UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Pablo Ernesto Vigneaux Wilton

Um Método Voltado à Representação de Conhecimento a Partir de Textos Acadêmicos Sobre Diabetes Mellitus

Florianópolis
2024

Pablo Ernesto Vigneaux Wilton

Um Método Voltado à Representação de Conhecimento a Partir de Textos Acadêmicos Sobre Diabetes Mellitus

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestre em Engenharia e Gestão do Conhecimento.

Orientador: Prof. Dr. Aires José Rover
Coorientador: Prof. Dr. Neri dos Santos

Florianópolis
2024

Vigneaux Wilton, Pablo Ernesto
Um Método Voltado à Representação de Conhecimento a Partir de Textos Acadêmicos Sobre Diabetes Mellitus / Pablo Ernesto Vigneaux Wilton ; orientador, Aires José Rover, coorientador, Neri Dos Santos, 2024.
200 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2024.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Mineração de Dados. 3. Inteligência Artificial. 4. Diabete. 5. Grafos de Conhecimento. I. Rover, Aires José. II. Dos Santos, Neri. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Pablo Ernesto Vigneaux Wilton

Um Método Voltado à Representação de Conhecimento a Partir de Textos Acadêmicos Sobre Diabetes Mellitus

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 25 de abril de 2024, pela banca examinadora composta pelos seguintes membros:

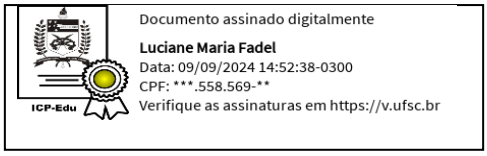
Prof. Aires José Rover, Dr.
Instituição Universidade Federal de Santa Catarina

Prof. Alexandre Gonçalves, Dr.
Instituição Universidade Federal de Santa Catarina

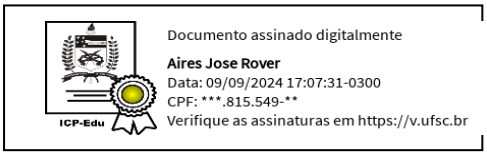
Prof. Fernando Alvaro Ostuni Gauthier, Dr.
Instituição Universidade Federal de Santa Catarina

Profa. Isabela Cristina Sabo, Dra.
Instituição Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Engenharia e Gestão do Conhecimento.



Coordenação do Programa de Pós-Graduação



Prof. Aires José Rover, Dr.
Orientador

Florianópolis, 2024.

Este trabalho é dedicado aos meus pais, Sonia Amanda Wilton Calé (in memoriam) e Pablo Gustavo Vigneaux Jouanne, a quem devo imensamente; aos meus amigos, professores e colegas que me acompanharam nesta jornada; a Luzcka, Prata, Esher, Chimú, Papaloro, Ackla, Quiltro, Cucho, Cucha, Chica, Pietra, Pietro, Panza, Oreja, Scar, Poly, Tigra, Nébula, Chicão, Max, Preta e tantos outros que fazem parte do meu coração e do meu ser.

AGRADECIMENTOS

Agradeço a Deus por tudo o que foi colocado no meu caminho e me fez ser quem eu sou hoje. Aos meus pais, dos quais sou fruto, e que foram meus primeiros educadores. Agradeço ao meu orientador Prof. Dr. Aires José Rover pelas ideias, orientações e confiança.

Para tantas pessoas, amigos, colegas e conhecidos com os quais compartilhei bons e maus momentos, com um agradecimento especial a Rodrigo Rafael Cunha. Agradeço aos meus pets que sempre me fizeram sorrir e apreciar a simplicidade do amor puro.

Agradeço à Universidade Federal de Santa Catarina pela oportunidade no curso de Pós-Graduação de Engenharia e Gestão do Conhecimento; pelos diversos eventos, e o Congresso Internacional de Conhecimento e Inovação (CIKI).

Agradeço ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pela bolsa que me acompanhou boa parte desta trajetória.

“O conhecimento torna-se mal se o objetivo não for virtuoso.”

(Platão)

RESUMO

O crescimento constante das publicações científicas apresenta desafios significativos na extração e organização do conhecimento gerado. Este trabalho aborda esses desafios no contexto de diabetes mellitus, explorando técnicas de Processamento de Linguagem Natural (NLP) para mineração de dados. A construção de um Grafