um suporte extremamente intuitivo e de fácil compreensão para representar qualquer domínio de conhecimento e, como apontam alguns pesquisadores, historicamente eles foram introduzidos pela primeira vez na década de 1970, a princípio apenas como um novo conceito de base de dados e de conhecimento relacionado.

Entretanto, foi a criação no ano de 2012 do Knowledge Graph (Grafo de Conhecimento) do Google, que conforme Singhal (2012), aperfeiçoou o sistema de busca para apresentar resumos de informações relevantes sobre uma determinada consulta de pesquisa, bem como, lista de tópicos relacionados, que fez ressurgir o interesse do uso de grafos para representar o conhecimento nos mais variados domínios. Portanto, a necessidade crescente de representar o conhecimento em um formato compreensível tanto para humanos como para máquinas, aprofundou as pesquisas tanto no âmbito acadêmico, quanto no comercial voltado para a recuperação e representação do conhecimento em grande escala.

Uma possível definição determina que os grafos de conhecimento, são estruturas abstratas de dados que possibilitam a representação de diferentes elementos, bem como, de seus relacionamentos na forma de nós e arestas, sendo que, estes devem conter aspectos significativos do conhecimento, anotados semanticamente em variadas fontes de dados.

Para Barrasa *et al* (2021), Grafos de Conhecimento essencialmente são conjuntos interligados de fatos que descrevem entidades, eventos ou coisas do mundo real e suas inter-relações, em um formato inelegível para humanos e máquinas. Já de acordo com Fensel *et al* (2020), os denominados Grafos de Conhecimento são grandes redes semânticas que integram diversas e heterogêneas fontes de informação visando a representação do conhecimento sobre certos domínios do discurso humano. Finalmente, para Liu e Han (2018), estes são conjuntos heterogêneos de dados significativos, que organizam o conhecimento humano sobre variados domínios de forma estruturada, onde o conhecimento é representado como entidades concretas e os conceitos abstratos se relacionam entre si. Por fim, uma definição sucinta diz que grafos de conhecimento são simplesmente, conjuntos de nós e arestas.

Portanto, os grafos de conhecimento possibilitam uma maneira de construir bases de dados mais interpretável para humanos e máquinas, fornecendo uma forma eficiente de desbloquear, armazenar e recuperar conhecimento heterogêneo em uma escala industrial. E, como apresentado nos resumos do corpus documental desta presente pesquisa, estes têm sido usados para formalizar o conhecimento de domínio clínico e de saúde de modo geral. A utilização do Grafo de Conhecimento médico tem despertado interesse da área acadêmica e de saúde por suas aplicações escaláveis e inteligentes. E, de acordo com LI *et al* (2020) é possível construir grafos de conhecimento médicos eficientes a partir de prontuários eletrônicos utilizando procedimentos sistematizados e diversos tipos de entidade, como doenças e sintomas. E, com a incorporação da representação semântica de entidades, é possível obter um bom desempenho em buscas de relações e agrupamentos de doenças.

Na Biomedicina conforme Nicholson e Greene (2020), os grafos de conhecimento têm a capacidade de auxiliar os pesquisadores na solução de diversos problemas biomédicos, tais como a busca por novas opções de tratamento utilizando medicamentos já existentes, colaborando com os esforços de diagnóstico de pacientes e identificação de associações entre doenças e biomoléculas. Logo, existem inúmeros exemplos nos quais grafos de conhecimento foram empregados para descobrir novas propriedades de drogas, prever interações entre drogas, identificar alvos moleculares com os quais uma medicação pode interagir, bem como, para identificar novas doenças e tratamentos com medicamentos já existentes, além da identificação de novos alvos e reutilização de medicamentos já disponíveis (KURBATOVA e SWIERS,

2021; GELETA, 2021). Portanto, a construção de uma base de conhecimento em grafo pode viabilizar uma solução prática que permita a integração de dados farmacêuticos e médicos de forma mais simples.

A figura 8 representa um grafo de conhecimento em sua forma mais básica, onde dois nós diferentes e uma aresta se relacionam para formar o denominado triplo ou triplo semântico ou ainda, triplo RDF. Os triplos semânticos essencialmente são entidades de dados atômicos constituídos em sua forma básica por um conjunto de três entidades que sistematiza uma dada declaração semântica na forma de expressões do tipo sujeito-predicado-objeto, ou seja, ele consiste em três entidades contidas em uma expressão. Onde, cada nó pode ser um sujeito e um objeto simultaneamente se utilizado em vários triplos, e o predicado representa a relação entre dois nós, ou um nó e um literal, sendo um o sujeito e o outro objeto.

Nesta representação o nó A e o nó B são duas entidades, onde o A é o sujeito, e o nó B é o objeto. Sendo que, a seta direcional ou aresta, representa o predicado que define a relação entre ambos.

Figura 8: Representação básica de um grafo de conhecimento.

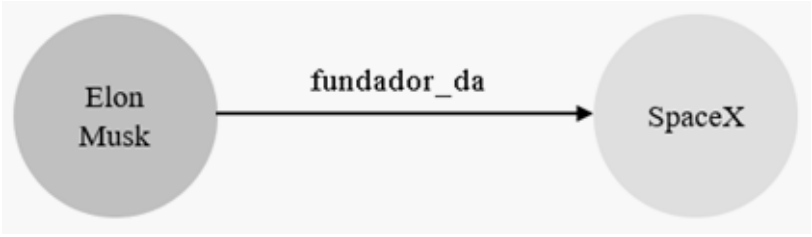
Fonte: Elaborado pela autora (2022).

Por exemplo, na expressão: “Elon Musk fundador da SpaceX” a extração do triplo pode ser dividida da seguinte forma:

- **Sujeito:** Elon Musk
- **Predicado**: Fundador da
- **Objeto**: SpaceX

Portanto, este triplo constitui um grafo de conhecimento básico, que incorpora a informação ou conhecimento derivado de uma expressão, em uma estrutura tripartida, que por sua vez, relaciona o empreendedor Elon Musk com a fundação da SpaceX.

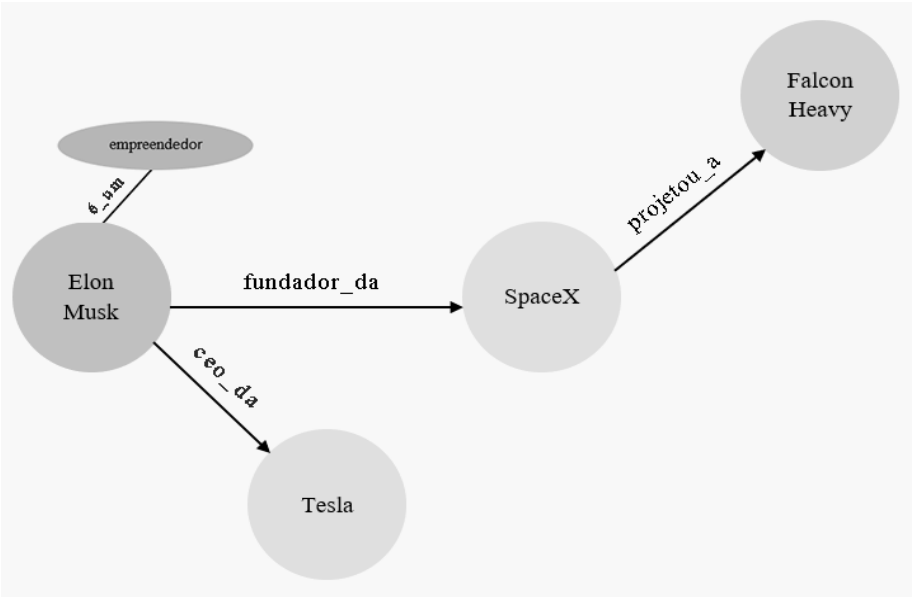
Figura 9: Representação básica de um triplo semântico.



Fonte: Elaborado pela autora (2022).

Obviamente, dentro de um grafo de conhecimento uma entidade pode ter infindas relações, o que permite representar o conhecimento de um domínio ou de múltiplos domínios em um formato hierarquicamente estruturado, permitindo responder questões cada vez mais abrangentes e desta forma viabilizar o acesso a uma gama enorme e complexa de conhecimento. Assim, novos relacionamentos entre nós ou entidades podem surgir e se expandirem não apenas da primeira entidade, mas de qualquer entidade contida em um grafo de conhecimento.

Figura 10: Representação da expansão de um grafo.



Fonte: Elaborado pela autora (2022).

Desta forma, o conhecimento representado nas entidades de interesse como nós no grafo, e com relacionamentos expressos por meio de arestas rotuladas e direcionadas, forma conjuntos interligados de fatos que descrevem entidades, eventos ou coisas e suas inter-relações, modelando um determinado domínio como uma ontologia, com tipos ou classes e subclasses hierárquicas, que descrevem um conceito.

Portanto, ao que se refere a recuperação e visualização da informação de dados não-estruturados, a construção de modelos de visualizações baseados em grafos de conhecimento tem despontado como uma alternativa viável e eficaz. Isso porque, estes são modelos baseados em conhecimento, e usados na vinculação entidades como pessoas, coisas, estados, entre outros. Bem como, na constituição de modelos estruturados em grafos, que objetive a extração e visualização das informações de dados complexos de domínio, exatamente como o são as anotações de receituário médico em prontuários eletrônicos.

Nesse sentido, as técnicas de modelagem de tópicos, classificação de texto e, de reconhecimento e extração de entidades nomeadas, constituem-se como ferramentas essenciais na construção de modelos de grafos de conhecimento e, por consequência, o desenvolvimento de modelos de recuperação e visualização da informação textual. Sobretudo quando estes estão baseadas em métodos de extração específicos de PLN (Processamento de Linguagem Natural), posto que, estes oferecem um meio computacional para sintetizar estes textos.

De um alto nível, o processamento de linguagem natural, nesse contexto, basicamente implica em alimentar um algoritmo com certa quantidade de anotações das quais ele “aprende” um conjunto de regras e padrões para identificar o que é significativo. Essas regras assumem a forma de funções probabilísticas e estatísticas (LIN *et al*. 2013).

Portanto, o PLN está na vanguarda dos métodos de extração do conhecimento expresso em linguagem textual, e, pode ser usado para a construção e atualização de um grafo de conhecimento. Que, por sua vez, pode ser integrado a outros grafos, e disponibilizado em linguagem natural, para a consulta e a geração de indexadores de consultas. Assim, as duas tecnologias (Grafos de Conhecimento e PLN) podem mutuamente se complementarem, se apoiarem e se aprimorarem para criação de visualizações de domínios específicos.

5 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural ou PLN é uma área da linguística computacional que essencialmente, se concentra na resolução de problemas relacionados à interpretação e geração automática da linguagem humana, em aplicações como: tradução automática, sumarização de textos, categorização textual, recuperação e extração da informação de textos livres. O PLN conforme Patel *et al* (2021) "[...] é um subcampo da linguística, ciência da computação, engenharia de informação e inteligência artificial preocupados com as interações entre computadores e linguagens humanas naturais (dados textuais, de fala e áudio)." Os estudos de PLN se concentram na busca de modos de programar computadores para que estes processem e analisem grandes quantidades de dados textuais. Objetivando, o desenvolvimento de modelos ou sistemas de geração da linguagem natural, bem como, de reconhecimento de fala, análise de sentimento e compreensão da linguagem oral.

Nos últimos anos com os avanços computacionais e, sobretudo, com aumento exponencial do poder de processamento computacional e, por conseguinte, o avanço do campo de aprendizado de máquina e inteligência artificial as técnicas de PLN deram um salto de qualidade e importância dentro da ciência da informação e da computação, na busca para resolver problemas linguísticos-computacionais nos mais variados domínios do conhecimento.

Compondo-se em um conjunto abrangente de técnicas de sumarização, tradução, geração e extração de dados de textos livres o processamento de linguagem natural, vem sendo massivamente aplicado para resolução de problemas relacionados à análise de documentos textuais, em variadas vertentes comerciais e industriais. E, há décadas tem sido amplamente estudado, desenvolvido e aplicado no âmbito acadêmico.

Atualmente, o PLN tem se mostrado muito eficaz na resolução e reconhecimento de entidades médicas (como: sintomas, doenças, medicamentos, tratamento, entre outros), bem como, na descoberta do conhecimento médico por meio da extração de padrões, entidades e relacionamentos de dados contidos em textos não-estruturados.

Embora o evidente progresso e importância atual das pesquisas e as múltiplas aplicações das técnicas e modelos de PLN, em uma visão retrospectiva-evolucionária o PLN teve suas origens ou bases, conforme vários autores, a partir do ano de 1950. Quando o matemático, cientista computacional e criptoanalista Alan Turing publicou um distinto artigo intitulado "Computing Machinery and Intelligence", onde propôs a formulação de um problema descrito como um jogo, denominado "jogo da imitação", no que agora, denomina-se como o "Teste de Turing". Onde, de acordo com Gunkel (2017) apud (Turing, 1996), o jogo ou teste, devia ser

realizado por três pessoas, sendo: A (um homem), B (uma mulher) e C (um interrogador, de qualquer sexo). Onde o objetivo do interrogador é apontar, com base em perguntas e respostas simplificadas, qual é o homem e, qual é a mulher. Ou seja, basicamente, o teste de Turing é proposto para ser aplicado com o objetivo de se aferir (por meio de um árbitro ou analista humano, independente), qual é a capacidade de um determinado algoritmo ou sistema de computador, de se passar por, ou imitar um ser humano em uma conversa real.

Como pontuado, as origens do PLN também têm ligações diretas com o campo da Linguística, especificamente, a partir de uma abordagem estruturalista da linguagem, onde especialmente vê-se a linguagem menos como um sistema hermético e caótico, e mais como um sistema interligado onde os elementos correlacionam-se, levando invariavelmente, à identificação de contextos, por meio de um conjunto de causas.

Historicamente, as primeiras aplicações de processamento de linguagem natural remontam ao ano de 1954, onde a IBM em parceria com a universidade americana Georgetown construíram um sistema de tradução automática, capaz de traduzir um número limitado de frases do idioma russo para o inglês. Houve, então, um considerável período de "obscurantismo" nesta área, até que na década de 1980, a PLN, através da IBM, se destacou novamente com a construção de sistemas de tradução automática baseada em métodos estatísticos. Onde, a tradução automática fundamentada em estatística, por meio de aprendizado de dados ou com dados, promoveu uma diminuição da necessidade de intervenção humana, para este fim. (LI *et al*, 2018; PATEL *et al*, 2021).

Com a gradual evolução das pesquisas acadêmicas com PLN, nas décadas de 1990 e 2000, a indústria começa a pesquisar e desenvolver aplicações, levando ao surgimento de sistemas comerciais de reconhecimento de voz e tradução automática. No início da década de 2010, pesquisadores de PNL, tanto no meio acadêmico, quanto na indústria, fazem os primeiros usos de redes neurais, para tarefas de PLN. Em 2015 o Google emprega métodos de redes neurais para resolução de tarefas de áudio, assim, os métodos de aprendizado profundo, promovem um salto na solução de tarefas de PLN. (LI *et al*, 2018; PATEL *et al*, 2021).

Em 2016, usando métodos de aprendizado profundo, o Google liberou uma versão aprimorada do Google Tradutor. E, em 2017/2018 surgiram modelos de PLN pré-treinados, usando uma nova arquitetura denominada Transformer, onde destaca-se o modelo BERT, desenvolvido pela equipe do Google. Por fim, em 2019 e 2020, surgem os modelos GPT (Generative Pre-Training Transformer), GPT-2 e GPT-3 que são modelos de processamento de linguagem natural autorregressivo, que executam uma ampla variedade de tarefas de PLN

(como o ChatGPT), por meio de aprendizagem profunda. Os modelos de arquitetura GPT se constituem no estado da arte do PLN. (LI *et al*, 2018; PATEL *et al*, 2021).

Logo, o PLN em seu estado da arte é baseado em arquitetura de aprendizado profundo, especialmente na arquitetura Transformer. O aprendizado profundo, basicamente implica em alimentar um algoritmo com grandes quantidades de dados referentes a um objeto informacional (por exemplo, corpus de anotações clínicas não-estruturadas), com o objetivo de fazer com o que o algoritmo aprenda um conjunto de regras ou padrões. E, deste modo, possa identificar informações que de fato possuam valor, possibilitando assim, a partir da recuperação das informações obtidas, a criação de uma maior clareza em relação ao conhecimento, e dessa forma, permitindo melhores tomadas de decisão (SHEN *et al*, 2021).

Essencialmente, o Transformer é um modelo de aprendizado profundo, que de acordo com Vaswani *et al.* (2017) baseia-se em um mecanismo denominado "self-attention" (ou “auto-atenção”), e, cujo a nova proposta de arquitetura de rede neural, pondera diferencialmente o significado de cada parte dos dados de entrada, e dispensa inteiramente a recorrência e as convoluções neurais, até então, marca característica de modelos pré-treinados de PLN.

Pontuando que as denominadas Redes Neurais Artificiais (do inglês Artificial Neural Networks ou ANNs), em síntese são um paradigma computacional que propõe mimetizar ou simular o funcionamento do cérebro humano, com o objetivo de possibilitar que um sistema de computador aprenda raciocínios ou conhecimentos humanos, a partir do processamento e análise de grandes volumes de dados observacionais.

Esse paradigma tem sido aplicado em larga escala, sendo o componente chave de aplicações de Inteligência Artificial em tarefas de reconhecimento e classificação de imagens, análise de texto e fala e, inúmeros outros problemas de PLN e computacionais nos mais variados domínios do conhecimento. Existem atualmente diferentes tipos de redes neurais, aplicáveis, que basicamente se diferem em relação a sua arquitetura e aplicações, entre elas, as Redes Neurais Convolucionais ou Convolutional Neural Networks (CNNs) e as Redes Neurais Recorrentes ou Recurrent Neural Networks ou (RNNs).

O Transformer surgiu no final de 2017, com a publicação de um artigo seminal denominado “Attention Is All You Need”, onde Vaswani e outros pesquisadores e especialista membros do Google Brain e Google Research, apresentam uma nova arquitetura de rede neural que superou em velocidade de treinamento e precisão os modelos de PNL de última geração existentes. Desde então, os Transformers se tornaram um componente determinante dos modelos de processamento de linguagem natural, baseados em aprendizado profundo.

Observa-se assim, que os métodos de Processamento de Linguagem Natural, especialmente os baseados em arquitetura de transformadores, estão na vanguarda dos modelos de extração da informação e do conhecimento expresso em linguagem natural, sejam eles de origem puramente textual, ou mesmo de áudio ou fala. Podendo ser empregados, entre outras, para a construção e/ou atualização de grafos de conhecimento, que, por sua vez, podem ser disponibilizados de forma agregada para a geração de resultados em sistemas de consultas.

É de consenso geral que é extremamente complexo o desenvolvimento de modelos de sistemas computacionais inteligentes capazes de realizar simultaneamente várias tarefas de processamento automático. Deste modo, é necessário categorizar problemas específicos e os dividir e, assim construir modelos específicos para lidar com problemas de modo separado. Desta forma, no âmbito das tarefas de PLN é muito eficaz a criação de modelos de domínio específicos para o reconhecimento e extração de entidades nomeadas e seus relacionamentos.

5.1 Reconhecimento e Extração de Entidades Nomeadas e Relacionamentos

Como pontuado anteriormente, a área de abrangência da linguística computacional e, por conseguinte, as tarefas de processamento de linguagem natural são muito abrangentes e, desta forma, contemplam inúmeras aplicações que buscam a resolução de variados problemas relacionados ao tratamento de textos e áudios. Dentre estas tarefas e subtarefas destacam-se as relacionadas à tradução de textos, a sumarização de textos, a análise de sentimentos, geração de textos e a de reconhecimento de entidades nomeadas e seus relacionamentos.

Neste contexto, destaca-se aqui a tarefa de PLN denominada de NER ou (Named Entity Recognition ou ainda, Reconhecimento de Entidade Nomeada). O Reconhecimento de Entidade Nomeada, conceitualmente, pode ser definido como: uma subtarefa de extração de informação, cujo objetivo principal é o de reconhecer e extrair uma ou várias entidades pré-determinadas em um dado documento de texto. Como afirma Campesato (2021) a recuperação das informações objetivamente requer a extração de dados e informações, e, envolve métodos específicos de indexação e classificação de documentos de base puramente textuais.

A extração de informação envolve subtarefas para identificar entidades e extrair dados em um documento de texto. Estas entidades pré-determinadas ou nomeadas podem ser: nomes pessoais, nomes de organizações, localidades, marcas ou quaisquer outras palavras especializadas como: datas, objetos, números, medidas, medicamentos, doenças, entre outras.

Desta forma, uma entidade nomeada é, essencialmente, uma menção ou apontamento de texto relacionado a um conceito específico extraído do mundo real. E, esse dado apontamento é usado para identificar e extrair informações significativas (entidades) em uma

parte específica de texto. Sendo que, uma entidade pode ser restrita a uma única palavra ou pode se constituir de grupo pré-determinado de palavras pertencentes a uma mesma categoria.

Ou seja, na frase:

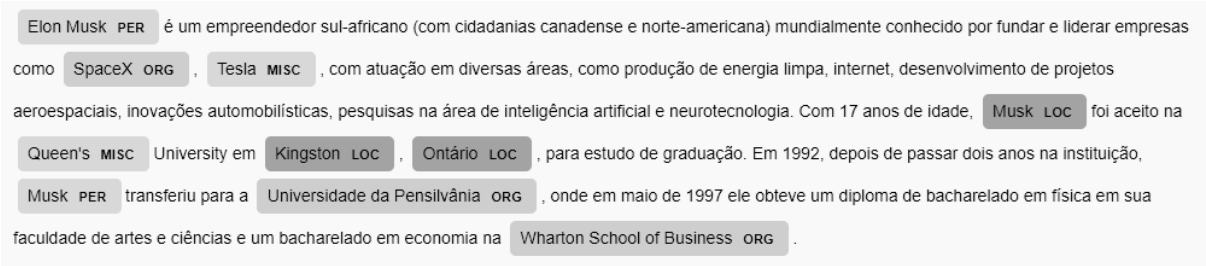
“O empreendedor Elon Musk é fundador da SpaceX”

• **SpaceX:** é uma entidade constituída por uma única palavra.

• **Elon Musk:** é uma entidade constituída por duas palavras da mesma categoria.

Invariavelmente, uma entidade nomeada sempre fará referência a um objeto em específico, de forma que, este dado objeto específico sempre será distinguível e/ou identificável, através da entidade nomeada correspondente. Ou seja, evocando novamente a frase “O empreendedor Elon Musk é fundador da SpaceX”, identifica-se as palavras Elon Musk e SpaceX como sendo entidades nomeadas, mas, contrariamente, não são identificadas com tal, as palavras empreendedor e fundador. A razão desta não identificação, é que as palavras empreendedor e fundador não apontam, ou seja, não fazem referência a um objeto específico, se tratando de nomes de atribuições de objetos generalizados.

Figura 11: Exemplo de entidades reconhecidas, destacadas em um texto.



Fonte: Elaborado pela autora (2022).

As tarefas de processamento de linguagem natural, quando se concentram em reconhecimento de entidade nomeada, como visto, é essencialmente uma tarefa que marca entidades em um dado texto. Deste modo, a aquisição de informação e de conhecimento é fundamentada em entidades extraídas como representação de referência de entidade, ou recursos de texto e de tipos de entidade e seus respectivos relacionamentos. Assim, estas tarefas são baseadas em algoritmos que buscam padrões e recursos específicos de um idioma e, são aplicados em grandes quantidades de textos. Para este fim usa-se modelos de algoritmos baseados em arquiteturas neurais denominadas de sequência a sequência (por exemplo, as

arquiteturas LSTM, CNN e BERT), para aprender recursos de nível de palavras e, para a codificação ou computação de graus de correspondências de entidades) (LI *et al*, 2021).

Para o desenvolvimento de modelos de reconhecimento de entidades nomeadas, há disponíveis, inúmeras ferramentas e modelos pré-treinados. Dentre estas ferramentas, uma amplamente utilizada é a biblioteca SpaCy, que é uma biblioteca de código aberto que integra modelos e aplicações para processamento avançado de linguagem natural, estando baseada na linguagem de programação Python.

O SpaCy é uma biblioteca muito completa e especializada que facilita a criação de aplicações reais de PLN, e disponibiliza modelos de arquitetura neural pré-treinados, dando suporte a modelos em português. Logo, esta referida biblioteca permite o treinamento de modelos de reconhecimento de entidades de um domínio específico, onde estas entidades são reconhecidas por meio de aprendizado profundo. Assim, tais entidades são reconhecidas por meio de exemplos, os quais são rotulados e integrados ao modelo pré-treinado disponibilizado pelo SpaCy, e, treinados para que o novo modelo possa reconhecer estas novas entidades.

5.2 PLN: Extração de Textos Não-Estruturados de Domínio Clínico

Com a crescente adoção dos prontuários eletrônicos do paciente, bem como, de variadas plataformas de tecnologia da informação no âmbito das instituições de saúde, inevitavelmente surgem novas demandas e questionamentos, em especial os relacionados a recuperação e usabilidade do grande e heterogêneo montante de dados e informações que são coletadas e armazenadas nessas plataformas.

E, especificamente surgem questões referentes a otimização da recuperação da informação dos dados não-estruturados provenientes de anamneses, e armazenadas de modo caótico, nos prontuários eletrônicos do paciente. Visto que, este tipo específico de dados é essencial para todo o contexto de atendimento do paciente, posto que, é nesse documento que estão as principais informações sobre o verdadeiro estado clínico do paciente. E, obviamente, a recuperação tradicional da referida informação, na busca de informações relevantes contidas nessas anotações demanda um tempo considerável, sobretudo se levados em consideração o montante de atendimentos diários, nas instituições de saúde.

Existe, portanto, uma crescente dificuldade e, por conseguinte, uma urgência no desenvolvimento de sistemas que otimizem o processamento e a recuperação desses dados. Nesse sentido, de acordo com Savova *et al* (2019) a área médica deve lançar mão ao máximo dos avanços recentes da tecnologia da informação, em especial da tecnologia voltadas ao

Processamento de Linguagem Natural "[...] para extrair a totalidade da riqueza de dados e informações armazenadas, e que se acumulam rapidamente em prontuários eletrônicos."

Conforme afirma Jackson *et al* (2016), "[...] as técnicas de PLN podem, efetivamente serem aplicadas em dados de anotações clínicas para processar grandes quantidades de texto não-estruturados e retornar informações estruturadas sobre seu real significado."

Deste modo, o Processamento de Linguagem Natural tem sido muito usado para analisar e entender a linguagem humana no domínio clínico, e existe uma gama crescente de inciativas para recuperar essa quantidade de dados textuais não estruturados de variados tipos de anotações médicas eletrônicas, que contêm informações valiosas sobre os cuidados do paciente. Assim, o PLN tem sido utilizado para extrair essas informações e transformá-las em dados estruturados que podem ser analisados para melhorar os resultados da saúde. E uma das tarefas de PLN mais usada é a tarefa de NER, que como visto, envolve a identificação e categorização de entidades mencionadas no texto (por exemplo, doenças, medicamentos e procedimentos), bem como classificação de texto clínico. O PLN tem o potencial de revolucionar a área da saúde, fornecendo insights a partir de dados clínicos não estruturados que eram anteriormente inacessíveis. E, que agora (como sugere as pesquisas científicas componentes do corpus documental deste presente trabalho) começam a ser de forma automática e inteligente processadas, recuperadas e visualizadas de modo rápido e claro.

Nesse contexto, como exemplo de uso, conforme Jagannatha *et al*. (2019), o PLN pode ser uma solução para fornecer informações precisas e rápidas sobre a detecção automatizada de eventos adversos de medicamentos. Esse processo é difícil e custoso quando realizado por meio de revisão puramente humana ou manual em prontuários eletrônicos. Visando sanar esta demanda os autores utilizaram o PLN para extrair menções de doenças e medicamentos, a fim de gerar previsões de eventos adversos de medicamentos. Eles utilizaram notas de prontuários eletrônicos que foram anotadas com entidades clínicas nomeadas, como medicamentos, dosagens, vias, durações, frequências e indicações, entre outros. Essas anotações foram usadas para treinar um modelo de reconhecimento de entidades nomeadas, que extraiu menções de doenças e medicamentos.

Outro exemplo de uso, como mostra o trabalho de Koleck *et al*. (2019) que conduziram uma revisão sistemática da literatura que investigou o uso de PLN para analisar informações sobre sintomas em textos não estruturados e propuseram recomendações para estudos futuros sobre o uso do PLN para examinar narrativas clínicas de texto livre em PEP. Em resumo, os autores afirmam que tarefas de PLN, como o NER e métodos de classificação podem ser utilizados e, já estão sendo avaliados para extrair informações de texto livre em PEP. Esses

métodos têm sido aplicados com êxito para recuperar dados de vários tipos de anotações e, uma ampla gama de sintomas em diversas especialidades clínicas.

Um último exemplo conforme a pesquisa de Kormilitzin *et al* (2021), que pontuam que os registros médicos em texto livre contêm informações ricas sobre a história dos pacientes, mas como são expressos em linguagem natural, apresentam desafios na utilização, em comparação com fontes de dados estruturados e prontos para uso. No entanto, os autores afirmam, que o PLN oferece novas oportunidades para lidar com textos médicos não estruturados permitindo que uma grande quantidade de dados seja processada e, interações sejam detectadas e informações significativas sejam extraídas. Os autores afirmam que um modelo de NER preciso e robusto é um componente essencial e fundamental para qualquer sistema de informação clínica, posto que, ele é capaz de identificar rapidamente conceitos médicos importantes, como nomes de medicamentos, frequência de administração, relatórios de sintomas, diagnósticos, entre outras.

Logo, entende-se que o emprego de técnicas de processamento de linguagem natural para o reconhecimento de dados específicos em texto não estruturados. E, sua posterior estruturação em grafos de conhecimento, a partir do reconhecimento e extração de entidades e relacionamentos, efetivamente, contribui para a geração de informações médicas mais compreensíveis e padronizadas em um nível sintático e semântico. Tornando as informações prioritárias mais acionáveis, o que tende a agilizar a compreensão do quadro geral de um determinado paciente em um menor tempo, impactando na qualidade do atendimento.

Como anteriormente pontuado, como parte da abordagem prática desta pesquisa propõe-se um modelo para a extração de entidades nomeadas e relacionamentos e, sua posterior estruturação e agregação em grafos de conhecimento. Para isso descreve-se no próximo capítulo um “passo a passo” sistematizado, partido da aquisição e tratamento de dados (sentenças de anotações de receituário médico), sua rotulagem visando a construção e treinamento de um modelo supervisionado de aprendizado de máquina, capaz de reconhecer e extrair pré-determinadas entidades e, seus respectivos relacionamentos. Com esse intuito, também são detalhados os métodos e ferramentas empregados para a conclusão desse modelo de recuperação e visualização de medicamentos descritos em prontuários eletrônicos.

6 MODELO DE RECUPERAÇÃO E VISUALIZAÇÃO DE INFORMAÇÕES DE MEDICAMENTOS

Esta seção compõe-se da descrição detalhada da vertente aplicada dessa pesquisa. Onde é delineado um modelo para a recuperação e visualização de informações de medicamentos descritos em prontuários eletrônicos. Tal modelo efetivamente é composto de procedimentos e fluxos de trabalho aplicados para recuperação e visualização em grafos de conhecimento de informações de anotações medicamentosas de receituários eletrônicos. Logo, apresenta-se nesta, procedimentos de reconhecimento e extração de entidades nomeadas e relacionamentos, bem como, métodos de construção, agregação e armazenamento de dados e informações em grafos de conhecimento, em banco de dados gráficos.

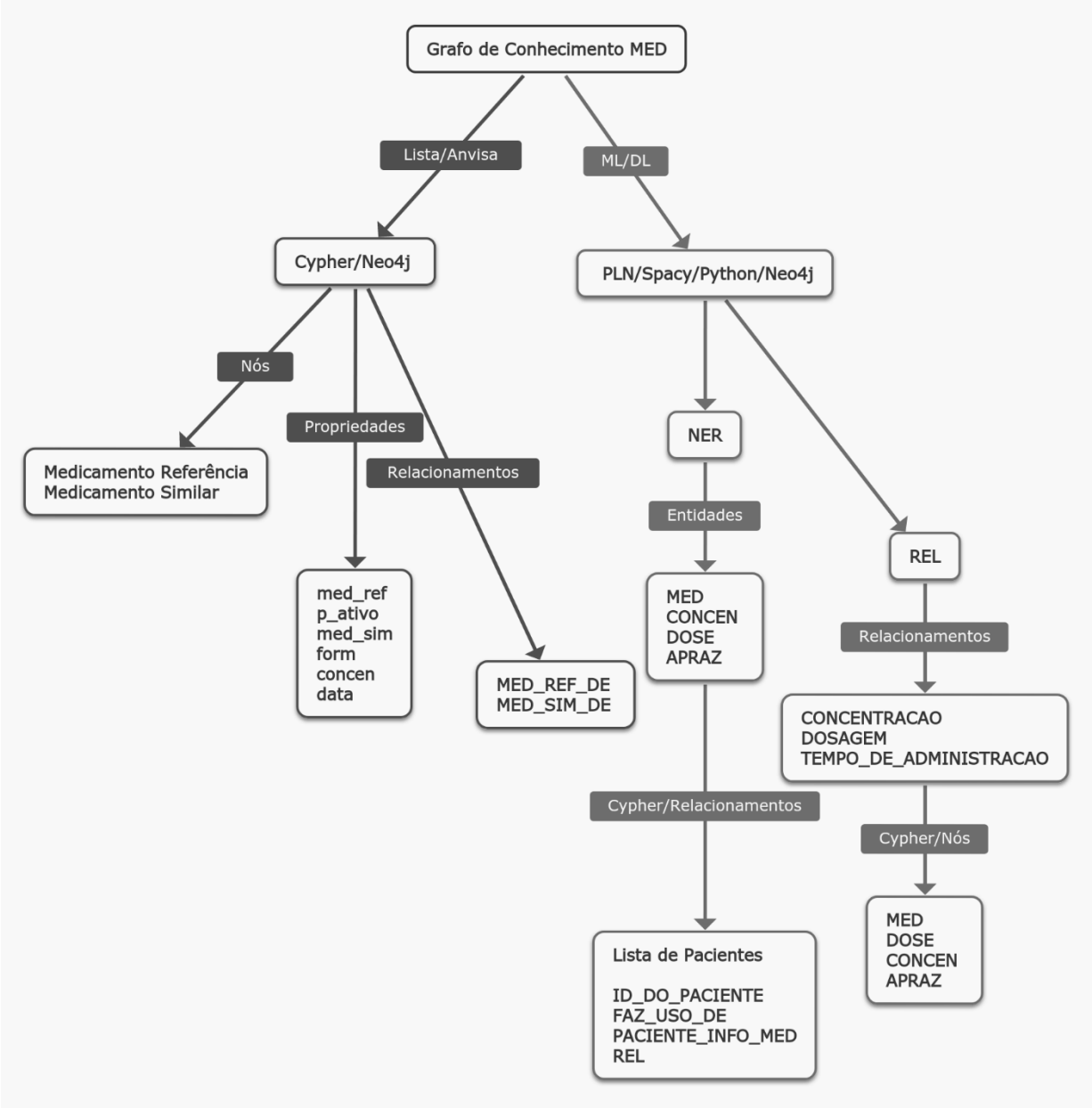
6.1 Aspectos Gerais do Modelo de Visualização de informações de Medicamentos

A vertente aplicada dessa pesquisa, como já pontuado, constitui-se basicamente no desenvolvimento, delineamento e descrição de um modelo de reconhecimento e extração de entidades nomeadas e relacionamentos. E, da construção e agregação de grafos de conhecimento para a visualização de informações de medicamentos, provenientes de dados de anotações não estruturadas de receituários advindos de prontuários eletrônicos.

Este modelo consiste em dois processos distintos que posteriormente juntam-se compondo o modelo a recuperação e de visualização de informações diversas de medicamentos. Sendo que o primeiro processo se trata do desenvolvimento de um modelo de reconhecimento e extração de entidades nomeadas e seus respectivos relacionamentos, modelo esse, baseado em PLN e aprendizado de máquina. E o segundo trata-se da construção e agregação de grafos de conhecimento de medicamentos de referência, seus similares, princípio ativo e outras informações, a partir de lista de registo de medicamentos disponibilizada pela Anvisa.

Nesse ponto, é importante reforçar que os dados usados nos referidos processos, advém respectivamente de resumos de modelos de anotações textuais de receituário provenientes de prontuários eletrônicos disponibilizados pelo do grupo de estudos HAIS, do Hospital das Clínicas da Faculdade de Medicina de Marília. E, de dados que compõem uma lista completa de medicamentos similares intercambiáveis, contendo medicamentos similares e seus respectivos medicamentos de referência, além de outras informações básicas, como: o princípio ativo de cada medicamento e seu respectivo fabricante. Esta lista foi atualizada em 11 de maio de 2020, e disponibilizada pela Agência Nacional de Vigilância Sanitária – Anvisa, em formato PDF.

Figura 12: Mapa das duas abordagens de desenvolvimento do grafo de conhecimento de medicamentos, deste trabalho.



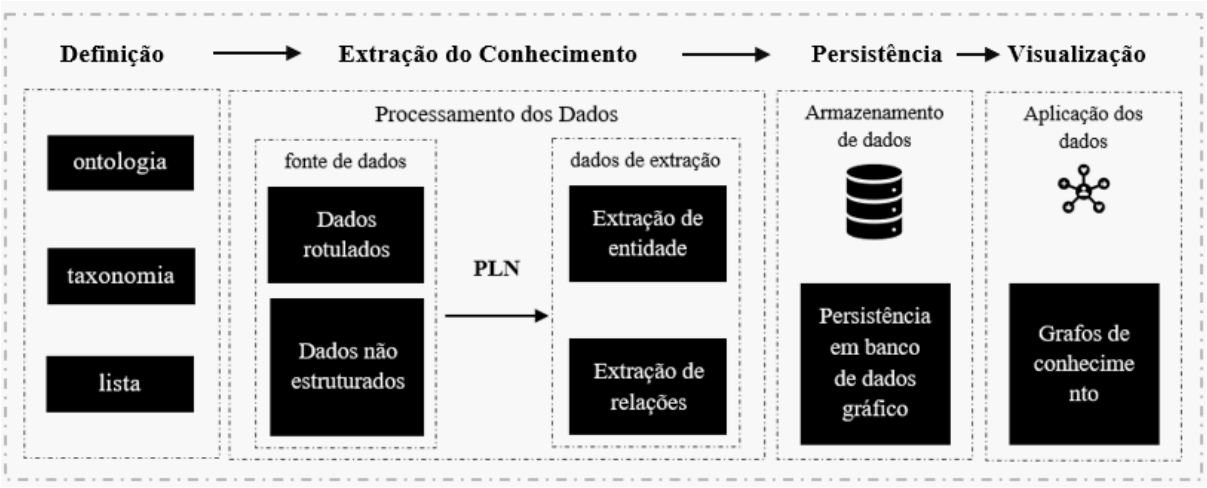
Fonte: Elaborado pela autora (2022).

Especificamente, o referido modelo consiste na descrição de métodos, processos e fluxos de trabalho para a customização de modelos de reconhecimento e extração de entidades nomeadas e da criação de grafos de conhecimento. E tem como objetivo orientar a construção de sistemas de recuperação e visualização de informação de dados de medicamentos provenientes de fontes não-estruturadas de anotações textuais de receituários eletrônicos, bem como, compor uma base de dados a ser usada para o desenvolvimento de um sistema de recomendação de medicamentos similares, em um sistema de prontuário eletrônico.

Como mostrado anteriormente na figura 14, neste trabalho foram utilizadas duas abordagens visando objetivamente a construção e agregação de grafos de conhecimentos contendo informações relacionadas a medicamentos. Onde, a primeira abordagem parte da referida lista de medicamentos de referência e similares disponibilizada pela Anvisa, contendo milhares de medicamentos. E, por meio do tratamento e transformação dos dados, e do uso de scripts da linguagem Cypher e do banco de dados Neo4j constroem-se um grafo de conhecimento com os nós, relacionamentos e propriedades de todos os medicamentos da lista.

Ainda como detalhado na figura 12, a outra abordagem usada foi baseada no desenvolvimento de modelos de aprendizado de máquina, para o reconhecimento e extração de entidades nomeadas e seus relacionamentos, denominados modelos NER e REL.

Figura 13: Fluxograma de representação do modelo de NER/REL.



Fonte: Elaborado pela autora (2022).

6.2 Resumo das Fases da Constituição dos Modelos de NER e REL

Na **primeira fase**, definiu-se as entidades de interesse a serem reconhecidas e extraídas dos referidos textos, que foi definida com base na frequência que estas aparecem nos textos e nas informações de medicamentos de interesse, assim, as entidades de interesse são 4 (medicamento, concentração, dosagem e intervalo de administração). Uma vez definida as entidades de interesse e relações foi construída uma coleção de dados textuais contendo anotações de receituário e, uma vez constituído e normalizado o dataset, fez-se a anotação ou rotulagem de NER (Named Entity Recognition ou Reconhecimento de Entidade Nomeada), bem como, de relacionamento de entidades, REL. Este processo de rotulagem pode ser feito de

modo manual ou de preferência de forma semiautomática, usando uma ferramenta de rotulagem de dados textuais de código aberto, como por exemplo: a ferramenta Label Studio.

Uma vez constituídos os conjuntos de dados rotulados de NER e de Extração de Relação, em uma **segunda fase**, foi feito o treinamento de um modelo de processamento de linguagem natural especializado para executar tarefas de extração de entidades (MED, CONCEN, DOSE E APRAZ) e de seus respectivos relacionamentos. A principal ferramenta utilizada para a criação do referido modelo de PLN foi a biblioteca SpaCy, que é uma biblioteca de software de código aberto, especializada para tarefas de processamento avançado de linguagem natural, como, tarefas de extração de informações em grande escala.

A biblioteca SpaCy é baseada na linguagem de programação Python e no PyTorch, que é uma estrutura que possibilita a prototipagem e a implementação de modelos de aprendizado de máquina, ela é de código aberto, e tem capacidade industrial sendo altamente escalável. Uma vez treinado e ajustado, este modelo de extração de entidade e relacionamento foi capaz de extrair as informações, que posteriormente foram salvas em um banco de dados gráfico denominado Neo4j, e representadas em uma estrutura constituída em grafos de conhecimento, contendo, nós com entidades e arestas com respectivos relacionamentos entre as entidades reconhecidas.

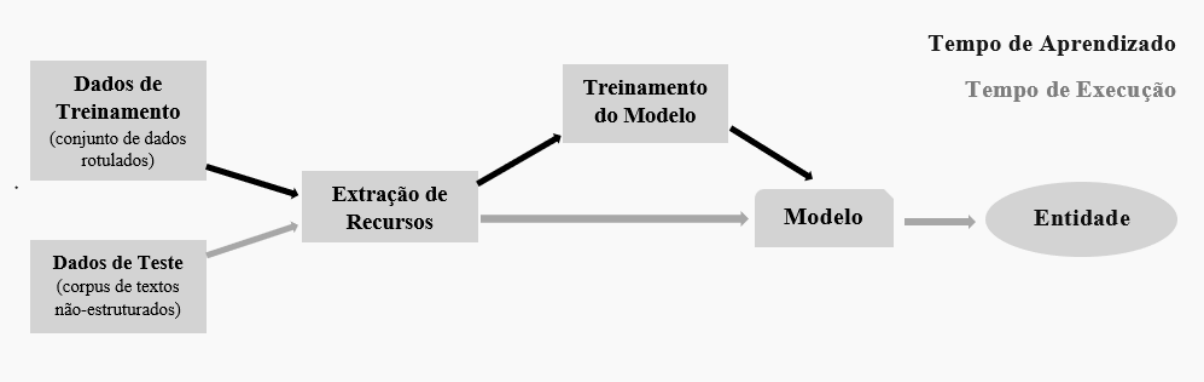
Na **terceira última fase** do desenvolvimento do modelo foi construído um grafo de conhecimento composto por medicamentos retirados da lista da Anvisa, que foi agregado ao grafo criado e armazenado na fase anterior, contendo informações medicamentosas. Bem como, para exemplo, foi proposto uma API para a recuperação das informações contidas no banco de dados gráfico, Neo4j, usando o GraphQL. Esta é uma linguagem de consulta e tempo de execução, usada em ambientes de execução dedicados a servidores, que viabiliza a construção de interfaces de aplicações do tipo API, com dados estruturados em grafos. Onde a prioridade seja recuperar e disponibilizar dados por meio de consultas flexíveis, possibilitando o acesso rápido e “amigável” das propriedades contidas em um dado recurso informacional, bem como, possibilitando que “se siga” ou se recupere de modo fácil as referências ou relações entre eles, sendo muito aplicada na recuperação de dados de grafos de conhecimento.

6.3 Fluxos de Trabalho para Reconhecimento de Entidade Nomeada

O Reconhecimento de Entidade Nomeada, essencialmente é uma tarefa que permite determinar referências em um dado texto e identificar entidades de informação como: pessoas, lugares, organizações, medicamentos ou qualquer outro objeto de referência. E atualmente para

a resolução destes problemas ou tarefas de processamento de linguagem natural, usa-se abordagem de “aprendizagem supervisionada”. Logo, técnicas de PLN baseadas em aprendizado de máquina são utilizadas para o desenvolvimento de modelos de análise e extração de textos não-estruturados, usando-se o NER.

Figura 14: Fluxos de trabalho para reconhecimento de entidade nomeada.



Fonte: Elaborado pela autora (2022).

Com base no fluxo de trabalho para reconhecimento de entidades nomeadas, o denominado fluxo de "Tempo de Aprendizado", consiste em um conjunto de dados usado para o treinamento de um "modelo" que deve ser treinado ou “aprendido dos dados”. O propósito é que o modelo de previsão se generalize a partir de um pequeno conjunto de exemplos de dados de medicamentos, e, desta forma, reconheça novos textos contendo dados de medicamentos.

No fluxo de tempo de aprendizado, os dados de treinamento consistem em dados de textos não-estruturados previamente anotados (rotulados) por humanos para as entidades nomeadas a serem aprendidas pelo treinamento do modelo. Exemplificando: no texto “O paciente fez uso de 30 gotas do medicamento Dipirona Sódica de 500Mg".

Anota-se as entidades:

- <medicamento>**Dipirona Sódica**</medicamento>
- <concentração>**500Mg**</concentração>
- <dosagem>**30 gotas**</dosagem>

A expectativa é que o modelo, uma vez treinado, aprenda com exemplos previamente rotulados, e, desta forma, reconheça e destaque as entidades nomeadas (a partir de novos textos arbitrários de entrada), como: medicamento, concentração e dosagem.

Ainda dentro do fluxo de tempo de aprendizado, a fase ou passo de “Extração de Recursos” trata-se dos recursos associados à “demarcação” de uma ou um grupo de palavras em uma sentença ou trecho de texto. Ou seja, palavras anteriores e posteriores a uma entidade, são características fortes para contextualizar entidades. Sendo um recurso útil que geralmente se usa para reconhecer entidades nomeadas que ocorrem em um dado texto. Assim, “trechos de textos” compostos de palavras e tipos de palavras são definidos usando as “tags” de parte de um discurso. Para isso pode-se definir um padrão de etiquetas IOB (Inside–Outside–Beginning) para denotar o “interior, o exterior e o início” de um pedaço de texto.

Figura 15: Exemplo de formatação IOB de marcação de tokens.

```
O   O
paciente    O
toma    O
2   B-DOSE
Cápsulas    I-DOSE
de  O
Neo B-MED
Moxilin I-MED
de  O
10  B-CONCEN
mg  I-CONCEN
de  O
4   B-APRAZ
em  I-APRAZ
4   I-APRAZ
horas   I-APRAZ
.   O
```

Fonte: Elaborado pela autora (2022).

Assim, no texto exemplo, tokenizado: “O paciente toma 2 cápsulas de Neo Moxilin de 10 Mg de 4 em 4 horas.”, como mostra a figura 16. O prefixo **I-** indica que a tag está dentro de um pedaço. Já o **O** indica que o token não pertence a nenhum pedaço. E o prefixo **B-** indica que a tag é o início de um pedaço, que segue imediatamente outro pedaço sem tags.

Finalizando o fluxo de tempo de aprendizado, o “Treinamento do Modelo" é o que algoritmos de Aprendizados de Máquina ou Aprendizado Profundo trata os dados de entrada. Ou seja, é nesse ponto que é realizado o treino do modelo de NER, “combinando os recursos”. Para este fim, existem algumas técnicas e/ou arquiteturas de ML/DL disponíveis, dentre estas destacam-se: CNN, RNN e BERT ou Transformers.

Ao que se refere ao “Tempo de Execução”, essencialmente, trata-se da análise de dados da entrada, onde um trecho de texto, ou um conjunto de textos de entrada no formato não-estruturado e não rotulados são inseridos no modelo gerando o texto de saída, que correspondente, com as entidades reconhecidas e destacadas pelo modelo que foi treinado no fluxo de trabalho do tempo de aprendizado. Como mostra a figura 16, o fluxo de tempo de execução reutiliza o módulo extração de recursos do fluxo de trabalho do tempo de aprendizado. Portanto, a tarefa de Reconhecimento de Entidade Nomeada é o primeiro e mais importante passo para a extração do conhecimento de domínio, contido no texto não-estruturado. Assim, um dado textual bruto é processado com o modelo de reconhecimento de entidade nomeada, desenvolvido com a ajuda de abordagens linguísticas e métodos matemáticos e estáticos. E o modelo de NER identifica uma dada entidade e a categoriza em uma classe mais adequada.

Nesse contexto, para facilitar e otimizar o tempo de treinamento de um modelo para resolução de tarefas de PLN existem várias ferramentas, E dentre todas as ferramentas utilizadas, como já pontuado, destaca-se o SpaCy. Esta é uma biblioteca de código, gratuita, que oferece um pacote completo para o processamento avançado de linguagem natural da linguagem, esta biblioteca baseada na linguagem de programação Python possibilita a criação de modelos de reconhecimento de entidades nomeadas por meio da customização de modelos pré-treinados disponibilizados em inúmeros idiomas, inclusive no idioma português.

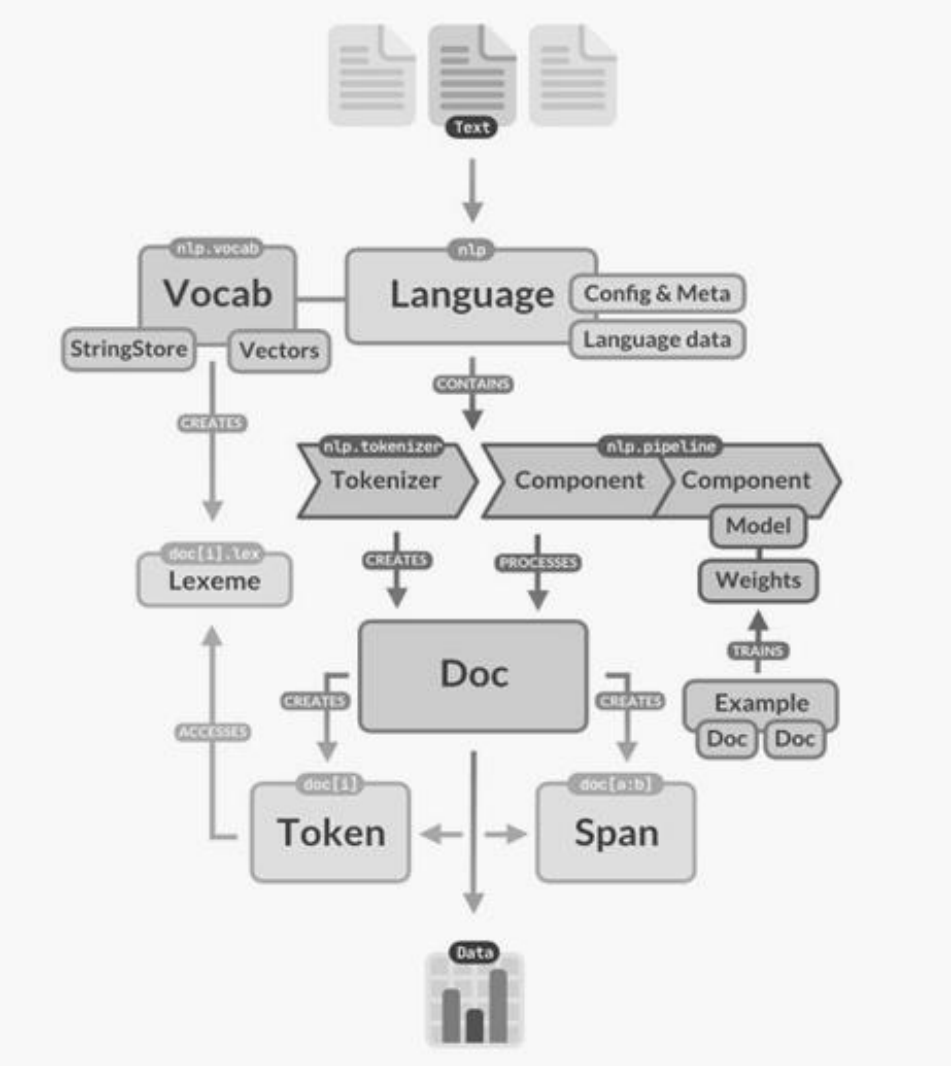
6.4 Reconhecimento de Entidade Nomeada com a Biblioteca SpaCy

Como pontuado, a biblioteca SpaCy disponibiliza métodos rápidos de pré-processamento de texto, incluindo modelos neurais pré-treinados, para serem utilizados em aprendizado por transferência. A biblioteca SpaCy também, compõem-se de vetores de palavras e suporta tarefas de tokenização em dezenas de idiomas.

Logo, ela permite uma rápida criação de modelos customizados de Rede Neural Convolucional (ConvNet / Convolutional Neural Network ou CNN) ou de modelos de arquitetura neural Transformer, para marcação, reconhecimento de entidade nomeada, promovendo uma fácil integração de aprendizado profundo. Assim, as ferramentas disponibilizadas pela biblioteca SpaCy permite a fácil realização de três tarefas básicas de PLN, sendo elas: a análise sintática de dependência (para análise e determinação de possíveis relações entre palavras em uma dada sentença), a marcação de uma parte do discurso (servindo para a identificação substantivos, verbos e outras partes componentes, deste) e, enfim, o

reconhecimento de entidades nomeadas (que identifica e ordena grupos nomes próprios, em categorias) (SPACY.IO, 2022).

Figura 16: Arquitetura geral da biblioteca SpaCy.



Fonte: SpaCy.io (2022).

Como mostra a figura 16, a arquitetura SpaCy é composta por diferentes objetos, dividindo-se em três grupos de classes principais, são elas: a Vocab, Language e a Doc.

A classe Vocab centraliza strings, vetores de palavras e atributos lexicais, isso, logo após o processamento de um determinado texto, assim, palavras e pontuação são armazenadas no objeto de Vocab. Esta classe conta com dois componentes de pipeline: a StringStore que mapeia strings de e, para valores de hash (ou seja, retorna valores hash). E o Vector para dados vetoriais codificados por string (SPACY.IO, 2022).

A classe Language é responsável pelo processamento de texto para transformá-lo em um objeto Doc. Assim, quando o texto é processado, ele é primeiro tokenizado e modificado e,

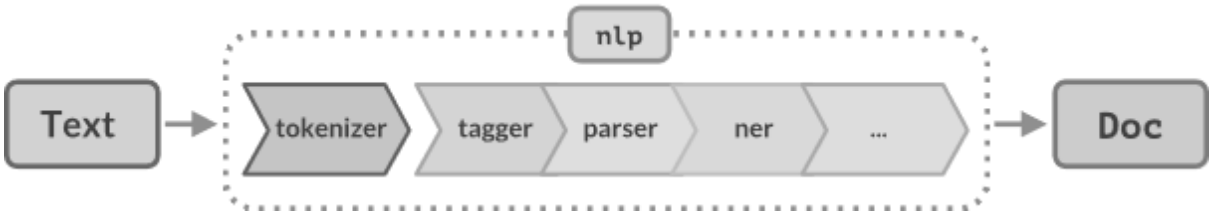
por fim, retornando o objeto Doc com novas informações. O tokenizer lida como o texto bruto e cria objetos Doc a partir de palavras. Ele é uma pipeline de processamento e é executado antes dos componentes. O objeto Example é uma coleção de anotações de treinamento, contendo dois objetos Doc: contendo os dados de referência e as previsões (SPACY.IO, 2022).

A classe Doc que conforme a documentação é um contêiner responsável pelo acesso de anotações linguísticas. A Doc é a classe mais importante, responsável pela serialização, bem como, pelos dados de treinamento, e pelo armazenamento dos dados tratados pelo Span, Token e Lexeme. Onde o Span é a fatia das anotações linguísticas acessadas pelo Doc; o Token lida com símbolo individual de caracteres de palavra, como espaço em branco e/ou símbolo de pontuação. E, o Lexeme é o oposto de token, pois não lida com o contexto ou análise de dependência, logo, lidando com as palavras sem contexto. (SPACY.IO, 2022).

Conforme a documentação, a classe de pipeline de processamento consiste em um ou mais componentes de pipeline. E referidos componentes de pipeline podem conter um modelo estatístico e pesos treinados ou, apenas fazer modificações baseadas em regras no arquivo Doc. O SpaCy fornece uma variedade de componentes integrados para diferentes tarefas de processamento de linguagem e também permite adicionar componentes personalizados. Os métodos disponíveis no SpaCy para NER atribuem um rótulo aos dados de texto e classificam os mesmos conforme previamente definido. O SpaCy também disponibiliza uma opção para adicionar classes arbitrárias aos sistemas de reconhecimento de entidades e atualizar ou customizar o modelo para incluir novos exemplos de entidades. Facilitando o treinamento de modelos de dados de domínio, para solucionar necessidades específicas. (SPACY.IO, 2022).

O SpaCy disponibiliza a pipeline de processamento de redes neurais (tok2vec), pré-treinadas em modelos de aprendizado profundo, baseadas em arquitetura CNN.

Figura 17: A pipeline de processamento tok2vec CNN do SpaCy.

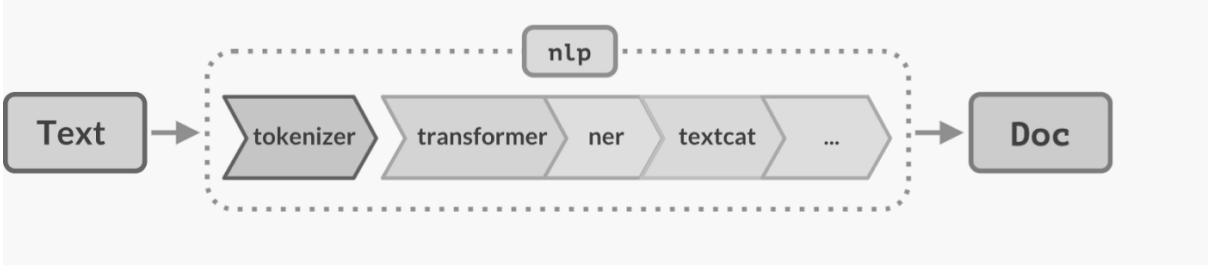


Fonte: SpaCy.io (2022).

O SpaCy, também oferece os Transformers (transformadores), que são um tipo de arquitetura específica para modelos de aprendizado profundo que constituem o estado da arte, de modelos neurais aplicados às resoluções de tarefas de processamento de linguagem natural.

A característica que delineia e define a arquitetura Transformers, como já pontuado, é o chamado mecanismo de “self-attention” ou mecanismo de “auto-atenção”. Onde, usa-se cada palavra, para aprender como está se relaciona com as outras palavras em uma sequência. Assim, os transformers computam representações complexas, no contexto tokenização de textos. Deste modo, a pipeline de processamento do SpaCy usa representações como recursos de entrada para melhorar previsões, conectando vários componentes a um único modelo de transformer.

Figura 18: A pipeline de processamento transformer do SpaCy.

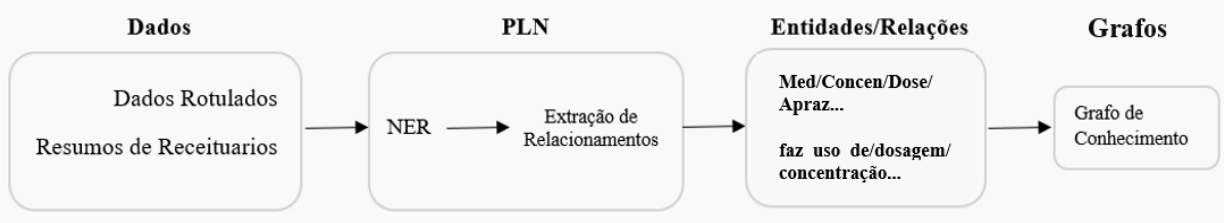


Fonte: SpaCy.io (2022).

Conforme destaca a documentação do SpaCy, sua arquitetura transformer interage com o PyTorch e a biblioteca de transformers do HuggingFace (comunidade de aprendizado de máquina e IA, que disponibiliza modelos e dados para construção de AI/IA), o que significa o acesso a milhares de modelos pré-treinados para seus pipelines. O que, para fins práticos, permite uma maior precisão dos modelos, com menor custo do tempo de treinamento e do tempo de execução de um modelo de aprendizado de máquina ou profundo.

Logo, este trabalho de pesquisa usa o SpaCy para compor o modelo de recuperação de informações medicamentosa descrita em modelos textuais extraídos de receituários de prontuários eletrônicos. Especificamente, usa a pipeline de processamento do SpaCy para treinar um modelo de NER, para a o reconhecimento e extração de entidades nomeadas e seus relacionamentos, e o posterior armazenamento dessas, em um banco de dados gráfico.

Figura 19: Fluxo da extração de NER e REL do modelo.



Fonte: Elaborado pela autora (2022).

Finalmente, esse modelo usa um banco de dados gráficos para armazenar as entidades reconhecidas e extraídas, bem como seus relacionamentos. Faz-se também, o uso deste para a construção e incorporação de um grafo de conhecimento contendo informações vinculadas, de medicamentos de referência e seus similares, e outras informações, como princípio ativo.

6.5 Banco de Dados Gráfico Neo4j para Criação de Grafos do Conhecimento

Os denominados “bancos de dados gráficos”, foram pensados e projetados para otimizar o armazenamento de redes de informação. Onde, as conexões entre vários elementos informativos, como pessoas, organizações, objetos e lugares, pudessem ser facilmente ligadas e consultadas. Conforme Lyon (2022), estes bancos de dados permitem modelar, armazenar e consultar dados como uma rede ou grafo, otimizando o desempenho em comparação aos outros sistemas de banco de dados, como bancos de dados relacionais. Isto porque, os bancos de dados relacionais não foram projetados para realizar consultas e analisar relacionamentos complexos entre várias partes. Ao contrário, um banco de dados gráfico é muito ágil para responder a solicitações de pesquisas em redes definidas por conexões. Logo, estes são atualmente muito utilizados na construção de mecanismos de recomendação, e de grafos de conhecimento. Portanto os bancos de dados gráficos são otimizados para realizarem consultas que podem seguir uma ordem completamente aleatória, perfazendo “saltos” e, desta forma, coletar todos os nós e arestas dentro de uma rede de informação.

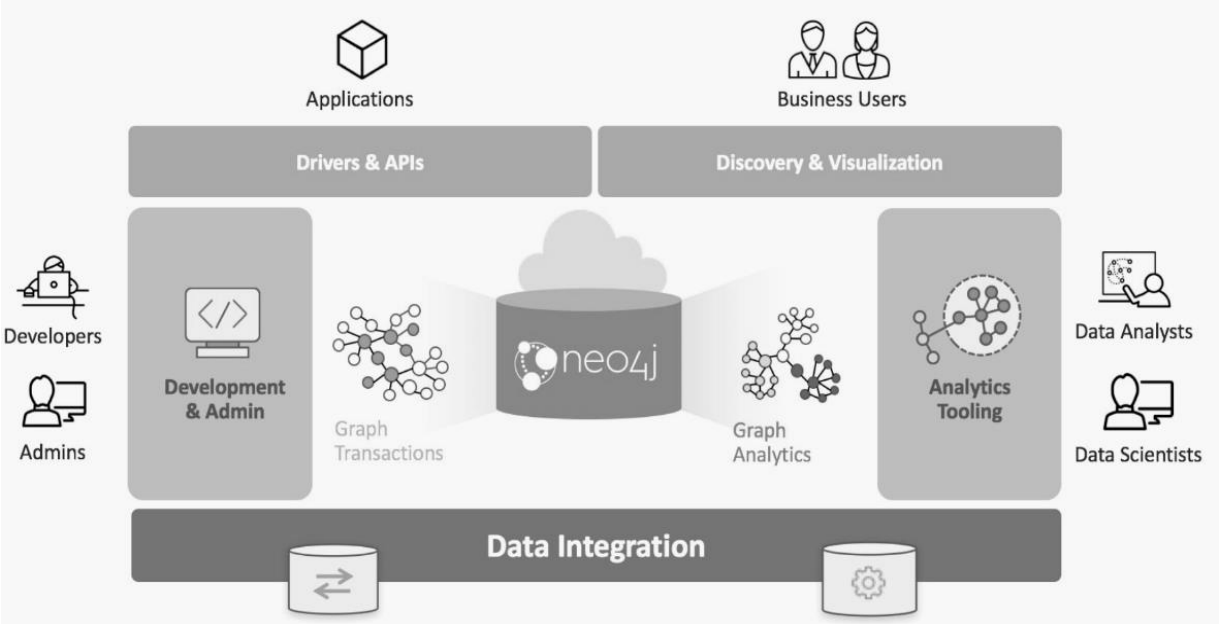
Dentre os variados bancos de dados gráficos disponibilizados o Neo4j se destaca como um dos mais populares. A plataforma Neo4j possibilita a criação de sistemas de informação bem arquitetados, com aplicações baseadas em estruturas de grafos, possibilitando uma alta conectividade entre dados heterogêneos. O SGBD (Sistema Gerenciador de Banco de Dados)

do Neo4j, entre outras, está disponível em uma versão gratuita, sob uma licença GPL3, comunitária e de código aberto. (NEO4J.COM, 2022).

O SGBD do Neo4j inclui os seguintes componentes e recursos:

- Banco de dados gráfico nativo.
- Análise de gráficos.
- Integração de dados.
- Linguagem Cypher para consulta de grafos cifrados.
- Ferramentas de visualização de dados para usuários não técnicos.

Figura 20: A plataforma gráfica Neo4j.



Fonte: Neo4j.com (2022).

Conforme mostrado na figura 20, a plataforma do Neo4j além de armazenar os dados como banco de dados gráfico geral, também oferece uma variedade de ferramentas e recursos voltados para aplicações de ciência e análise de dados.

O Neo4j facilita a criação de aplicações de incorporação de gráficos e de processamento de linguagem natural, pois seus recursos oferecem suporte a filtragem de recomendações e para a construção, agregação e incorporação de grafos de conhecimento. Portanto, com base nos dados de integração, e com o uso da sua linguagem Cypher, em conjunto com a linguagem Python, o banco de dados Neo4j possibilita a conexão de dados evidenciando ligações que não podem ser realizadas facilmente pelos bancos de dados convencionais.

O processo de descoberta da informação e do conhecimento, permitida pelos bancos de dados gráficos, de forma otimizada e flexível, facilita uma inferência das relações entre os dados. Assim, constrói-se redes de relacionamento entre entidades variadas e heterogêneas, que se ligam e evidenciam novas descobertas de conhecimento até então não evidentes (NEO4J.COM, 2022).

Posto isso, este modelo faz uso do banco de dados gráfico Neo4j para criar, armazenar e consultar o grafo de conhecimento contendo informações de medicamentos descritos em modelos de receituários de prontuário eletrônicos.

7 APRESENTAÇÃO E INTERPRETAÇÃO DOS RESULTADOS

Por meio dos resultados obtidos, foi possível inferir que o crescimento e acúmulo de documentos de textos médicos/clínicos gerados e armazenados em sistemas de prontuários eletrônicos do paciente, cada vez mais, tendem a se tornarem um problema. Especificamente, no que se refere a recuperação da informação contida em textos em formato não-estruturados, que compõem mais de dois terços das informações armazenadas nestes.

Com a análise aprofundada das literaturas de referência componente do corpus documental desta pesquisa, bem como, dos trabalhos científicos desenvolvidos ao longo do processo, e dos estudos técnicos empreendidos. Foi possível verificar uma forte tendência para o uso de técnicas e tecnologia da informação baseadas em processamento de linguagem natural, e aprendizado de máquina, bem como, do uso de grafos de conhecimento, para recuperar e representar a informação e o conhecimento no domínio da saúde.

Neste esforço que visa melhorar a recuperação e visualização da informação nestes sistemas de prontuários eletrônicos, este trabalho científico sugere que, efetivamente, é muito promissor o uso de técnicas e recursos de PLN como, o Reconhecimento de Entidades Nomeadas, bem como, o uso de estruturas de Grafos de Conhecimento, para a solução de tarefas de recuperação e estruturação do conhecimento gerado e contido em prontuários eletrônicos. E, especificamente, mostra-se muito oportuno na recuperação e visualização de informações de dados de medicamentos descritos em prontuários eletrônicos do paciente.

7.1 Etapas e Resultados: Recuperação e Visualização da Informação

Como pontuado anteriormente, a parte prática e/ou aplicada desta pesquisa percorre uma série de etapas sequenciais. Sendo que a primeira destas referidas etapas foi iniciada com a construção e formalização de um conjunto de dados de anotações contendo variadas sentenças descrevendo informações de receituários médicos, onde são, entre outros, registrados dados de especial importância como: nome de medicamentos, concentração de referidos medicamentos, dosagem e o tempo e/ou intervalo de administração do medicamento.

Por fatores associados aos custos de processamento, este conjunto de dados foi limitado a 1300 entradas de sentenças de dados de medicamentos. Onde 1000 são usadas para treinamento do modelo e 300 para o teste do modelo. Esta divisão entre dados de treino e teste conforme Gholamy *et al*. (2018), deve ser feita com a relação 70/30 ou 80/20, posto que, ao aprender de uma dependência de dados, empiricamente, estas relações contribuem para evitar o overfitting, ou seja, o sobre-ajuste do modelo. Assim, primeiro treina-se o modelo no conjunto

de treinamento e, em seguida, usa-se os dados do conjunto de teste para avaliar a precisão e eficácia do modelo resultante. Posto isso, o conjunto de dados deste trabalho usou a relação percentual 70/30 de treino/teste, onde principalmente os tipos ou nomes de medicamentos do conjunto de testes são diferentes dos usados no conjunto de dados de treinamento.

Figura 21: Exemplos de dados usados nos conjuntos de sentenças, para treino e testes.

```
"O paciente usa 2 capsulas de glimepirida de 25 mg de 4 em 4 horas. A se
"O paciente usa o medicamento Cebrilin de 20 ml, 30 gotas de 12/12 hs El
"Ela toma 15 gts de Fludalibbs de 10 ml 4/4 hs O paciente diz que sentir
"O paciente usa 1 comprimido de cloridrato de ranitidina de 15 mg de 6 e
"O paciente usa 25 gts do medicamento Confilify de 20 ml 6/6 hs A senhor
"O Sr. usa 30 gotas do medicamento aripiprazol de 10 ml. 4/4 hs O Sr. de
"A paciente tomou 20 gotas de Benicaranlo de 10 ml de 2 em 2 horas. O Sr
"O paciente faz uso de 2cp de C-Platin de 30 mg 1 por dia. O Sr. declara
"Ele faz uso de 30 gotas do medicamento Ocylin de 10 ml de 8 em 8 horas.
"A senhora toma 30 gotas de Bremetix de 20 ml 2 vezes por dia. Ele infor
"O Sr. tomou 3 gts Dormium de 14 ml 4 vezes por dia. A paciente disse qu
"O paciente faz uso de 2 COMPR valsartana + besilato de anlodipino de 10
"Ele faz uso de 3 gts Dormium de 30 ml 1 vez por dia. Ele relata que sen
"A paciente usa 1 Cápsula do medicamento cloridrato de doxorrubicina de
```

Fonte: Elaborado pela autora (2022).

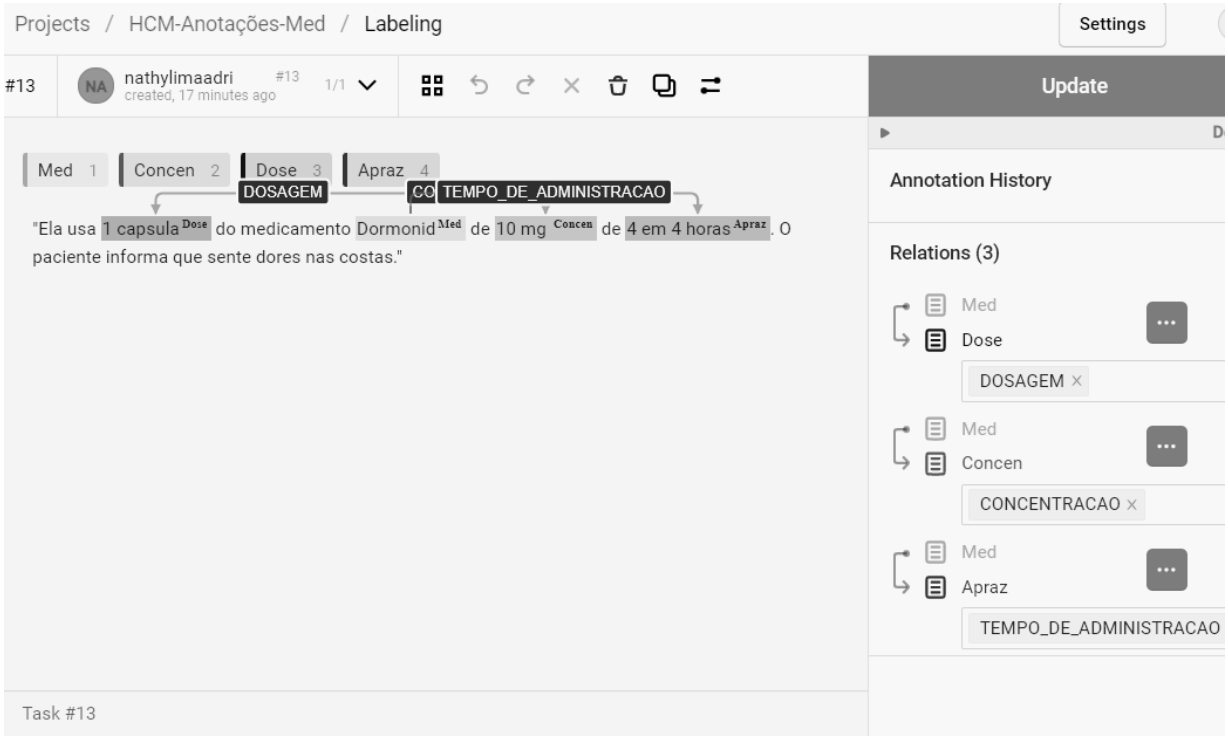
Uma vez de posse desses conjuntos de dados, procedeu-se à anotação ou rotulação dos dados, posto que, este é um modelo de aprendizado supervisionado. As anotações dos dados deram-se no formato IOB, para se adaptar às conformidades do modelo BERT do SpaCy.

Quando finalizado o processo de rotulagem das entidades constituintes deste modelo de Reconhecimento de Entidade Nomeada, ou seja, as anotações das entidades do tipo: MED, CONCEN, DOSE e APRAZ. Os conjuntos de dados foram salvos com a extensão .tsv que posteriormente (como mostrado nos *pipelines* descritos no apêndice) são convertidos para o formato .json e posteriormente para o formato .spacy ficando no formato final para uso.

Este trabalho dispõe de dois modelos de aprendizado de máquina, o modelo de NER cujo os primeiros passos foram descritos anteriormente, e o modelo de REL, ou seja, um modelo de extração de relacionamento de ou entre entidades nomeadas.

Para a construção deste modelo foi usado os conjuntos de dados citados anteriormente, com o acréscimo de um novo conjunto similar ao de teste, mas com a denominação dev. Assim, para o treinamento deste modelo usa-se os conjuntos de dados de treino/teste/dev. Estes conjuntos de dados foram rotulados usando a ferramenta Label Studio (como mostrado na figura 24), que agiliza e melhora o padrão das anotações ou rotulagens. Estes arquivos depois de devidamente rotulados com demarcações de entidades e seus respectivos relacionamentos foram salvos com a extensão .json e posteriormente convertidos para a extensão .spacy.

Figura 22: Anotações de NER e REL com a ferramenta Open-source, Label-Studio.



Fonte: Elaborado pela autora (2022).

Figura 23: Anotações de NER e REL no formato Json.


```
[{"id":20,"annotations":[{"id":20,"completed_by":1,"result":[{"value":{"start":20,
"end":29,"text":"1 capsula","labels":["Dose"]},"id":"0NGXkdNawl","from_name":
"label","to_name":"text","type":"labels","origin":"manual"},{"value":{"start":46,
"end":57,"text":"Olmetecanlo","labels":["Med"]},"id":"W4AqUDD7pt","from_name":
"label","to_name":"text","type":"labels","origin":"manual"},{"value":{"start":63,
"end":67,"text":"25mg","labels":["Concen"]},"id":"v-crgMYGPc","from_name":"label",
"to_name":"text","type":"labels","origin":"manual"},{"value":{"start":72,"end":84,
"text":"6 em 6 horas","labels":["Apraz"]},"id":"j3F9kW8lIm","from_name":"label",
"to_name":"text","type":"labels","origin":"manual"},{"from_id":"W4AqUDD7pt","to_id"
:"0NGXkdNawl","type":"relation","direction":"right","labels":["DOSAGEM"]},{
"from_id":"W4AqUDD7pt","to_id":"v-crgMYGPc","type":"relation","direction":"right",
"labels":["CONCENTRACAO"]},{"from_id":"W4AqUDD7pt","to_id":"j3F9kW8lIm","type":
"relation","direction":"right","labels":["TEMPO_DE_ADMINISTRACAO"]}],
"was_cancelled":false,"ground_truth":false,"created_at":
"2022-12-24T22:59:49.077077Z","updated_at":"2022-12-24T22:59:49.077077Z",
"lead_time":94.437,"prediction":{},"result_count":0,"task":20,"parent_prediction":
null,"parent_annotation":null}],"file_upload":"3b6038b7-dataset_med_treino.txt",
"drafts":[],"predictions":[],"data":{"text":"\"O Sr.  faz uso de  1 capsula do
medicamento  Olmetecanlo  de  25mg  de 6 em 6 horas.  Ele  relata que  sente dores
```


Fonte: Elaborado pela autora (2022).

A partir da rotulação e das conversões necessárias dos conjuntos de dados, os modelos NER e REL foram treinados usando a arquitetura transformer do SpaCy, e por meio do mesmo

foram testados alcançando um nível aceitável de precisão em suas previsões. Como evidenciam os percentuais ou taxa de precisão mostrada nas figuras 24 e 25.

Figura 24: Precisão, Recall e Fscore do modelo NER.

```
ℹ Pipeline: ['transformer', 'ner']
ℹ Initial learn rate: 0.0
E    #       LOSS TRANS...  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  -------------  --------  ------  ------  ------  ------
  0       0         751.13    963.28    0.00    0.00    0.00    0.00
 12     200       87598.82  73167.13   99.07   98.76   99.38    0.99
 25     400           0.00      0.00   98.88   98.51   99.25    0.99
 37     600           0.00      0.00   98.88   98.51   99.25    0.99
```

Fonte: Elaborado pela autora (2022).

Figura 25: Precisão, Recall e Fscore do modelo REL.

```
ℹ Pipeline: ['transformer', 'relation_extractor']
ℹ Initial learn rate: 0.0
E    #       LOSS TRANS...  LOSS RELAT...  REL_MICRO_P  REL_MICRO_R  REL_MICRO_F  SCORE
---  ------  -------------  -------------  -----------  -----------  -----------  ------
  0       0           0.80           1.25         7.26        99.57        13.54    0.14
  5     100          36.04          27.23        87.36        88.60        87.98    0.88
 10     200           0.08           0.48        91.16        91.03        91.09    0.91
 15     300           0.01           0.12        91.40        90.88        91.14    0.91
```

Fonte: Elaborado pela autora (2022).

As nomenclaturas dos dados de performance vistas na figura 24: ents_p, ents_r e ents_f são a precisão, recall e fscore para a tarefa NER, conforme o SpaCy (2022) estes elementos são calculados tendo por base a entidade. Ou seja, o SpaCy considera todas as entidades no documento ou corpus, para encontrar verdadeiro positivo, falso positivo e falso negativo. O mesmo vale para os elementos de performance do modelo REL: rel_micro_p, rel_micro_r, rel_micro_f. Como mostram as figuras anteriores, a precisão dos modelos treinados alcançou níveis acima de 91% de precisão no modelo REL e ainda maior no modelo NER.

A saber, Precisão, Recall e Fscore são métricas de avaliação. Onde a métrica de Precisão quantifica o número de previsões de classes positivas (neste caso de uso: entidades e relacionamentos) que verdadeiramente pertencem à classe positiva. A métrica de Recall, por sua vez, quantifica o número de previsões de classes positivas feitas de todos os exemplos positivos no conjunto de dados. Finalmente, a métrica de Fscore combina Precisão e Recall para representar ou ponderar o valor central, usando a média harmônica.

Como mostrado na figura 12 e anteriormente pontuado, para a conclusão desse trabalho em sua vertente aplicada percorreu-se dois caminhos, o primeiro tratou-se da recuperação de informações pontuais de dados não estruturados (sentenças) contidos em anotações de

receituário eletrônico. Isso, por meio da tarefa de PLN denominada Reconhecimento de Entidade Nomeada, onde o modelo de NER e REL foram treinados, salvos e aplicados para reconhecer e extrair as entidades: MED, CONCEN, DOSE, e APRAZ, bem como, os relacionamentos: DOSAGEM, CONCENTRACAO e TEMPO_DE_CONCENTRACAO.

Assim, uma vez reconhecidas e extraídas as entidades e respectivos relacionamentos de um conjunto/exemplo com 300 entradas de dados (sentenças demarcadas por ID de pacientes). Foi usada a linguagem de consulta Cypher para construir scripts para salvar e persistir os dados extraídos, em uma estrutura de grafo de conhecimento no banco de dados Neo4j. Esse grafo de conhecimento gerado a partir da extração de entidades e relacionamentos via PLN, foi agregado a um outro grafo que contêm uma lista de medicamentos de referência, similares e outras informações, contendo um total de 5.317 medicamentos e seus princípios ativos.

O referido grafo de conhecimento foi gerado a partir de uma lista denominada “Lista de Medicamentos Similares e seus respectivos medicamentos de referência”, disponibilizada pela Anvisa. Esta lista disponibilizada no formato PDF foi convertida para CSV e rearranjada para compor 3 CSVs contendo todas a informações e, por meio de scripts Cypher foi convertido em um grafo de medicamentos, que enfim junta-se ao grafo med-paciente para construir as visualizações que compõem um banco de dados com milhares de nós, relacionamentos e propriedades.

Entende-se que a extração, agrupamento e vinculação de informações a partir da recuperação dos dados de medicamentos descritos em receituários médicos, é uma abordagem muito promissora. Modelos pré-treinados de processamentos de linguagem natural baseados em aprendizado profundo, que requerem um número reduzido de parâmetros para o treinamento de um modelo de ajuste fino de reconhecimento de entidades nomeadas e seus relacionamentos. Ainda, com a análise aprofundada das literaturas de referência, bem como, com os trabalhos científicos desenvolvidos ao longo do processo e dos estudos técnicos empreendidos. Foi possível analisar e testar técnicas e tecnologias para a construção e uso de grafos de conhecimento, para recuperar e visualizar a informação e o conhecimento no domínio dos receituários de medicamentos.

Assim, também foi possível criar um modelo de previsão de reconhecimento de entidades para extrair, agrupar e vincular informações, a partir da recuperação dos dados de medicamentos, por meio do mapeamento de entidades nomeadas e seus relacionamentos reconhecidas e extraídas em exemplos textuais não estruturados de receituários.

Logo, neste esforço que visa melhorar a recuperação e visualização da informação nestes sistemas de prontuários eletrônicos. Os resultados parciais sugerem que, efetivamente, é

muito promissor o uso de técnicas e recursos de PLN como, o Reconhecimento de Entidades Nomeadas, bem como, o uso de estruturas de Grafos de Conhecimento, para a solução de tarefas de recuperação e visualização do conhecimento gerado e contido em prontuários eletrônicos.

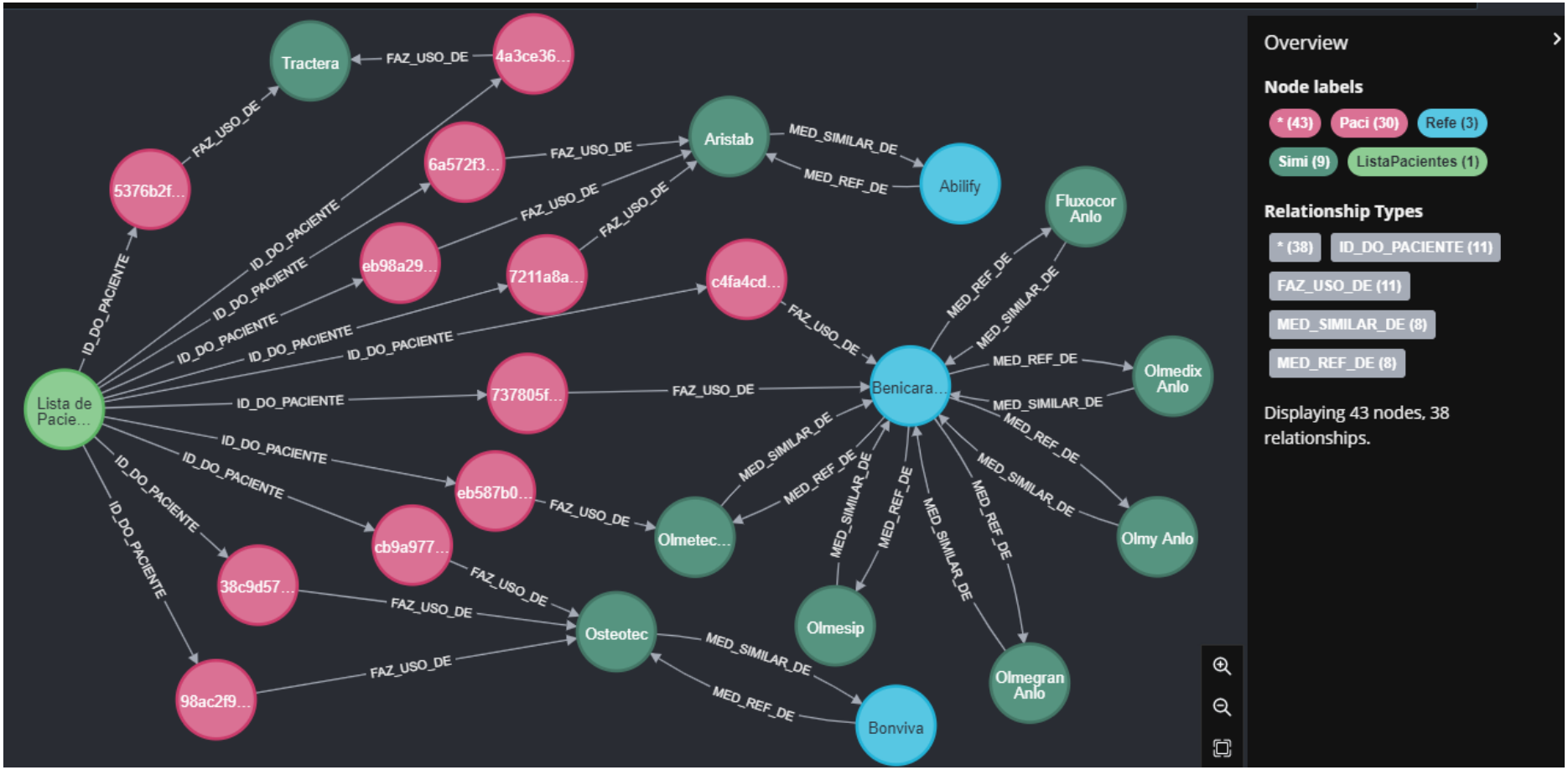
Figura 26: Entidades nomeadas sendo reconhecidas e extraídas de um texto não estruturado de exemplo de anotações de medicamentos, por meios do modelo de NER.

Sentenças		Entidades	Rótulos
O medicamento Abilify de 20 mg foi indicado como terapia para o tratamento agudo de episódios associados ao transtorno bipolar do tipo I. Tomar 1 comprinido de 4/4 hs O mecanismo de ação do aripiprazol, como ocorre com outras drogas eficazes no tratamento de transtorno bipolar, com\n2 cp do medicamento Confilify de 35Mg de 5 em 5 horas. No entanto, foi proposto que a eficácia do Sensaz, aripiprazol é mediada por efeitos no sistema nervoso central. A atividade de Abilify principalmente devida à droga inalterada, aripiprazol, e em menor medida ao seu, dehidro-aripiprazol. sendo o Aripiprazol o o Kavium Aipri bem como do Harip e Toarip.	0	Abilify	MED
	1	20 mg	CONCEN
	2	1 comprinido	DOSE
	3	4/4 hs	APRAZ
	4	aripiprazol	MED
	5	2 cp	DOSE
	6	Confilify	MED
	7	35Mg	CONCEN
	8	5 em 5 horas	APRAZ
	9	Sensaz	MED
	10	, aripiprazol	MED
	11	Abilify	MED
	12	aripiprazol	MED
	13	dehidro-aripiprazol.	MED

Fonte: Elaborado pela autora (2022).

Além disso, o uso de métodos e tarefas de PLN e Grafos, demonstraram, até aqui, como essas técnicas podem ser aplicadas para melhorar a eficiência na recuperação de informações de prontuários eletrônicos. O que indica que possibilitará a implementação de uma base de conhecimento de medicamentos em um banco de dados gráficos, que permitirá a recuperação e exploração de informações de medicamentos de maneira eficiente e escalável. Posto isso, os resultados alcançados estão alinhados com os objetivos da pesquisa e mostram grande potencial para a aplicação prática do modelo para a recuperação e visualização de dados de medicamentos.

Figura 27: Grafo de Conhecimento de medicações e pacientes, armazenada no Neo4j.

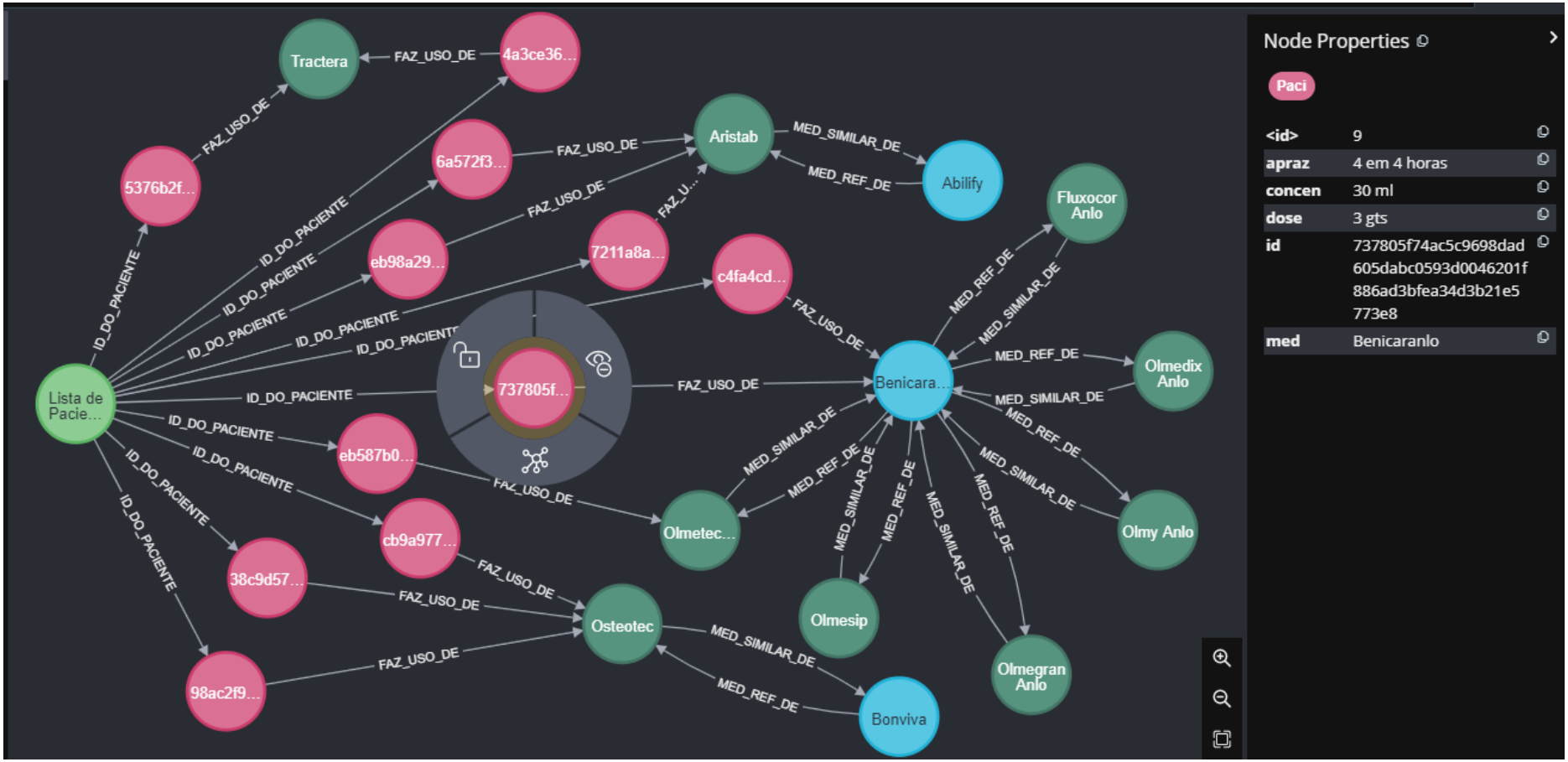


Fonte: Elaborado pela autora (2022).

A figura 27, (retirada do banco de dados gráfico Neo4j) mostra um grafo de conhecimento contendo três tipos específicos de nós, onde o nó de cor rosa identifica um determinado paciente, o nó azul representa o medicamento de referência e o nó verde representa os medicamentos similares. O grafo também é composto por certas relações denominadas arestas, onde a aresta **faz_uso_de**, evidencia a relação do paciente e o medicamento, por sua vez, a aresta **med_similar_de**, mostra a relação do medicamento de referência representado por **med_ref_de** com o medicamento similar que contém o mesmo princípio ativo, e a aresta do **id_do_paciente** representa a relação de um paciente específico a um determinado medicamento, bem como, outras propriedades.

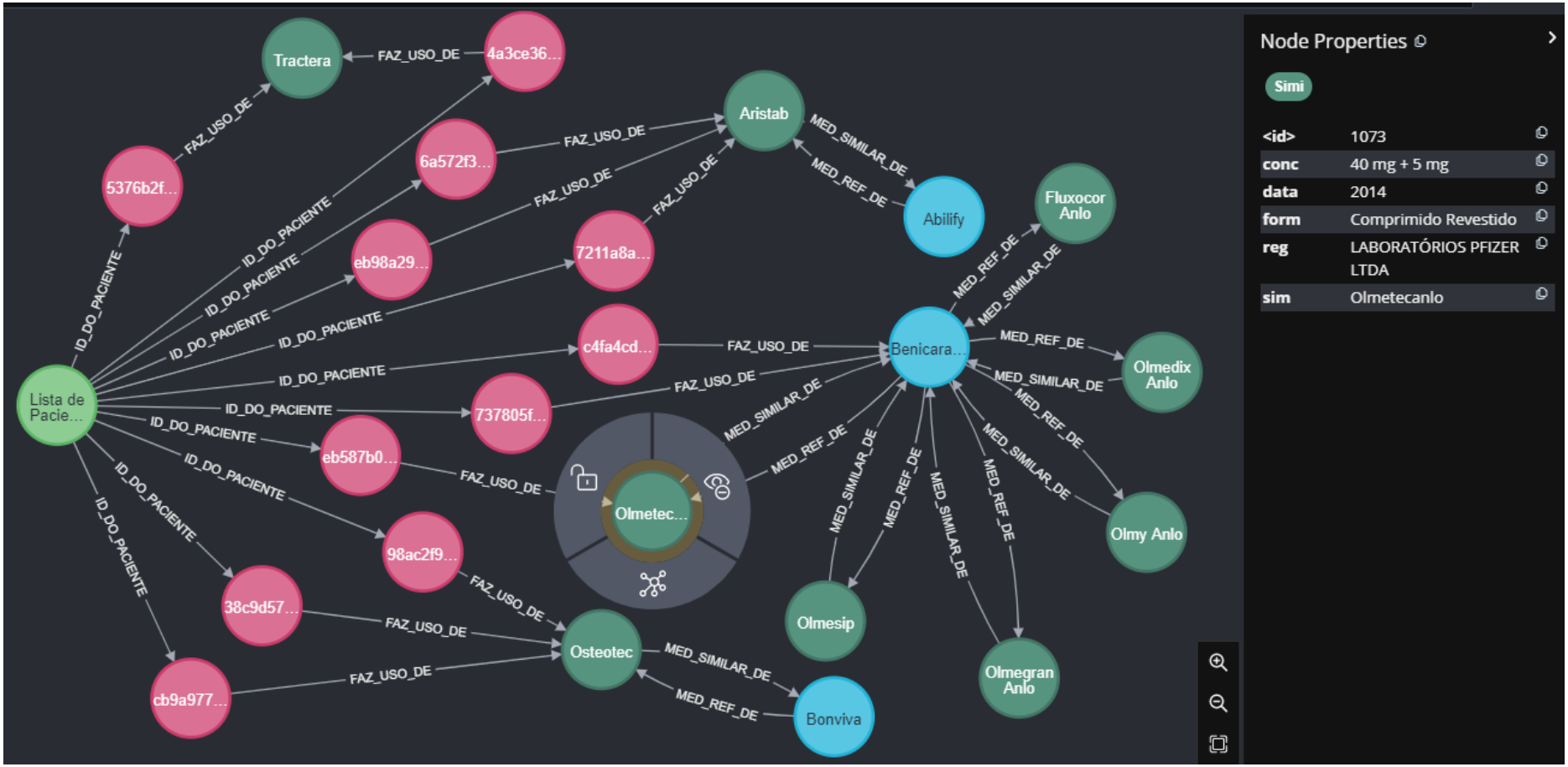
A figura 28, mostra as propriedades dos nós que identificam os pacientes. Onde é possível observar três propriedades medicamentosas, sendo elas: **med**, **concen** e **dose**. Sendo que, a propriedade ou entidade **med** referente ao medicamento "**Benicaranlo**" corresponde a medicação ministrada a um paciente específico e a relação com o seu **id**, a propriedade denominada **concen**, descreve a concentração da medicação, referenciada com "30 ml" sendo a concentração em miligramas e, finalmente, a propriedade **dose**, evidentemente, refere-se a dosagem ministrada representada com "3 gts" sendo a quantidade de gotas, e o tempo de intervalo representado como sendo de 4 em 4 horas.

Figura 28: Grafo de Conhecimento, informações de medicamento de um determinado paciente.



Fonte: Elaborado pela autora (2022).

Figura 29: Grafo de Conhecimento, propriedades dos medicamentos similares.

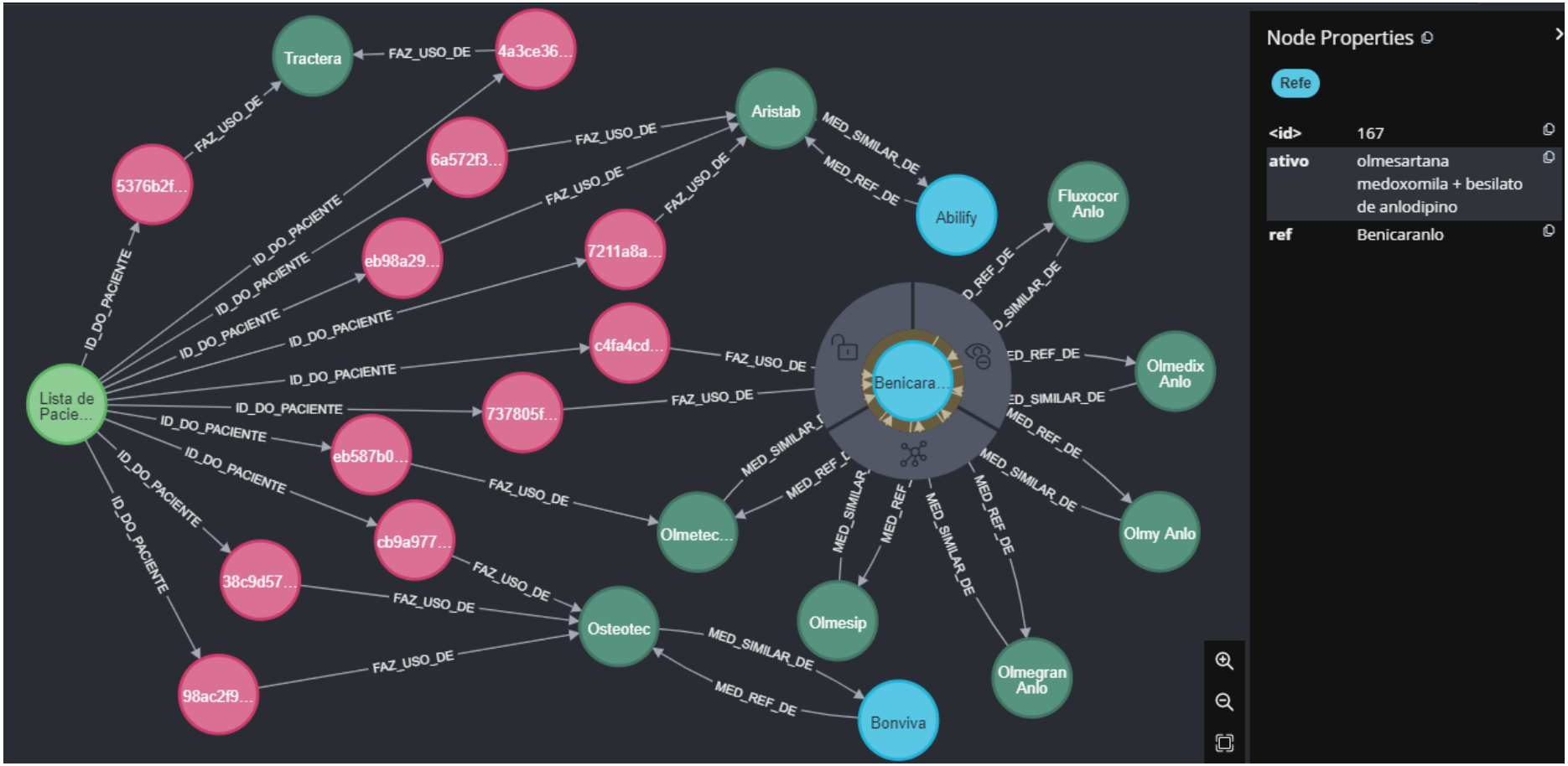


Fonte: Elaborado pela autora (2022).

A figura 29, destaca os nós de cor verde, que descrevem os medicamentos similares, onde suas arestas, vinculam-se ao medicamento de referência, e são, eventualmente, vinculados por pacientes que fazem uso destes. Como podem ser observados, estes, invariavelmente, contêm propriedades que descrevem a concentração, data de registro na Anvisa, formato, nome da empresa proprietária e o nome do medicamento.

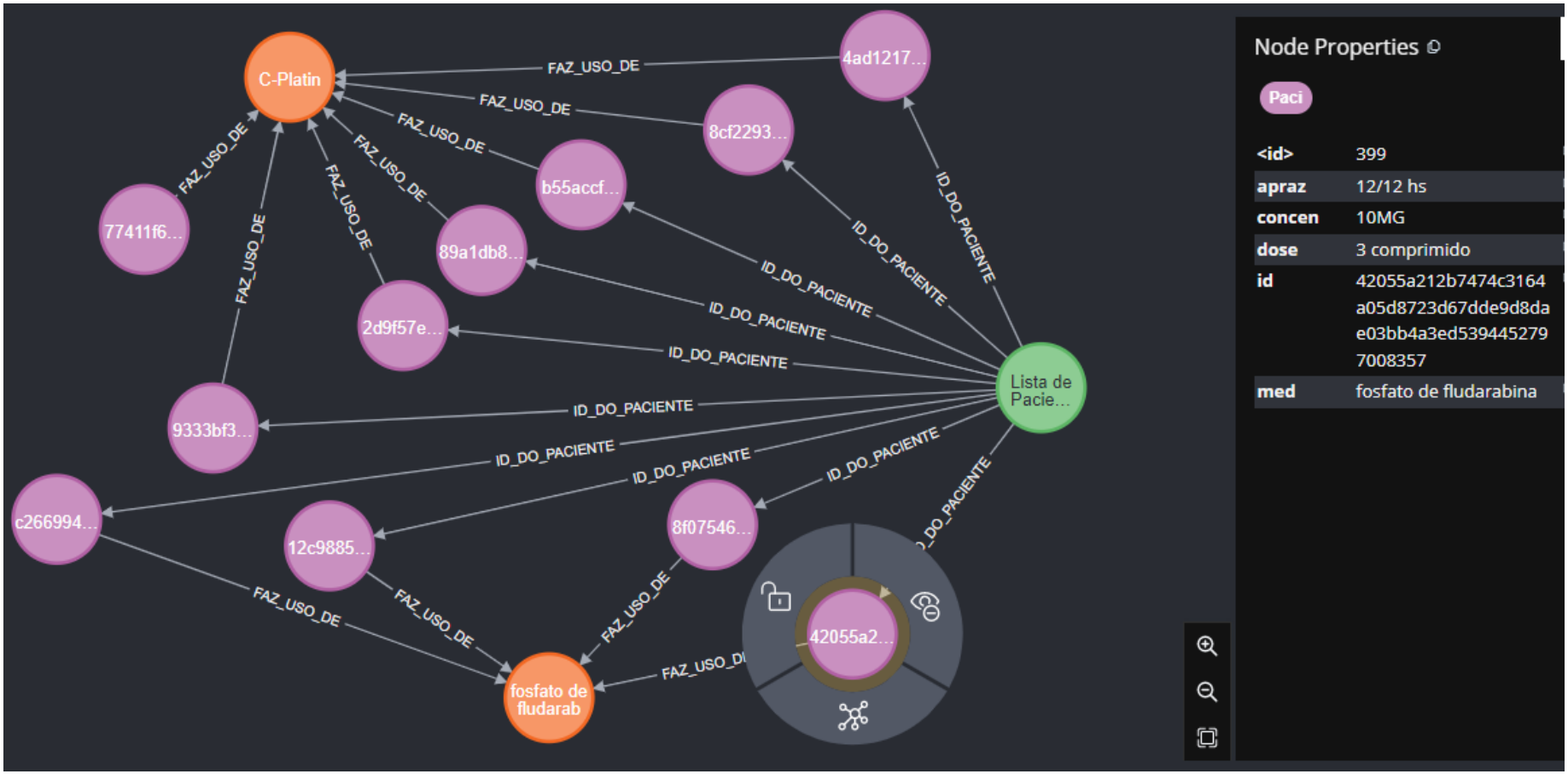
Abaixo, a figura 30, relaciona um determinado paciente, a uma medicação de referência, cujo relacionamento se dá com todos os seus medicamentos similares. Ela traz apenas duas propriedades, que são: o princípio ativo e o nome do medicamento.

Figura 30: Grafo de Conhecimento, relacionamentos e propriedade de um determinado medicamento de referência.



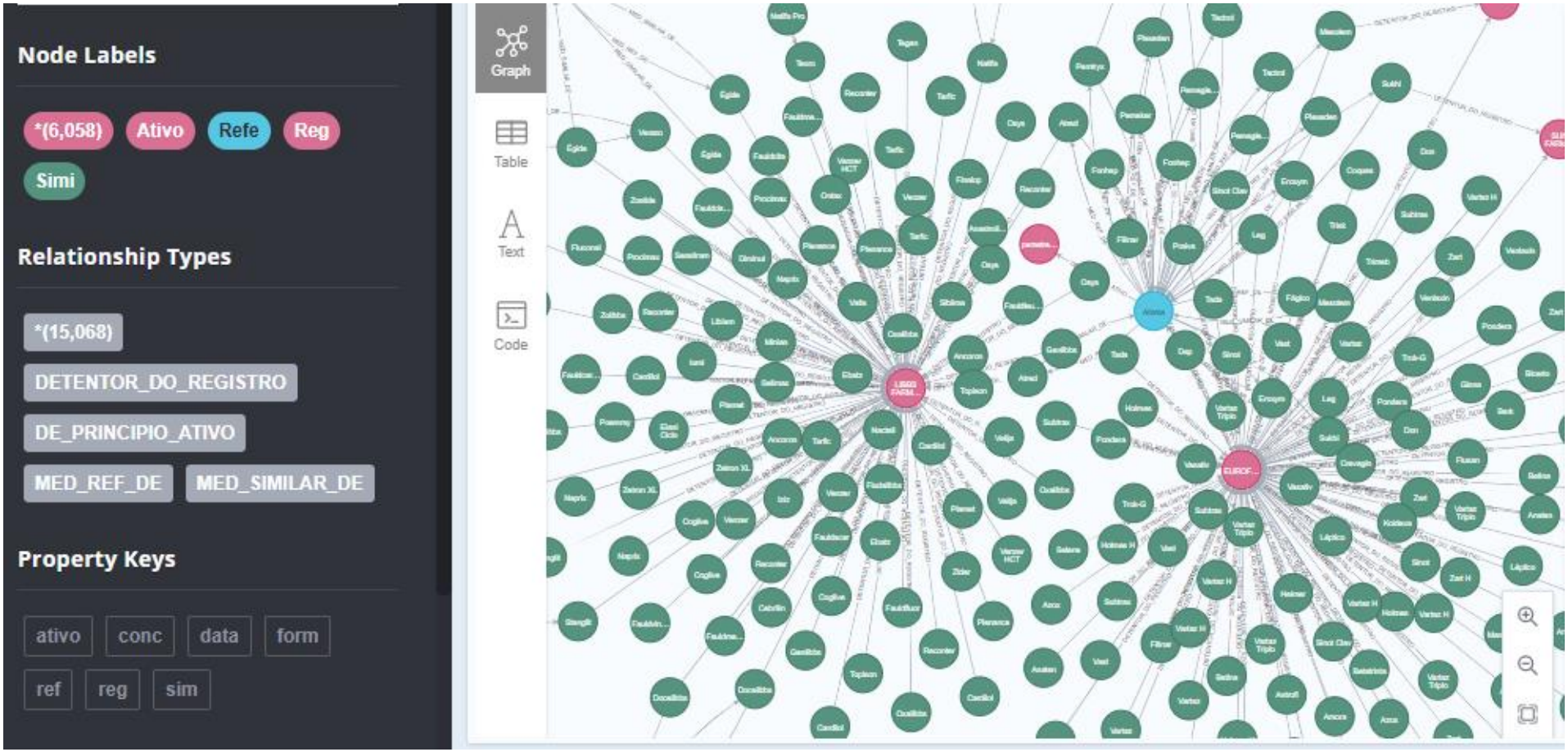
Fonte: Elaborado pela autora (2022).

Figura 31: Grafo de Conhecimento, outro exemplo de visualização, relacionado pacientes e medicamentos usados, e propriedades dos medicamentos e informações de administração.



Fonte: Elaborado pela autora (2022).

Figura 32: Grafo de Conhecimento, composto por todos os medicamentos de referência, seus similares e princípio ativo. De acordo com a lista da Anvisa 2021.



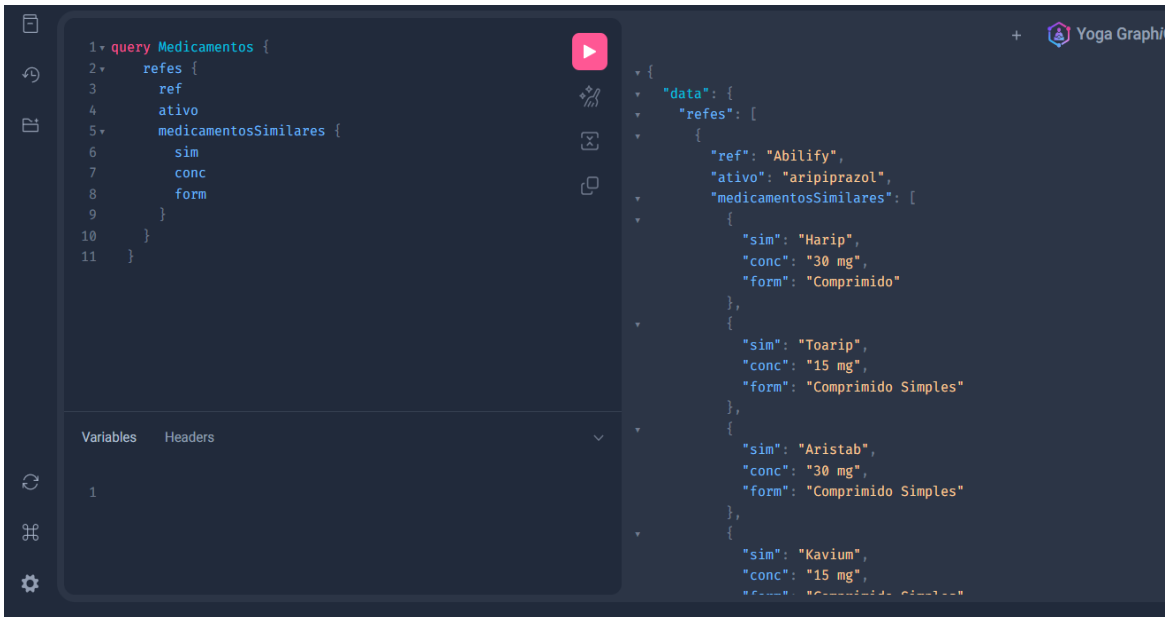
Fonte: Elaborado pela autora (2022).

A figura 32 (acima) traz a visualização de um banco de dados constituído por todos os medicamentos de referência, seus similares e princípio ativo, bem como, contempla todas as informações que consta da lista de medicamentos com 254 páginas contendo os medicamentos referência e similares aprovados e registrados pela Anvisa. Esta lista foi disponibilizada por esta instituição com atualizações feitas nos anos 2020/2021.

Especialmente, este banco de dados gráfico foi desenvolvido visando contribuir com a equipe do grupo de estudos HAIS, pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília. Tal contribuição está relacionada ao futuro desenvolvimento de um sistema que visa possibilitar que médicos no ato de receitar um determinado medicamento possa optar por um medicamento similar, ao invés de um medicamento de referência ("medicamento de marca"). Objetivamente, esse sistema funcionaria como um sistema de recomendação de medicamentos similares visando incentivar a indicação (por parte dos médicos) de medicamentos similares ou simplesmente por seu princípio ativo, notadamente mais baratos e, portanto, mais acessíveis a população em geral.

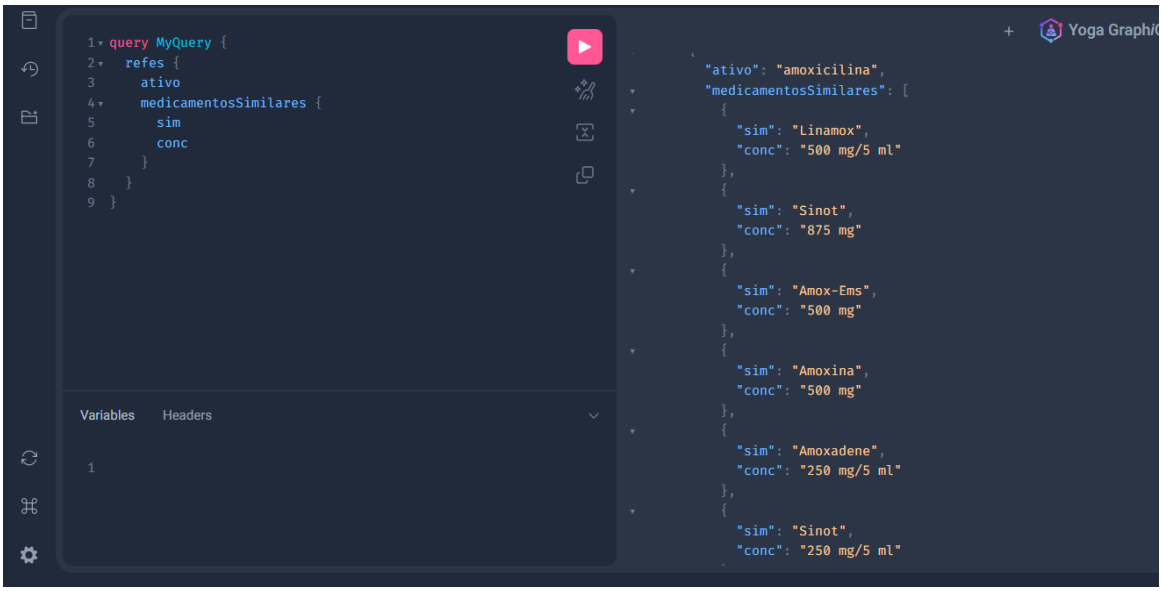
Tecnicamente, o desenvolvimento do referido "sistema de recomendação" de medicamentos similares pode ser, possibilitado, pelo banco de dados constituído por todos os medicamentos contidos no grafo de conhecimento desenvolvido nesta pesquisa, com medicamentos referência/similares. E, com o uso de uma ferramenta de consulta de banco de dados gráfico. Neste trabalho, como exemplo faz-se uso do GraphQL, que é uma linguagem de consulta para API, que possibilita a consulta de informações ou dados altamente relacionados.

Figura 33: Visualização do GraphQL consultando a API de medicamentos.



Fonte: Elaborado pela autora (2022).

Figura 34: GraphQL, outra visualização de consulta a partir do princípio ativo.



Fonte: Elaborado pela autora (2022).

Portanto, evidencia-se que com o uso do banco de dados gráficos Neo4j e da ferramenta GraphQL, bem como, com o uso de tarefas e métodos atuais de PLN, como o Reconhecimento de Entidades Nomeadas baseados em modelos de aprendizado profundo e Grafos de Conhecimento. É possível recuperar informações específicas possibilitando e potencializando a criação de variadas aplicações que possibilitam a otimização da recuperação e visualização da informação de medicamentos contidos em sistemas de prontuário eletrônico do paciente.

Figura 35: GraphQL, outra visualização de consulta a partir do princípio ativo.

Fonte: Elaborado pela autora (2022).

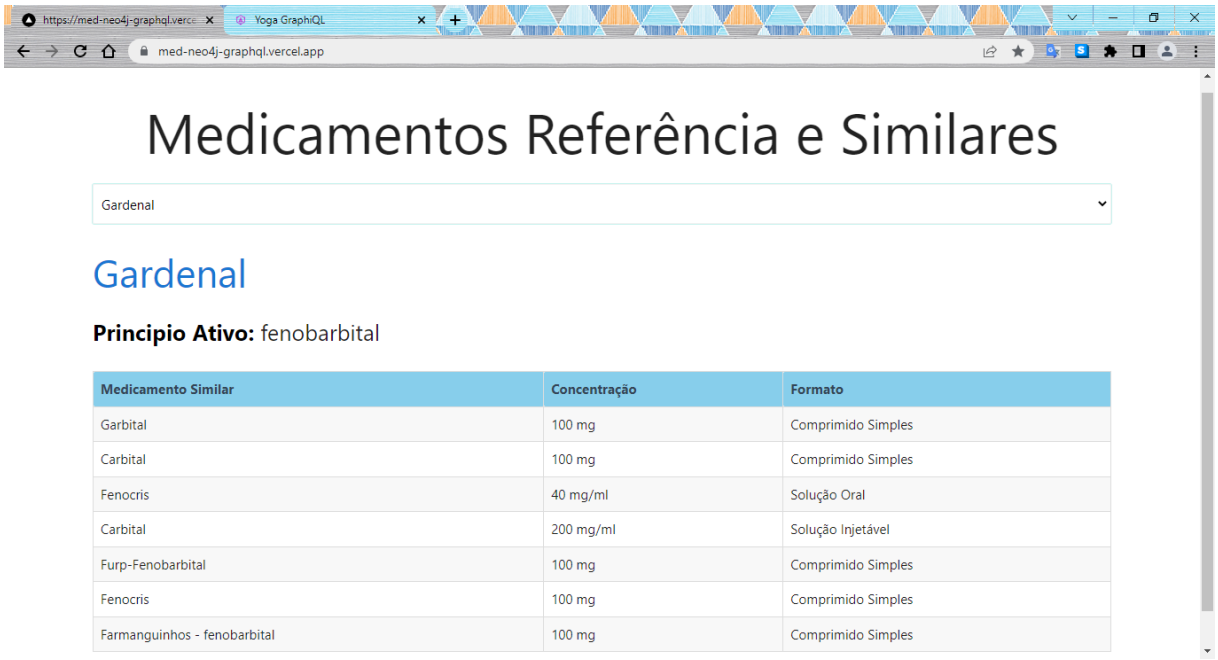
A figura 35 acima mostra uma IU de um app de recuperação de informação, mais especialmente um sistema de recomendação de medicamentos similares e princípio ativo. Esse aplicativo foi desenvolvido juntamente com a API anteriormente descrita.

Onde está aplicação faz uso (via consultas da API) dos dados de medicamentos armazenados no já referido banco de dados gráfico. Contendo todos os medicamentos de referência, seus similares e princípio ativo, bem como, contempla todas as informações que consta da lista de medicamentos disponibilizada pela Anvisa.

Ainda no exemplo acima na figura 35, o usuário consulta o medicamento de referência denominado: Benicaranlo, o sistema retorna o princípio ativo e o nome de todos os medicamentos similares, bem como, sua concentração e formato.

A figura 36 abaixo mostra outro exemplo de recomendação agora com o medicamento referência Gardenal.

Figura 36: GraphQL, outra visualização de consulta e recomendação.



Fonte: Elaborado pela autora (2022).

Esta aplicação pode ser incorporada ao sistema de prontuário eletrônico do HCM e estabelecer um sistema de recomendação de medicamentos similares atendendo a uma demanda especificada pelo grupo de TI do referido hospital.

8. DISCUSSÕES E ARGUMENTAÇÃO

Como já apontado, os Grafos de Conhecimento desempenham um papel fundamental no cenário atual de estruturação e representação da informação, e de desenvolvimento de Inteligência Artificial e Processamento de Linguagem Natural. Ao capturar entidades e as relações semânticas entre entidades de maneira intuitiva e flexível, eles oferecem uma abordagem poderosa e eficiente para modelar, integrar e explorar informações complexas. Neste capítulo, é discutido o porquê optou-se pelo uso de Grafos de Conhecimento neste trabalho em específico, em detrimento de outras formas de representação de dados. Destaca-se as vantagens e benefícios dessa abordagem para esta pesquisa, enfatizando seus pontos fortes que tornaram os grafos uma escolha eficiente para resolver a problemática desta.

8.1 Problemas Abordados e o Uso de Grafos de Conhecimento

O problema abordado neste trabalho de pesquisa foi a recuperação de informações de medicamentos compostas por textos não estruturados e semiestruturados. Posto que, os métodos tradicionais lutam para extrair dados estruturados de texto não estruturado, exigindo esforços manuais e possibilitando erros e atrasos. Sendo necessária uma abordagem automatizada e confiável para extrair entidades relevantes de medicamentos e seus relacionamentos.

Assim, a pesquisa abordou o problema da recuperação de informação de medicamentos, especificamente, os relacionados a recuperação e visualização de informações medicamentosas, como: nome de medicamentos, princípio ativo, medicamentos referência e seus similares, concentração, forma do medicamento, dosagem e tempo de administração, entre outros.

Para a solução desse problema optou-se como visto, por explorar a aplicação de técnicas de Processamento de Linguagem Natural (especificamente, a tarefa de NER) para extrair entidades de medicamentos de textos não estruturados de receituários. Bem como, optou-se, por questões de aplicabilidade, pelo uso de Grafos de Conhecimento para representar as entidades extraídas e seus relacionamentos. E, empregou-se técnicas de agregação de grafos para aprimorar a representação e recuperação de informações de medicamentos.

Isso visando contribuir com o desenvolvimento de um sistema de recomendação de medicamentos similares que pode ser empregado para facilitar a recomendação de medicamentos similares e/ou princípio ativo, em detrimento de medicamentos de marca, em sistemas de prontuários eletrônicos, especificamente, no âmbito do HCM.

O desenvolvimento desse sistema de recomendação de medicamentos similares está sendo idealizado pela equipe de TI do HCM, especificamente pelo grupo de estudos HAIS,

pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília. Bem como, pelo grupo de estudos GIHC, componente da linha de pesquisa em “Informação e Tecnologia” da UNESP Marília. A ideia por trás do sistema é aproveitar o conhecimento contido em um banco de dados gráfico para identificar medicamentos que compartilham o mesmo princípio ativo e, assim, podem ser considerados alternativas mais econômicas para um determinado paciente.

Concluiu-se que para o desenvolvimento do referido sistema, o uso de grafos é uma abordagem superior para a recuperação e representação de informações sobre medicamentos em relação a outras abordagens, como RDF, bancos de dados relacionais e tabelas. Isso ocorre porque os grafos de conhecimento fornecem uma maneira mais flexível e intuitiva de modelar relacionamentos complexos entre diferentes entidades no domínio de medicamentos. Conforme Tian (2022) de modo geral, os modelos de grafo de conhecimento são capazes de representar as mesmas informações utilizando menos nós e arestas em comparação aos modelos RDFs.

Uma das principais vantagens do uso de grafos de conhecimento é a capacidade de representar relacionamentos complexos entre entidades. No domínio dos medicamentos, existem muitas entidades inter-relacionadas, como medicamentos, princípios ativos, medicamentos de referência, concentrações e formas. Um grafo de conhecimento pode representar essas entidades e seus relacionamentos de forma abrangente e intuitiva, facilitando a compreensão dos relacionamentos, e a navegação entre elas.

Além disso, a representação em estrutura de grafos permite a integração e/ou agregação de dados de várias fontes, o que é importante no domínio dos medicamentos, onde as informações costumam estar espalhadas por diferentes bancos de dados e sistemas. Ao integrar esses dados em um único grafo de conhecimento, fica mais fácil recuperar e analisar informações sobre medicamentos sem a necessidade de pesquisar em várias fontes.

Outra vantagem de usar grafos é a sua capacidade de lidar com ambiguidade e incerteza nos dados. No domínio dos medicamentos, invariavelmente, há nomes diferentes para o mesmo medicamento ou medicamentos similares com nomes comerciais diferentes. Os grafos de conhecimento podem lidar facilmente com essas situações, representando os diferentes nomes como nós separados e vinculando-os à mesma entidade de medicamento. Assim, essa capacidade de vincular e representar relações entre diferentes informações sobre medicamentos. Por exemplo, conectando um medicamento específico com seu princípio ativo, similares, concentração, etc. Fornece uma compreensão mais completa e contextualizada do medicamento, permitindo a recuperação de informações altamente relacionadas.

Em contraste, outras abordagens como bancos de dados relacionais e tabelas que têm limitações na modelagem de relacionamentos complexos entre entidades. Já que, essas

abordagens requerem um esquema predefinido e não permitem facilmente a integração de dados de várias fontes. Eles também lutam para lidar com a ambiguidade e a incerteza nos dados, o que pode levar a erros na recuperação de informações sobre medicamentos. Como aponta Reina *et al*. (2021) e Aldwairi (2022), ao contrário de uma solução de banco de dados relacional que exige um esquema predefinido, um sistema de banco de dados baseado em grafos oferece flexibilidade de esquema. Isso significa que um elemento de dados não precisa estar presente em todas as entidades, sendo uma opção viável para armazenar informações com uma estrutura de dados indefinida e que se expande e evolui ao longo do tempo.

Desta forma, ao longo da pesquisa observou-se que o uso de grafos de conhecimento, de modo geral, e, especificamente no problema enfrentado nesta pesquisa, se mostrou uma abordagem superior para a recuperação e representação de informações sobre medicamentos e relacionamentos. Portanto, os grafos de conhecimento forneceram uma maneira mais flexível e intuitiva de modelar relacionamentos entre entidades de medicamentos, permitindo a integração de dados de fontes diferentes, e lidando de modo ótimo com ambiguidade e incerteza nos dados.

Em última análise, a agregação de grafos também pode ser usada para integrar dados médicos (como receituários, relatórios, etc.) derivados de diferentes fontes e agrupar conceitos médicos relacionados em agrupamentos. Isso pode, por fim, ajudar a informar a tomada de decisão clínica com o melhoramento da análise e recuperação de informações.

Assim, entende-se que a combinação do processamento de linguagem natural, grafos de conhecimento e agregação de grafos é uma abordagem eficiente e promissora para superar os desafios para a recuperação de informações diversas sobre medicamentos e outras.

8.2 Aumento de Dados e Escalabilidade

Os bancos de dados gráficos fornecem uma base escalável para armazenar e consultar gráficos de conhecimento. Eles utilizam modelos de dados baseados em gráficos e técnicas de escalabilidade horizontal e computação distribuída para lidar com conjuntos de dados de grande escala de forma eficaz. De acordo com Hall (2022), a escalabilidade horizontal refere-se à capacidade de aumentar a capacidade operacional por meio da adição de servidores adicionais. A distribuição, por sua vez, é a capacidade de espalhar um banco de dados gráfico por vários servidores. Ou seja, um banco de dados gráfico pode distribuir seus dados, e é capaz de conectar dois vértices (entidades) com uma borda, mesmo que esses vértices estejam armazenados em servidores diferentes. Existem vários bancos de dados gráficos que suportam a distribuição e a escalabilidade horizontal, sendo o Neo4j um exemplo desses bancos de dados.

Logo, ao armazenar dados em um formato de grafos de conhecimento, esses bancos de dados facilitam a passagem eficiente e a exploração de relacionamentos, permitindo consultas complexas em grandes quantidades de dados interconectados. Como resultado, analistas e pesquisadores podem descobrir informações valiosas, identificar padrões e tomar decisões baseadas em dados, mesmo diante de enormes conjuntos de dados. E desenvolvedores de sistemas podem criar aplicações para indexação de informações de modo mais eficiente e fácil.

O uso de grafos de conhecimento armazenados em bancos de dados gráficos aumenta a escalabilidade na representação e exploração gráfica. Ao estruturar e organizar os dados em um formato de grafo, fica mais fácil dimensionar a representação para acomodar tamanhos de dados crescentes. Bancos de dados gráficos como o Neo4j, conforme Lyon (2022), são projetados para lidar com grafos de conhecimento em grande escala, e, fornecem a infraestrutura necessária para armazenar e consultar gráficos de conhecimento com eficiência, permitindo a exploração perfeita de relacionamentos e conexões complexas e, escalabilidade de dados e processamento.

Os bancos de dados gráficos podem ser escalados para lidar com grandes volumes de informações de maneira rápida e eficiente. Em termos práticos, a equipe de TI do HCM pode usar o banco de dados gráfico para armazenar informações estruturadas em grafos de conhecimento, sobre medicamentos e pacientes. Usando esse banco de dados, eles podem facilmente responder a perguntas como: quem são os pacientes e quais medicamentos são mais usados, quais são suas interações, contra indicações e reações adversas, entre outras.

Com a possibilidade de escalabilidade oferecida pelos bancos de dados gráficos atuais, a expansão do grafo de conhecimento de medicamentos pode se dar de modo indefinido, e incluir outras ligações como: categoria terapêutica, efeito colateral, interação medicamentosa, doenças tratadas, sintomas, bem como, informações sobre profissionais de saúde, especialidades médicas entre outras. Ajudando a fornecer uma visão mais abrangente do cenário da saúde com a adição de mais elementos, como diretrizes de tratamento, informações demográficas e outros, sempre criando relações diferentes entre eles.

9 CONSIDERAÇÕES FINAIS

Esta pesquisa percorreu uma série ordenada de fases que de modo gradual contribuíram para o embasamento teórico e técnico, que, por conseguinte, formaram a base aplicada de conhecimento utilizado para o desenvolvimento prático desta pesquisa.

A primeira fase desta pesquisa é descrita no primeiro capítulo (introdução) e, indica o caminho metodológico percorrido nesta. Desde a definição e justificativa da proposta de pesquisa, os objetivos gerais e específicos, e, metodologia aplicada. Entende-se que esta foi uma fase definidora que orientou todo o projeto contribuindo para que os objetivos fossem alcançados. Especialmente a construção do corpus documental básico que descortinou algumas das principais tendências de pesquisas voltadas para a recuperação da informação de PEPs.

Ainda na primeira fase da pesquisa procedeu-se com o aprofundamento sistemático da literatura e, como foi apresentado no capítulo 3 foram realizados resumos destacando os pontos e resultados principais de cada um dos artigos e pesquisas selecionadas e analisadas para o embasamento e orientação teórica desta pesquisa. Apresentou-se também as conclusões obtidas da análise do corpus documental. Especialmente, classificando e/ou associando tais pesquisas a uma ou mais contribuições para os esforços de pesquisa, que visam suprir as demandas informacionais surgidas com a implantação e uso do prontuário eletrônico do paciente.

Os capítulos 2, 3, 4 e 5 foram elaborados, respectivamente, com o objetivo de promover uma visão geral e teórica do prontuário eletrônico enquanto documento médico-social, exclusivamente em seus aspectos informacionais. Bem como, pretendeu-se fazer uma breve introdução aos principais conceitos e prerrogativas da Ciência da Informação destacando sua natureza inter-colaborativa e analisando a vertente da CI que investiga os sistemas de organização do conhecimento, e inferir as contribuições desta para sanar as demandas informacionais apresentadas nos sistemas de prontuários eletrônicos. Por fim, foram descritos os conceitos principais da recuperação, representação e visualização da informação e do conhecimento. Destacando os aspectos principais dos Grafos de Conhecimento e PLN.

Já os capítulos 6, 7 e 8 apresentam de forma linear as etapas e/ou passos seguidos para a construção de visualizações de informações de medicamentos e os resultados obtidos. Isso com o objetivo de demonstrar um modelo possível para a extração de informações contidas em anotações livres em prontuários eletrônicos, mais especificamente, anotações referentes a medicamentos. O modelo também abrange ao uso de tarefas de Processamento de Linguagem Natural e Grafos de Conhecimento. Bem como, de outras ferramentas usadas para a

persistência, recuperação e uso da informação e do conhecimento, e sua possível aplicação na recuperação da informação contida nesses sistemas de prontuários eletrônicos.

Conforme informações e resultados apresentados nos capítulos 6, 7 e 8, é possível inferir que a aplicação de métodos, técnicas e ferramentas de PLN aliada ao uso de estruturas de Grafos de Conhecimento pode melhorar a recuperação da informação em sistemas de prontuários eletrônicos. Em especial na extração de informações de anotações médicas não estruturadas. Finalmente neste último capítulo são apresentadas as conclusões e contribuições da pesquisa.

9.1 Conclusões

A presente pesquisa buscou investigar o domínio da recuperação da informação nos prontuários eletrônicos, especificamente, a recuperação e estruturação dos dados textuais de medicamentos. Observa-se, que os atuais recursos e ferramentas computacionais de processamento de linguagem natural e de estruturação em grafos de conhecimento tem demonstrado resultados extremamente positivos sendo objeto de pesquisas e de estudos em diversos campos, em especial na área da saúde. Assim, são desenvolvidas aplicações voltadas para dados gerados e armazenados em prontuários eletrônicos do paciente nas mais diversas áreas, como: clínica geral, psiquiatria, psicologia e outros campos do conhecimento ligados à saúde, como a área de atendimentos primários, farmacêutica e/ou de medicamentos.

Entretanto, mesmo com o crescente interesse na pesquisa voltada para o melhoramento de sistemas e abordagem de recuperação e visualização desses valiosos dados, observou-se uma evidente necessidade de aprimoramento constante para a recuperação e integração de informações atualmente bloqueadas nesses sistemas de prontuários digitais, e assim, evidenciou-se a importância destes dados. Deste modo, com o aprofundamento da pesquisa bibliográfica, e com o estudo das técnicas e tecnologias abordadas nesta pesquisa, foi possível concluir que, especialmente, devido ao caráter interdisciplinar e abrangente do campo da Ciência da Informação (que tem como foco processos e fluxos essencialmente voltados para a organização, gestão e recuperação da informação) esta tem importantes contribuições a fazer, aos PEPs.

Assim, vê-se claramente inúmeras possibilidades de contribuições efetivamente necessárias, relacionadas à área CI e da saúde, perante o prontuário eletrônico do paciente. Nesse contexto, e conforme afirma Sant’Ana (2016), os pesquisadores da Ciência da Informação têm a possibilidade e o dever de contribuir em ambientes que contam com a

presença de acesso e uso intensivo de dados, buscando elementos que possibilitam a construção de estruturas de referências que permitam identificar características em contextos específicos.

Portanto, diante da interdisciplinaridade e, da natureza e tradição colaborativa da Ciência da Informação, há contribuições necessárias a serem feitas pela CI por meio de pesquisas voltadas às resoluções de tarefas de processamento de linguagem natural e de grafos de conhecimento, e da efetiva aplicação destas na recuperação e visualização da informação advindas de prontuários eletrônicos. Sendo que os principais aspectos de pesquisa a serem abordados são os relacionados ao fluxo de trabalho de coleta e recuperação da informação, tendo como foco o aprimoramento constante desta imprescindível ferramenta médico/social.

9.2 Contribuições

Por fim, entende-se que esta pesquisa traz além do próprio valor intrínseco da pesquisa, contribuições para a equipe de TI do HCM, especificamente ao grupo de estudos HAIS, pertencente ao Hospital das Clínicas da Faculdade de Medicina de Marília. Bem como, ao grupo de estudos GIHC, componente da linha de pesquisa em “Informação e Tecnologia” da UNESP Marília. Isso por meio do modelo desenvolvido, que apresenta métodos para o treinamento de modelos de reconhecimento de entidade nomeada de medicamentos e uma base de conhecimento médico estruturada em grafos de conhecimento e armazenada em um banco de dados gráficos. O que possibilita a construção de aplicações de recuperação da informação de medicamentos descritos em prontuários eletrônicos do paciente.

A pesquisa também contribui:

- Com a pesquisa do uso de técnicas atuais de reconhecimento e extração de entidades nomeadas e seus relacionamentos;
- Com a demonstração do uso de estrutura de grafos de conhecimento aplicados a informações de medicamentos;
- Com a metodologia de pesquisa voltada para o melhoramento dos processos de recuperação da informação contidas em sistemas de prontuários eletrônicos.

Posto isso, conclui-se que o estudo realizado evidencia a importância atual e futura dos prontuários eletrônicos, e a pertinência de estudos contínuos visando os aprimoramentos desta ferramenta. Em especial estudos voltados para a recuperação e usos do conhecimento contido nesses documentos fundamentais para a melhoria e otimização do sistema de saúde pública.

REFERÊNCIAS

ALDWAIRI, Monther; JARRAH, Moath; MAHASNEH, Naseem; Graph-based data management system for efficient information storage, retrieval and processing; Information Processing & Management, Vol. 60, 2023.

ANSI/NISO z39.19-2005 (r2010); Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies; p. 17, 2010.

ANTONIOU, Grigoris; HARMELEN, Frank; Web Ontology Language: OWL; p. 1-27, 2013.

ALVARENGA, Lídia; Representação do Conhecimento na Perspectiva da Ciência da Informação em Tempo e Espaço Digitais; Universidade Federal de Santa Catarina, Florianópolis, Brasil, 2003.

ÁVILA, Bráulio Coelho; Representação do Conhecimento Usando Frame; USP- São Carlos, 1991.

AVILA, C. V. S., Rolim, T. V., da Silva, J. W. F., & Vidal, V. M. P. (2019, June). MediBot: Um chatbot para consulta de riscos e informações sobre medicamentos. In Anais Estendidos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde (pp. 1-6). SBC.

BARRASA, Jesús; HODLER, Amy E.; WEBBER, Jim; Knowledge Graphs: Data in Context for Responsive Businesses; editora O'Reilly Media; 2021.

BENÍCIO, Diego Henrique Pegado. Aplicação de mineração de texto e processamento de linguagem natural em prontuários eletrônicos de pacientes para extração e transformação de texto em dados-estruturados. 2020.

BERMAN AN, BIERY DW, GINDER C, HULME OL, MARCUSA D, LEIVA O, Wu WY, CARDIN N, Hainer J, Bhatt DL, Di Carli MF, Turchin A, Blankstein R. Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project. Clin Cardiol. 2021.

BERNERS-LEE T, FIELDING R, MASINTER L (2005) Uniform resource identifier (URI): generic syntax, disponível em http://tools.ietf.org/html/rfc3986. Acesso em: 10 jun. 2022.
BRASIL. Resolução CFM nº 1638, de 10 de julho de 2002. Diário Oficial da União nº 153, seção I, 09/08/2002, p. 184-5, 2022.

BISCALCHIN, Ricardo; Os Sistemas de Organização do Conhecimento e os Desafios Frente a Geração Google; p. 1-7, 2019.

BORKO, H. Information Science: What is it?; p.3-5, Jan. 1968; Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-01---information-science---what-is-it.pdf>. Acesso em 02/11/2022.

BRASIL. Resolução CFM nº 1638, de 10 de julho de 2002. Diário Oficial da União nº 153, seção I, 09/08/2002, p. 184-5, 2022.

BUFREM, Leilah Santiago; Configurações da pesquisa em ciência da informação; p. 1-13, 2013.

CAMPESATO. Oswald; Natural Language Processing and Machine Learning for Developers; editora. Mercury Learning and Information; 2021.

CARLAN, E. Sistemas de organização do conhecimento: uma reflexão no contexto da Ciência da Informação, 2010.

CARVALHO, Ricardo César; Aplicação de mineração de dados em informações oriundas de prontuários de paciente. Informação em Pauta, Fortaleza, v. 3, p. 161-181, 2018.

CORRÊA, Fábio; LACERDA, Mariana Emery de; ZIVIANI, Fabrício; FRANÇA, Renata de Souza; RIBEIRO, Jurema Suely de Araújo Nery; Tecnologias de apoio a Gestão do Conhecimento: uma abstração por conceito, taxonomia e tipologia; p. 1-25, 2018.

CHASE, Herbert S. *et al*. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC medical informatics and decision making, v.17, n. 1, p. 1-8, 2017.

CHEN, Y., Ma, T., Yang, X., Wang, J., Song, B., & Zeng, X. (2021). MUFFIN: multi-scale feature fusion for drug–drug interaction prediction. Bioinformatics, 37(17), 2651-2658.

DORILEO, E. A.; PONCIANO, M.; COSTA, T. Estruturação da evolução clínica para o prontuário eletrônico do paciente. In: CBIS-X Congresso Brasileiro de Informática em Saúde, Florianópolis. 2006. Disponível em: https://www.researchgate.net/profile/Joaquim-Felipe/publication/238075164_Estruturacao_da_Evolucao_Clinica_para_o_Prontuario_Eletronico_do_Paciente/links/548985850cf2ef3447929a03/Estruturacao-da-Evolucao-Clinica-para-o-Prontuario-Eletronico-do-Paciente.pdf. Acesso em: 10 jul. 2022.

FENSEL, Dieter; ŞIMŞEK, Umutcan; ANGELE, Kevin; HUAMAN, Elwin; KÄRLE, Elias; PANASIUK, Oleksandra; TOMA, Ioan; UMBRICH, Jürgen; WAHLER, Alexander; Knowledge Graphs Methodology, Tools and Selected Use Cases; editora. Springer Nature Switzerland, pag 2-6, 2020.

FUJITA, Mariângela Spotti Lopes; GUIM, Vera Lucia Ribeiro; As linguagens de indexação e a análise de domínio; v.3, p. 1-10, 2015.

GELETA, David; NIKOLOV, Andriy; EDWARDS, Gavin; GOGLEVA, Anna; JACKSON, Richard; JANSSON, Erik; LAMOV, Andrej; NILSSON, Sebastian; PETTERSSON, Marina; POROSHIN, Vladimir; ROZEMBERCZKI, Benedek; SCRIVENER, Timothy; UGHETTO, Michael; PAPA, Eliseo; Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development; p. 1-24, 2021.

GALVÃO, M. C. B. *et al*. Linguagens empregadas em prontuários do paciente frente aos processos de organização e recuperação da informação no contexto da saúde. 2008. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/804. Acesso em: 10 jul. 2022.

GALVÃO, M. C. B.; RICARTE, I. L. M. Prontuário do paciente. Rio de Janeiro: Guanabara Koogan, 2012.

GAVRILOVA, Tatiana A.; LESHCHEVA, Irina A.; Building Collaborative Ontologies: A Human Factors Approach; p. 1-23, 2015.

GHOLAMY, Afshin; KREINOVICH, Vladik; KOSHELEVA, Olga; Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation; p. 1-7, 2018

GUNKEL, David J; Comunicação e inteligência artificial: novos desafios e oportunidades para a pesquisa em comunicação; São Paulo, n. 34, p. 05-19, 2017.

HALL, Dan; Graph Database Scalability; 2022.

HEFLIN, Jeff; an introduction to the owl web ontology language;A Semantic Web Primer. MIT Press, Cambridge, MA, p. 1-24, 2022. Disponível em: <http://www.cse.lehigh.edu/~heflin/IntroToOWL.pdf>; Disponível em: Acesso em: 06 ago. 2022.

JACKSON, Richard G. *et al*. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ open, v. 7, n. 1, p. 12, 2017.

JAGANNATHA, Abhyuday *et al*. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug safety, v. 42, n. 1, p. 99-111, 2019.

JAKUS, Grega; MILUTINOVIC, Veljko; OMEROVIC, Sanida; TOMAZIC, Sašo; Concepts, Ontologies, and Knowledge Representation; editora Springer New York Heidelberg Dordrecht London, 2013., p. 1-71, 2013.

JI, Shaoxiong; PAN, Shirui; CAMBRIA, Erik; MARTTINEN, Pekka; YU, Philip S. A Survey on Knowledge Graphs: Representation, Acquisition and Applications; p. 1-27, 2021.

JOUFFROY, J., Feldman, S. F., Lerner, I., Rance, B., Burgun, A., & Neuraz, A. (2021). Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR medical informatics, 9(3), e17934.

JUHN, Young; LIU, Hongfang. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. Journal of Allergy and Clinical Immunology, v. 145, n. 2, p. 463-469, 2020.

KOLECK, Theresa 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, v. 26, n. 4, p. 364-379, 2019.

KORMILITZIN, Andrey *et al*. Med7: a transferable clinical natural language processing model for electronic health records. Artificial Intelligence in Medicine, v. 118, p. 102086, 2021.

KURBATOVA, Natalja; SWIERS, Rowan; Disease ontologies for knowledge graphs; BMC Bioinformatics, p. 1-7, 2021.

LI, Deng; YANG, Liu; Deep Learning in Natural Language Processing; editora. Springer, 2018.

LI, Linfeng *et al*. Real-world data medical knowledge graph: construction and applications. Artificial intelligence in medicine, v. 103, p. 101817, 2020.

LIMA, Gercina Ângela de. Organização e representação do conhecimento e da informação na web: teorias e técnicas; p57-97, 2020.

LIMA, Júnio César de; CARVALHO, Cedric Luiz de; Resource Description Framework (RDF); p. 1-28, 2005.

LIN, X., Quan, Z., Wang, Z. J., Ma, T., & Zeng, X. (2020, July). KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction. In IJCAI (Vol. 380, pp. 2739-2745).

LIU, Feifan; WENG, Chunhua; YU, Hong. Natural language processing, electronic health records, and clinical research. In: Clinical Research Informatics. Springer, p. 293-310, London, 2012.

LIU, Zhiyuan; HAN, Xianpei; Deep Learning in Natural Language Processing - Deep Learning in Knowledge Graph; published by: Springer Nature Singapore Pte Ltd; p. 117-145; 2018.

LYON, William; Fullstack GraphQL Applications With React, Node.js, and Neo4j; p.51-75 2022.

LOPES, Dener Cesar Ferreira; Grafos de Conhecimento: perspectivas e desafios para a organização e representação do conhecimento; p. 28-29 2020.

MARTHA, A. S.; BARRA, P. S. C.; CAMPOS, C.J.R. Recuperação de Informações em Textos Livres de Prontuários do Paciente. In: CBIS-IX Congresso Brasileiro de Informática em Saúde. 2004. Disponível em: http://telemedicina.unifesp.br/pub/sbis/cbis2004/trabalhos/arquivos/636.pdf. Acesso em: 15 jul. 2022.

MENDES, Paula Raphisa; REIS, Raquel Martins dos; MACULAN, Benildes Coura Moreira dos Santos; Tesauros no Acesso a Informação: Uma Retrospecção; p. 49-66, 2015.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula; Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens; Instituto de Informática Universidade Federal de Goiás, p. 1-21, 2007.

MOREIRA, W.; Sistemas de organização do conhecimento: aspectos teóricos, conceituais e metodológicos; p.101-103, 2018.

NASCIMENTO, F. M. S. Uso Estratégico da Ontologia para Organização e Gestão da Informação Jurídica; 2018.

NASCIMENTO, Felipe Mozart de Santana; PINHO, Fábio Assis; Sistemas de organização do conhecimento: semelhanças e diferenças; p. 1-20, 2020

NEO4J.COM; Visão geral do Neo4j; Disponível em: <https://neo4j.com/graphacademy/training-overview-40/02-overview40-neo4j-graph-platform/>. Acesso em: 08 set. 2022.

NEUMAMM, Fernanda Bugarin de Andrade; Camuzi, Ranieri Carvalho; Cordeiro, Benedito Carlos; Implementação da prescrição eletrônica em um hospital público municipal; p. 1-9, v. 12, n. 1, 2023.

NICHOLSON, David N; GREENE, Casey S; Constructing knowledge graphs and their biomedical applications; editora Elsevier B.V, p.1414-1428, 2020.

OLIVEIRA, A. M. de.; Filipin, M. D. V.; Reis, D. A.; Monteiro, L. M. M.; Cubayachi, C.; Pharmaceutical services strengthen the brazilian Unified Health System: experience of the practice of technical and clinical activities that result in safety; 2022.

OLIVEIRA, Hellen Carmo de; CARVALHO, Cedric Luiz de; Gestão e Representação do Conhecimento; p. 1-20, Goiás, 2008.

OLIVEIRA, Suellen de Alcântara; FAVARETTO, Fábio. Qualidade da Informação do Prontuário Eletrônico do Paciente no Processo de Apoio à Decisão Clínica. Journal of Health Informatics, v. 13, n. 1, 2021. Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/767. Acesso em: 06 jun. 2021.

PATEL, Ankur A; ARASANIPALAI, Ajay Uppili; Applied Natural Language Processing in the Enterprise; editora O'Reilly, 2021.

PETASIS, Georgios; KARKALETSIS, Vangelis; PALIOURAS, Georgios; KRITHARA, Anastasia; ZAVITSANOS, Elias; Ontology Population and Enrichment: State of the Art; p. 1-34, 2011.

RASHID, Pshtiwan Qader; Semantic Network and Frame Knowledge Representation Formalisms in Artificial Intelligence; Eastern Mediterranean University; 2015.

REINA, Santiago; RINCÓN, Mariano, TOMÁS, Rafael; An overview of graph databases and their applications in the biomedical domain; 2021.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. Informação & Informação, Londrina, v. 21, n. 2, p. 116-142, 2016. Disponível em: https://www.readcube.com/articles/10.5433%2F1981-8920.2016v21n2p116. Acesso em: 05 jun. 2022.

SANTOS, Plácida Leopoldina Ventura Amorim da Costa. Catalogação, formas de representação e construções mentais. Tendências da Pesquisa Brasileira em Ciência da Informação; p. 1-24, 2015.

SAVOVA, Guergana K. *et al*. Use of natural language processing to extract clinical câncer phenotypes from electronic medical records. Cancer research, v. 79, n. 21, p. 5463-5470, 2019.

SCHRODT, Jens *et al*. Graph-Representation of Patient Data: a Systematic Literature Review. p. 1-7, 2020.

SERAFIM, J. E. F., Matos, L. E. O., & Unfer, T. C. (2022). Informações proativas emitidas por um centro de informações sobre medicamentos na pandemia de COVID-19 no período de 2020 a 2022. Revista Informação na Sociedade Contemporânea, 6, e28708-e28708.

SHEN, Yi-Cheng; HSIA, Te-Chun; HSU, Ching-Hsien. Analysis of Electronic Health Records Based on Deep Learning with Natural Language Processing. Arabian Journal for Science and Engineering, 2021, 1-11.

SILVA, Daniel N. R.; ZIVIANI, Artur; PORTO, Fábio; Aprendizado de máquina e inferência em Grafos de Conhecimento; p. 93-122, Fortaleza, CE, 2019.

SILVA, Jonathas Luiz Carvalho, GOMES, Henriette Ferreira; Conceitos de Informação na Ciência da Informação: percepções analíticas, proposições e categorizações; p. 145-157, 2015.

SINGHAL, Amit; Introducing the Knowledge Graph: things, not strings; Disponível em: <https://blog.google/products/search/introducing-knowledge-graph-things-not/>. Acesso em: 05 maio. 2022.

SOUZA, Amanda D; DE ALMEIDA, Maurício B. de. Análise de dados clínicos textuais de Prontuários Eletrônicos do Paciente para integração com terminologias médicas padronizadas. Disponível em: https://tinyurl.com/4pec7ark. Acesso em: 12 jul. 2022.

SOUZA, Maria da Paixão Neres de; Abordagem inter e transdisciplinar em ciência da informação; p. 75-90, 2007. Disponível em: <https://repositorio.ufba.br/bitstream/ufba/145/1/Para%20entender%20a%20ciencia%20da%20informacao.pdf>. Acesso em 02/11/2022.

SOUZA, Renato Tarciso Barbosa de.; ARAÚJO, Rogerio Henrique de. Júnior; Estudo do ecossistema de Big Data para conciliação das demandas de acesso, por meio da representação e organização da informação; p.187-198, 2018.

SILVA AP, Santos HD, Rotta AL, Baiocco GG, Vieira R, Urbanetto JS. Risco de queda relacionado a medicamentos em hospitais: abordagem de aprendizado de máquina. Acta Paul Enferm. 2023;36:eAPE00771.

SPACY.IO; SpaCy 101: Everything you need to know; 2022; Disponível em: <https://spacy.io/usage/spacy-101>. Acesso em: 10 set. 2022.

SUN, Bo *et al*. Using NLP in openEHR archetypes retrieval to promote interoperability: a feasibility study in China. BMC Medical Informatics and Decision Making, v. 21, n. 1, p. 1-12, 2021.

TIAN, Yuanyuan; The World of Graph Databases from An Industry Perspective; p. 1-8, 2022.

TURKI, Houcemeddine *et al*. Representing COVID-19 information in collaborative knowledge graphs: the case of Wikidata. Semantic Web, n. Preprint, p. 1-32, 2022.

VALENTIM, Marta Lígia Pomim; *et al*. GESTÃO, MEDIAÇÃO E USO DA INFORMAÇÃO. Editora UNESP/Cultura Acadêmica, p. 144. 2010.

VITAL, Luciane Paula; CAFÉ, Ligia Maria Arruda; Práticas de Elaboração de Taxonomias: análise e recomendações; p. 1-16, 2007.

VOGEL, Michely Jabala Mamede; KOBASHI, Nair Yumiko; Tesauro Funcional para Organização de Arquivos Administrativos; p. 1-15, 2019.

W3C; Semantic Web; Disponível em: <https://www.w3.org/standards/semanticweb/>; Acesso em: 10 jun. 2022.

WANG, Lei *et al*. Construction of a knowledge graph for diabetes complications from expertreviewed clinical evidences, p 29–35, 2020.

WU, Zongsheng; XUE, Ru; SHAO, Meiyun. Knowledge graph analysis and visualization of AI technology applied in COVID-19. Environmental Science and Pollution Research, v. 29, n. 18, p. 26396-26408, 2022.

ZENG, Zexian *et al*. Natural language processing for EHR-based computational phenotyping. IEEE/ACM transactions on computational biology and bioinformatics, v. 16, n. 1, p. 139-153, 2018.

ZIMMERMANN, I. R., de Oliveira, E. F., Vidal, Á. T., Santos, V. C. C., & Petramale, C. A. (2015). A qualidade das evidências e as recomendações sobre a incorporação de medicamentos no Sistema Único de Saúde: uma análise retrospectiva. Revista Eletrônica Gestão e Saúde, (4), 3043-3065.

ZHANG, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., & Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. Journal of biomedical informatics, 115, 103696.

APÊNDICE A: Pipelines do Modelo BERT/RoBerta para reconhecimento e extração de entidades nomeadas e relacionamentos de medicamentos descritos em PEP.

Preparando o ambiente de execução e instalando dependencias.

```
# Importando e mondando Google Drive
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
!pip install -U pip setuptools wheel
# ATENÇÃO! reinicie o tempo de execução após a instalação desta dependência.
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pip in /usr/local/lib/python3.8/dist-packages (22.3.1)
Requirement already satisfied: setuptools in /usr/local/lib/python3.8/dist-packages (65.6.3)
Requirement already satisfied: wheel in /usr/local/lib/python3.8/dist-packages (0.38.4)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager
```

```
# Instalando a Biblioteca Spacy 3.2.1
!pip install -U spacy==3.2.1
!python -m spacy download pt_core_news_lg # Baixando modelo pré-treinada em português, do spacy.
!pip install spacy[transformers]
# Importando biblioteca.
import spacy
```

```
# Conferindo Especificações do ambiente de execução.
!nvidia-smi
!python -m spacy info
```

```
Sun Jan  1 22:43:04 2023
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.03    Driver Version: 460.32.03    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   42C    P0    25W /  70W |      3MiB / 15109MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
/usr/local/lib/python3.8/dist-packages/spacy/util.py:833: UserWarning: [W095] Model 'en_core_web_sm' (3.4.1) was trained with spaCy
  warnings.warn(warn_msg)

============================== Info about spaCy ==============================

spaCy version    3.2.1
Location         /usr/local/lib/python3.8/dist-packages/spacy
Platform         Linux-5.10.133+-x86_64-with-glibc2.27
Python version   3.8.16
Pipelines        pt_core_news_lg (3.2.0), en_core_web_sm (3.4.1)
```

Construindo modelo de Reconhecimento de Entidade Nomeada (NER) Baseado em BERT.

Customizando um modelo BERT (Bidirectional Encoder Representations from Transformers) para NER (Named Entity Recognition - Reconhecimento de Entidade Nomeada) com spaCy, para prever entidades associadas a medicamentos como: Medicamento, Concentração, Dosagem e Tempo de Administração.

```
# Convertendo arquivos de anotações de entidades para treino e teste,
# no formato IOB (Inside, Outside, Beginning) () de .tsv para .json
!python -m spacy convert /content/drive/MyDrive/dados/iob_med_train.tsv ./ -t json -n 1 -c iob
!python -m spacy convert /content/drive/MyDrive/dados/iob_med_test.tsv ./ -t json -n 1 -c iob
```

```
ℹ Auto-detected token-per-line NER format
⚠ Document delimiters found, automatic document segmentation with `-n`
disabled.
✓ Generated output file (1 documents): iob_med_train.json
ℹ Auto-detected token-per-line NER format
⚠ Document delimiters found, automatic document segmentation with `-n`
```

```
disabled.
✓ Generated output file (1 documents): iob_med_test.json
```

```
# Convertendo arquivos de anotações de entidades de .json para .spacy
!python -m spacy convert /content/iob_med_train.json ./ -t spacy
!python -m spacy convert /content/iob_med_test.json ./ -t spacy
```

```
✓ Generated output file (1 documents): iob_med_train.spacy
✓ Generated output file (1 documents): iob_med_test.spacy
```

```
# Baxando arquivo de componentes e configuração de ajustes do modelo.
!python -m spacy init fill-config /content/drive/MyDrive/dados/base_config.cfg ./config_spacy.cfg
```

```
✓ Auto-filled config with all values
✓ Saved config
config_spacy.cfg
You can now add your data and train your pipeline:
python -m spacy train config_spacy.cfg --paths.train ./train.spacy --paths.dev ./dev.spacy
```

Treinado o e salvando o modelo NER

```
# Executar comando de treinamento.
!python -m spacy train -g 0 /content/config_spacy.cfg --output ./output
```

```
ℹ Saving to output directory: output
ℹ Using GPU: 0

=========================== Initializing pipeline ===========================
[2023-01-01 23:00:34,083] [INFO] Set up nlp object from config
INFO:spacy:Set up nlp object from config
[2023-01-01 23:00:34,092] [INFO] Pipeline: ['transformer', 'ner']
INFO:spacy:Pipeline: ['transformer', 'ner']
[2023-01-01 23:00:34,095] [INFO] Created vocabulary
INFO:spacy:Created vocabulary
[2023-01-01 23:00:34,096] [INFO] Finished initializing nlp object
INFO:spacy:Finished initializing nlp object
Some weights of the model checkpoint at neuralmind/bert-base-portuguese-cased were not used when initializing BertModel: ['cls.pre
- This IS expected if you are initializing BertModel from the checkpoint of a model trained on another task or with another archit
- This IS NOT expected if you are initializing BertModel from the checkpoint of a model that you expect to be exactly identical (i
[2023-01-01 23:01:07,189] [INFO] Initialized pipeline components: ['transformer', 'ner']
INFO:spacy:Initialized pipeline components: ['transformer', 'ner']
✓ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['transformer', 'ner']
ℹ Initial learn rate: 0.0
E    #       LOSS TRANS...  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  -------------  --------  ------  ------  ------  ------
  0       0         751.13    963.28    0.00    0.00    0.00    0.00
 12     200       87598.82  73167.13   99.07   98.76   99.38    0.99
 25     400           0.00      0.00   98.88   98.51   99.25    0.99
 37     600           0.00      0.00   98.88   98.51   99.25    0.99

Aborted!
```

Usando o modelo customizado para prever as entidades: *MED (Mdicamentos), CONCEN (Concentração), DOSE (Dosagem) e APRAZ (Tempo ou Intervalo para administração).*

```
import spacy # importando biblioteca spacy.
# importando o modelo com maior pontuação.
nlp = spacy.load("/content/output/model-best")

texto = [
'''O medicamento Abilify de 20 mg foi indicado como terapia para o tratamento agudo de episódios  associados ao transtorno bipolar
2 cp do medicamento Confilify de 35Mg de 5 em 5 horas. No entanto, foi proposto que a eficácia do Sensaz, aripiprazol é mediada po

# iniciando doc e percorrendo texto.
for doc in nlp.pipe(texto, disable=["tagger", "parser"]):
    print([(ent.text, ent.label_) for ent in doc.ents])
```

```
[('Abilify', 'MED'), ('20 mg', 'CONCEN'), ('1 comprinido', 'DOSE'), ('4/4 hs', 'APRAZ'), ('aripiprazol', 'MED'), ('2 cp', 'DOSE'),
```

```
# Visualizando com display do spacy.
spacy.displacy.render(doc, style="ent", jupyter=True)
```

O medicamento Abilify MED de 20 mg CONCEN foi indicado como terapia para o tratamento agudo de episódios associados ao transtorno bipolar do tipo I. Tomar 1 comprinido DOSE de 4/4 hs APRAZ O mecanismo de ação do aripiprazol MED , como ocorre com outras drogas eficazes no tratamento de transtorno bipolar, com

2 cp DOSE do medicamento Confililfy MED de 35Mg CONCEN de 5 em 5 horas APRAZ . No entanto, foi proposto que a eficácia do Sensaz MED , aripiprazol MED é mediada por efeitos no sistema nervoso central. A atividade de Abilify MED principalmente devida à droga inalterada, aripiprazol

Treinando modelo customizado para previsão de Relacionamentos de Entidades

```
# Clonar componentes de treinamento para relacionamento de entidades.
!python -m spacy project clone tutorials/rel_component
```

```
✓ Cloned 'tutorials/rel_component' from explosion/projects
/content/rel_component
✓ Your project is now ready!
To fetch the assets, run:
python -m spacy project assets /content/rel_component
```

```
%cd /content/rel_component #entrar na pasta rel_component.
```

```
!spacy project run train_gpu # comando para treinar transformadores de treino.
!spacy project run evaluate # comando para avaliar no conjunto de dados de teste.
```

```
INFO:spacy:Finished initializing nlp object
Downloading: 100% 481/481 [00:00<00:00, 788kB/s]
Downloading: 100% 878k/878k [00:00<00:00, 5.01MB/s]
Downloading: 100% 446k/446k [00:00<00:00, 3.91MB/s]
Downloading: 100% 1.29M/1.29M [00:00<00:00, 8.13MB/s]
Downloading: 100% 478M/478M [00:10<00:00, 49.9MB/s]
Some weights of the model checkpoint at roberta-base were not used when initializing RobertaModel: ['lm_head.bias', 'lm_head.den
- This IS expected if you are initializing RobertaModel from the checkpoint of a model trained on another task or with another a
- This IS NOT expected if you are initializing RobertaModel from the checkpoint of a model that you expect to be exactly identic
[2023-01-01 23:53:05,121] [INFO] Initialized pipeline components: ['transformer', 'relation_extractor']
INFO:spacy:Initialized pipeline components: ['transformer', 'relation_extractor']
✓ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['transformer', 'relation_extractor']
ℹ Initial learn rate: 0.0
E    #       LOSS TRANS...  LOSS RELAT...  REL_MICRO_P  REL_MICRO_R  REL_MICRO_F  SCORE
---  ------  -------------  -------------  -----------  -----------  -----------  ------
  0       0           0.80           1.25         7.26        99.57        13.54    0.14
  5     100          36.04          27.23        87.36        88.60        87.98    0.88
 10     200           0.08           0.48        91.16        91.03        91.09    0.91
 15     300           0.01           0.12        91.40        90.88        91.14    0.91

Aborted!

================================== evaluate ==================================
Running command: /usr/bin/python3 ./scripts/evaluate.py training/model-best /content/rel_component/data/test.spacy False

Random baseline:
threshold 0.00   {'rel_micro_p': '7.25', 'rel_micro_r': '100.00', 'rel_micro_f': '13.52'}
threshold 0.05   {'rel_micro_p': '7.15', 'rel_micro_r': '93.30', 'rel_micro_f': '13.28'}
threshold 0.10   {'rel_micro_p': '7.16', 'rel_micro_r': '88.75', 'rel_micro_f': '13.26'}
threshold 0.20   {'rel_micro_p': '7.34', 'rel_micro_r': '80.77', 'rel_micro_f': '13.46'}
threshold 0.30   {'rel_micro_p': '7.33', 'rel_micro_r': '70.66', 'rel_micro_f': '13.28'}
threshold 0.40   {'rel_micro_p': '7.47', 'rel_micro_r': '62.11', 'rel_micro_f': '13.34'}
threshold 0.50   {'rel_micro_p': '7.46', 'rel_micro_r': '51.85', 'rel_micro_f': '13.04'}
threshold 0.60   {'rel_micro_p': '7.43', 'rel_micro_r': '41.31', 'rel_micro_f': '12.60'}
threshold 0.70   {'rel_micro_p': '7.53', 'rel_micro_r': '31.48', 'rel_micro_f': '12.15'}
threshold 0.80   {'rel_micro_p': '7.20', 'rel_micro_r': '20.09', 'rel_micro_f': '10.60'}
threshold 0.90   {'rel_micro_p': '7.49', 'rel_micro_r': '10.40', 'rel_micro_f': '8.71'}
threshold 0.99   {'rel_micro_p': '5.94', 'rel_micro_r': '0.85', 'rel_micro_f': '1.49'}
threshold 1.00   {'rel_micro_p': '11.11', 'rel_micro_r': '0.14', 'rel_micro_f': '0.28'}

Results of the trained model:
threshold 0.00   {'rel_micro_p': '7.25', 'rel_micro_r': '100.00', 'rel_micro_f': '13.52'}
threshold 0.05   {'rel_micro_p': '90.77', 'rel_micro_r': '91.03', 'rel_micro_f': '90.90'}
threshold 0.10   {'rel_micro_p': '91.16', 'rel_micro_r': '91.03', 'rel_micro_f': '91.09'}
threshold 0.20   {'rel_micro_p': '91.14', 'rel_micro_r': '90.88', 'rel_micro_f': '91.01'}
threshold 0.30   {'rel_micro_p': '91.14', 'rel_micro_r': '90.88', 'rel_micro_f': '91.01'}
threshold 0.40   {'rel_micro_p': '91.14', 'rel_micro_r': '90.88', 'rel_micro_f': '91.01'}
threshold 0.50   {'rel_micro_p': '91.40', 'rel_micro_r': '90.88', 'rel_micro_f': '91.14'}
threshold 0.60   {'rel_micro_p': '91.93', 'rel_micro_r': '90.88', 'rel_micro_f': '91.40'}
threshold 0.70   {'rel_micro_p': '93.27', 'rel_micro_r': '90.88', 'rel_micro_f': '92.06'}
threshold 0.80   {'rel_micro_p': '96.08', 'rel_micro_r': '90.74', 'rel_micro_f': '93.33'}
threshold 0.90   {'rel_micro_p': '98.76', 'rel_micro_r': '90.74', 'rel_micro_f': '94.58'}
threshold 0.99   {'rel_micro_p': '99.84', 'rel_micro_r': '88.18', 'rel_micro_f': '93.65'}
threshold 1.00   {'rel_micro_p': '0.00', 'rel_micro_r': '0.00', 'rel_micro_f': '0.00'}
```

```
from pathlib import Path
import random
import typer

import spacy
from spacy.tokens import DocBin, Doc
from spacy.training.example import Example

# Atenção:
# É necessário transferir os arquivos (rel_pipe e rel_model) da pasta: rel_component/scripts, para o diretório raiz,
# para poder acessar scripts necessários para a previsão de relacionamentes de entidades.
from rel_pipe import make_relation_extractor, score_relations
from rel_model import create_relation_model, create_classification_layer, create_instances, create_tensors


# Carregando o modelo de extração de relações de entidades.
nlp2 = spacy.load("/content/drive/MyDrive/rel_component/training/model-best")

# Pegando as entidades geradas pelo modelo NER e inserindo-as no modelo de RE (Extração de Relação).
for name, proc in nlp2.pipeline:
  doc = proc(doc) # é necessário que o modelo de ner tenha sido executado.
# Analisa sentenças do texto  e extrai a relação para cada par de entidades encontradas.
for value, rel_dict in doc._.rel.items():
  for sent in doc.sents:
    for e in sent.ents:
      for b in sent.ents:
        if e.start == value[0] and b.start == value[1]:
          if rel_dict['DOSAGEM'] >=0.9 : # Extrai as relações de DOSAGEM prevista com percentual de probabilidade >= a 90%.
            print(f" entities: {e.text, b.text} --> predicted relation: {rel_dict}")
          if rel_dict['CONCENTRACAO'] >=0.9 : # O mesmo com as relações de CONCENTRACAO.
            print(f" entities: {e.text, b.text} --> predicted relation: {rel_dict}")
          if rel_dict['TEMPO_DE_ADMINISTRACAO'] >=0.9 : # O mesmo com as relações de TEMPO_DE_ADMINISTRACAO.
            print(f" entities: {e.text, b.text} --> predicted relation: {rel_dict}")

      entities: ('aripiprazol', '1 comprinido') --> predicted relation: {'DOSAGEM': 0.9423704, 'CONCENTRACAO': 0.0008167112, 'TEMPO_DE_A
      entities: ('aripiprazol', '2 cp') --> predicted relation: {'DOSAGEM': 0.98170125, 'CONCENTRACAO': 0.0013308827, 'TEMPO_DE_ADMINIS
      entities: ('Confilify', '2 cp') --> predicted relation: {'DOSAGEM': 0.9821806, 'CONCENTRACAO': 0.0014535998, 'TEMPO_DE_ADMINISTRA

import pandas as pd
from spacy import displacy

text = """O medicamento Abilify de 20 mg foi indicado como terapia para o tratamento agudo de episódios  associados ao transtorno bipola
2 cp do medicamento Confilify de 35Mg de 5 em 5 horas. No entanto, foi proposto que a eficácia do Sensaz, aripiprazol é mediada por efei

columns = {}
entities = []
labels = []
sentences=text
for token in doc.ents:
  print(token,token.label_)
  entities.append(token.text)
  labels.append(token.label_)
columns["Sentenças"] =text
columns["Entidades"]= entities
columns["Rótulos"] = labels
displacy.render(doc, style='ent', jupyter=True,)
dataframe = pd.DataFrame(columns)

for value, rel_dict in doc._.rel.items():
  for sent in doc.sents:
    for e in sent.ents:
      for b in sent.ents:
        # Prevendo entidades e relações simutaneamente.
        if e.start == value[0] and b.start == value[1]:
          if rel_dict['DOSAGEM']>=0.70 :
            print(f" entities: {e.text, b.text} --> predicted relation: {DOSAGEM-{rel_dict['DOSAGEM']}} ---> predicted entity: {e.label_
          if rel_dict['CONCENTRACAO']>=0.70 :
            print(f" entities: {e.text, b.text} --> predicted relation:{CONCENTRACAO-{rel_dict['CONCENTRACAO']}} ---> predicted entity:
          if rel_dict['TEMPO_DE_ADMINISTRACAO']>=0.70 :
            print(f" entities: {e.text, b.text} --> predicted relation:{TEMPO_DE_ADMINISTRACAO-{rel_dict['TEMPO_DE_ADMINISTRACAO']}} ---

# Criando um dataframe com a sentença e a previsão ou reconhecimento de entidades.
print("\n")
dataframe.groupby("Sentenças").apply(lambda x : x[:]).drop("Sentenças",axis=1)
```

```
Abilify MED
20 mg CONCEN
1 comprinido DOSE
4/4 hs APRAZ
aripiprazol MED
2 cp DOSE
Confilify MED
35Mg CONCEN
5 em 5 horas APRAZ
Sensaz MED
, aripiprazol MED
Abilify MED
aripiprazol MED
dehidro-aripiprazol. MED
Aripiprazol, Kavium MED
e Aipri MED
do Harip e Toarip MED
```

O medicamento Abilify **MED** de 20 mg **CONCEN** foi indicado como terapia para o tratamento agudo de episódios associados ao transtorno bipolar do tipo I. Tomar 1 comprinido **DOSE** de 4/4 hs **APRAZ** O mecanismo de ação do aripiprazol **MED** , como ocorre com outras drogas eficazes no tratamento de transtorno bipolar, com

2 cp **DOSE** do medicamento Confilify **MED** de 35Mg **CONCEN** de 5 em 5 horas **APRAZ** . No entanto, foi proposto que a eficácia do Sensaz **MED** , aripiprazol **MED** é mediada por efeitos no sistema nervoso central. A atividade de Abilify **MED** principalmente devida à droga inalterada, aripiprazol **MED** , e em menor medida ao seu dehidro-aripiprazol. **MED** sendo o Aripiprazol, Kavium **MED** e Aipri **MED** e do Harip e Toarip **MED** .

Sentenças		Entidades	Rótulos
O medicamento Abilify de 20 mg foi indicado como terapia para o tratamento agudo de episódios associados ao transtorno bipolar do tipo I. Tomar 1 comprinido de 4/4 hs O mecanismo de ação do aripiprazol, como ocorre com outras drogas eficazes no tratamento de transtorno bipolar, com\n2 cp do medicamento Confilify de 35Mg de 5 em 5 horas. No entanto, foi proposto que a eficácia do Sensaz, aripiprazol é mediada por efeitos no sistema nervoso central. A atividade de Abilify principalmente devida à droga inalterada, aripiprazol, e em menor medida ao seu, dehidro-aripiprazol. sendo o Aripiprazol o o Kavium Aipri bem como do Harip e Toarip.	**0**	Abilify	MED
	1	20 mg	CONCEN
	2	1 comprinido	DOSE
	3	4/4 hs	APRAZ
	4	aripiprazol	MED
	5	2 cp	DOSE
	6	Confilify	MED
	7	35Mg	CONCEN
	8	5 em 5 horas	APRAZ
	9	Sensaz	MED
	10	, aripiprazol	MED
	11	Abilify	MED
	12	aripiprazol	MED
	13	dehidro-aripiprazol.	MED
	14	Aripiprazol, Kavium	MED

Grafos de Conhecimento - Modelo de estruturação e visualização de dados

Usando os modelos de Transformers personalizados para Reconhecimento de Entidade Nomeada e Extração de Relacionamento de entidades de medicamentos descritos em prontuários eletrônicos, construir um grafo de conhecimento com/no banco de dados gráficos Neo4J.

```
# !pip install -U pip setuptools wheel


# pip install -U spacy==3.2.1
# python -m spacy download pt_core_news_lg
# pip install spacy[transformers]

# import spacy


# Carregando o melhor modelo de NER de previsão entidades de medicamentos.
nlp = spacy.load("/content/drive/MyDrive/output/model-best")
```

```
import pandas as pd

# Esta função lê um conjunto de dados com 300 sentenças contendo techos de receituário médico.
def get_setencas():
  df = pd.read_csv("/content/drive/MyDrive/dataset_pacientes.csv",sep='"',header=None)
  documentos = []
  for index,row in df.iterrows():
    documentos.append(str(row[0]))
  return documentos

documentos = get_setencas()
corpus = documentos[:]
```

Usando o modelo de NER para extrair entidades do conjunto de dados.

```
import hashlib

# Função para extrair entidades nomeadas: (MED, CONCEN, DOSE e APRAZ).
def extract_ents(corpus,nlp):
  docs = list()
  for doc in nlp.pipe(corpus, disable=["tagger", "parser"]):
      dictionary=dict.fromkeys(["text", "annotations"])
      dictionary["text"]= str(doc)
      dictionary['text_sha256'] =  hashlib.sha256(dictionary["text"].encode('utf-8')).hexdigest()
      annotations=[]

      for e in doc.ents:
        ent_id = hashlib.sha256(str(e.text).encode('utf-8')).hexdigest()
        # "Statartando" análise de entidades a partir da label "MED" - "Sensaz".
        # E criando hash de id para cada conjunto de entidades extraída, de cada sentença, individualmente.
        ent = {"start":e.start_char,"end":e.end_char, "label":e.label_,"label_upper":e.label_.upper(),"text":e.text,"id":ent_id}
        if e.label_ == "MED":
          ent["Sensaz"] = str(e.text[0])

        annotations.append(ent)

      dictionary["annotations"] = annotations
      docs.append(dictionary)
  print(annotations)
  return docs

parsed_ents = extract_ents(corpus,nlp)
```

```
[{'start': 11, 'end': 17, 'label': 'DOSE', 'label_upper': 'DOSE', 'text': '2 comp', 'id': '0664597b669ee7a1a4069d98e90d17688e11e62
```

```
# visualizar detalhes...
parsed_ents
```

```
 'text': '10 mg',
 'id': 'a384ff3ff47a6f7e201ee9a9d7f779e90c4d93aaa015bc34410e9cc96e798e06'},
{'start': 38,
 'end': 52,
 'label': 'APRAZ',
 'label_upper': 'APRAZ',
 'text': '2 vezes ao dia',
 'id': '646d71b3c30a84599501cc76880693cb12181437691c567ca412d50f78e32d92'}],
'text_sha256': 'b7595d4064b42825ce26e325b98ae5553ba0c8dc9028abefbc9f139aee98e883'},
{'text': 'A paciente  faz uso continuamente de  1CP  Diovan Amlo Fix  de  15 Mg  6/6 hs  O Sr.  disse que  sente dores nas
costas.',
 'annotations': [{'start': 38,
 'end': 41,
 'label': 'DOSE',
 'label_upper': 'DOSE',
 'text': '1CP',
 'id': 'a611a272bc5999700be0785651ecf6239adf5fef00aaddf74ffe36cb9d9957a6'},
{'start': 43,
 'end': 58,
 'label': 'MED',
 'label_upper': 'MED',
 'text': 'Diovan Amlo Fix',
 'id': '4c0fa264be6a3413e31e6f1226d08e7a0c6249973aa5076b97be8cd6a90f0213',
 'Sensaz': 'D'},
```

Usando o modelo de relacionamento para extrair as relações previstas, a partir das entidades extraídas anteriormente.

```
# importando bibliotecas.
from pathlib import Path
import random
import typer

import spacy
from spacy.tokens import DocBin, Doc
from spacy.training.example import Example

# para acessar scripts necessários para a previsão de relacionamentes de entidades.
from rel_pipe import make_relation_extractor, score_relations
from rel_model import create_relation_model, create_classification_layer, create_instances, create_tensors

# Atenção: se necessário reinicie o tempo de execução após a instalação dos deps.
# carregando o modelo de previsão de relacionamento: DOSAGEM, CONCENTRACAO e TEMPO_DE_ADMINISTRACAO
nlp2 = spacy.load("/content/drive/MyDrive/rel_component/training/model-best")

# Função para extrair relacinomamentos de entidades a partir dos modelos e do corpus de sentenças.
def extract_relations(corpus,nlp,nlp2):
  predicted_rels = list()
  for doc in nlp.pipe(corpus, disable=["tagger", "parser"]):
    source_hash = hashlib.sha256(doc.text.encode('utf-8')).hexdigest()
    for name, proc in nlp2.pipeline:
          doc = proc(doc)

    for value, rel_dict in doc._.rel.items():
      for e in doc.ents:
        for b in doc.ents:
          if e.start == value[0] and b.start == value[1]:
            max_key = max(rel_dict, key=rel_dict. get)
            print(max_key)
            e_id = hashlib.sha256(str(e).encode('utf-8')).hexdigest()
            b_id = hashlib.sha256(str(b).encode('utf-8')).hexdigest()
            if rel_dict[max_key] >=0.9 :
              print(f" entities: {e.text, b.text} --> predicted relation: {rel_dict}")
              predicted_rels.append({'head': e_id, 'tail': b_id, 'type':max_key, 'source': source_hash})
  return predicted_rels

predicted_rels = extract_relations(corpus,nlp,nlp2)
```

Neo4j - Conectando com o banco de dados on-lie do **neo4j sandbox** e usando a linguagem de consulta **Cypher** do **Neo4j** para criar **grafos de conhecimento** com as entidades e os relacionamentos extraídos do conjunto de 300 sentenças de receituários, usadas neste teste.

```
!pip install neo4j
```

```
# Alocando daraset, NER e REL...

documentos = get_setencas()
corpus = documentos[:]
parsed_ents = extract_ents(corpus,nlp)
predicted_rels = extract_relations(corpus,nlp,nlp2)
```

```
DOSAGEM
DOSAGEM
 entities: ('Tevavinbla', '2cp') --> predicted relation: {'DOSAGEM': 0.98765546, 'CONCENTRACAO': 0.003246911, 'TEMPO_DE_ADMINIST
CONCENTRACAO
 entities: ('Tevavinbla', '10 ML') --> predicted relation: {'DOSAGEM': 0.0040398003, 'CONCENTRACAO': 0.9977997, 'TEMPO_DE_ADMINI
TEMPO_DE_ADMINISTRACAO
DOSAGEM
CONCENTRACAO
CONCENTRACAO
DOSAGEM
TEMPO_DE_ADMINISTRACAO
CONCENTRACAO
DOSAGEM
CONCENTRACAO
TEMPO_DE_ADMINISTRACAO
DOSAGEM
 entities: ('cloridrato de paroxetina', '1 capsula') --> predicted relation: {'DOSAGEM': 0.9943328, 'CONCENTRACAO': 0.006345295,
CONCENTRACAO
 entities: ('cloridrato de paroxetina', '15 Mg') --> predicted relation: {'DOSAGEM': 0.0062042847, 'CONCENTRACAO': 0.99765885, '
TEMPO_DE_ADMINISTRACAO
 entities: ('cloridrato de paroxetina', '2 vezes ao dia') --> predicted relation: {'DOSAGEM': 0.0030477494, 'CONCENTRACAO': 0.00
DOSAGEM
CONCENTRACAO
TEMPO_DE_ADMINISTRACAO
DOSAGEM
TEMPO_DE_ADMINISTRACAO
CONCENTRACAO
DOSAGEM
CONCENTRACAO
DOSAGEM
DOSAGEM
 entities: ('Deeplin', '2 capsulas') --> predicted relation: {'DOSAGEM': 0.99241525, 'CONCENTRACAO': 0.0054166606, 'TEMPO_DE_ADM
CONCENTRACAO
 entities: ('Deeplin', '15ML') --> predicted relation: {'DOSAGEM': 0.0071643894, 'CONCENTRACAO': 0.9979814, 'TEMPO_DE_ADMINISTRA
TEMPO_DE_ADMINISTRACAO
DOSAGEM
CONCENTRACAO
CONCENTRACAO
DOSAGEM
TEMPO_DE_ADMINISTRACAO
CONCENTRACAO
DOSAGEM
CONCENTRACAO
DOSAGEM
DOSAGEM
 entities: ('glimepirida', '20 gotas') --> predicted relation: {'DOSAGEM': 0.99587184, 'CONCENTRACAO': 0.0049184486, 'TEMPO_DE_A
CONCENTRACAO
 entities: ('glimepirida', '30 ml') --> predicted relation: {'DOSAGEM': 0.011232655, 'CONCENTRACAO': 0.9980385, 'TEMPO_DE_ADMINI
DOSAGEM
DOSAGEM
CONCENTRACAO
CONCENTRACAO
DOSAGEM
TEMPO_DE_ADMINISTRACAO
CONCENTRACAO
DOSAGEM
CONCENTRACAO
TEMPO_DE_ADMINISTRACAO
```

```
# consulta e função básica de conexão a um banco previamente criado no neo4j sandbox.
from neo4j import GraphDatabase
import pandas as pd

host = 'bolt://3.219.247.60:7687'
user = 'neo4j'
password = 'maples-wages-rest'
driver = GraphDatabase.driver(host,auth=(user, password))

def neo4j_query(query, params=None):
  with driver.session() as session:
    result = session.run(query, params)
    return pd.DataFrame([r.values() for r in result], columns=result.keys())
```

Usando Cypher para criar e persistir os nós, relacionamentos e propriedades no banco gráfico neo4j.

```
neo4j_query("""
MATCH (n) DETACH DELETE n;
""")

# Cria um primeiro nodo principal de lista de pacientes.
neo4j_query("""
MERGE (l:ListaPacientes {name:"Lista de Pacientes"})
RETURN l
""")
```

```
#add entidades para o KG: Med, Concen, Dose, Apraz
neo4j_query("""
MATCH (l:ListaPacientes)
UNWIND $data as row
MERGE (p:Paciente{id:row.text_sha256})
SET p.text = row.text
MERGE (l)-[:ID_DO_PACIENTE]->(p)
WITH p, row.annotations as entities
UNWIND entities as entity
MERGE (e:Entidades {id: entity.id})
ON CREATE SET
              e.name = entity.text,
              e.label = entity.label_upper
MERGE (p)-[m:FAZ_USO_DE]->(e)
ON CREATE SET m.count = 1
ON MATCH SET m.count = m.count + 1
WITH e as e
CALL apoc.create.addLabels( id(e), [ e.label ] )
YIELD node
REMOVE node.label
RETURN node
""", {'data': parsed_ents})

#Add relations to KG
neo4j_query("""
UNWIND $data as row
MATCH (source:Entidades {id: row.head})
MATCH (target:Entidades {id: row.tail})
MATCH (paci:Paciente {id: row.source})
MERGE (source)-[:REL]->(r:Relacoes {type: row.type})-[:REL]->(target)
MERGE (paci)-[:PACIENTE_INFO_MED]->(r)
""", {'data': predicted_rels})
```

```
#Adicionar propriedade 'nome' à entidade Paciente
res = neo4j_query("""
MATCH (p:Paciente)
RETURN p.id as id, p.name as name
""")
```

Testando consulas no banco criado no Neo4j-Sandbox

```
# Contando número de nós de pacientes.
query=("""
MATCH (n:Paciente)
RETURN count(n) as Número_de_Entradas
""")

res = neo4j_query(query)
res
```

	Número_de_Entradas
0	300

```
# Consultando e retornado lista medicamentos mais usados, dentre os 300 "pacientes".

query = """
MATCH (paci:Paciente)-[:FAZ_USO_DE]->(m:MED)
RETURN m.name as medicamento, count(paci) as qtd_paciente
ORDER BY qtd_paciente DESC
LIMIT 10
"""
res = neo4j_query(query)
res
```

	medicamento	qtd_paciente
0	Fludalibbs	13
1	Uni Norflox	12
2	glimepirida	11
3	Tractera	11
4	Deeplin	11

```
# "Identificação de paciente"
paciente_id = "087e3b5b8dbd6e2263837d9579b02d9d97223b6bfd60e838ba1fc8709413fd2f"
```

```
# Consultado dados de um paciente especifico.
query = """
MATCH (p2:Paciente {id:$id})-[:FAZ_USO_DE]->(m:MED),
(p2:Paciente {id:$id})-[:FAZ_USO_DE]->(c:CONCEN),(p2:Paciente {id:$id})-[:FAZ_USO_DE]->(d:DOSE),
(p2:Paciente {id:$id})-[:FAZ_USO_DE]->(a:APRAZ)
RETURN DISTINCT p2.id as Id, m.name as Med, c.name as Concen, d.name as Dose, a.name as Apraz
"""
res = neo4j_query(query,{"id":paciente_id})
res
```

	Id	Med	Concen	Dose	Apraz
0	087e3b5b8dbd6e2263837d9579b02d9d97223b6bfd60e8...	Tractera	10ml	2 comp	8 em 8 horas

```
# Consultando todos os "pacientes".

query = """
MATCH (p2:Paciente)-[:FAZ_USO_DE]->(m:MED),
(p2:Paciente)-[:FAZ_USO_DE]->(c:CONCEN),(p2:Paciente)-[:FAZ_USO_DE]->(d:DOSE),
(p2:Paciente)-[:FAZ_USO_DE]->(a:APRAZ)
RETURN DISTINCT p2.id as id, m.name as med, c.name as concen, d.name as dose, a.name as apraz
"""
res = neo4j_query(query,{"id":paciente_id})
res
```

	id	med	concen	dose	apraz
0	fbca4b9779a6bfa92bfc08e4d45f2ead80c934a2581192...	aripiprazol	200 ml	2 Cápsulas	6 em 6 horas
1	e806546873479144a5a7ef22e5221ea35dda4dac858afc...	Tanisea	3 ML	2 comp	6 em 6 horas
2	b5bfe90de3f8f459d0e231724e66d1bf1cb7d3cccd02d6...	Uni Norflox	600 ml	20 gotas	6 em 6 horas
3	88c619649c5b490ea57a3ee802491dc4c39bfb0f8cdefd...	Diovan Amlo Fix	30MG	1 Cápsula	6 em 6 horas
4	f947f16110d5285b7bb9f608ed2627f9272c1ad2576414...	Tractera	10ML	1 COMPR	6 em 6 horas
...	...	...	...	...	...
365	418792081653l8f6b4bf1222b54ee469f36f6e473d582e...	Aristab	5 MG	1 cp	4 em 4 horas
366	8bcf21b4757e5e8fb1cd708721cc5361052df156256d50...	Nemoxil	20 Ml	2 capsulas	4 em 4 horas
367	3a41843c52cecfa18cd0aecb0e57a6c49201826170b7ab...	Cebrilin	30 ml	2 capsulas	4 em 4 horas
368	380fda752af2b334c54d589d0a5a4ff10514c36951b193...	Dormonid	15ML	2 Cápsulas	4 em 4 horas
369	4a1bd48870214f3e70faf89abd1da7840610ed073c77bb...	Osteotec	100ml	1 cp	4 em 4 horas

370 rows × 5 columns

```
import pandas as pd

# salvando a consulta anterior como .csv.
df = pd.DataFrame(res)
df.to_csv(r'/content/drive/MyDrive/exportar_dataframe.csv', index=False, header=True)


# lerndo dados salvos.
df = pd.read_csv("/content/drive/MyDrive/exportar_dataframe.csv",sep=',')
df.head(6)
```

Atenção: apenas só mais um exemplo, obviamnte se executar apagara todos os elementos do banco de dados.

É possivel criar outro grafo com o csv gerado na consultas das entidades etraida de cada paciente. Enfim...

```
# Apagando e Recriando nó de lista pacientes.

neo4j_query("""
```

```
MATCH (n) DETACH DELETE n;
""")

neo4j_query("""
MERGE (l:listaPacientes {name:"Lista de Pacientes"})
RETURN l
""")


# Criando nós, relacionamentos e propriedades de pacientes.

neo4j_query("""
LOAD CSV WITH HEADERS FROM 'https://docs.google.com/spreadsheets/d/e/2PACX-1vSZozMwNyeeH0l5KuneHDfww8h-7I28-R0y86dWsKsiqTsz2pA-U1RXhxPaP
MATCH (l:listaPacientes)
MERGE (p:Paci{id:row.id})
MERGE (l)-[:ID_DO_PACIENTE]->(p)
SET p.med = row.med,
    p.concen = row.concen,
    p.dose = row.dose,
    p.apraz = row.apraz
MERGE (m:Med{med:row.med}) MERGE (p)-[:FAZ_USO_DE]->(m)
""")
```


✓ 0s conclusão: 22:35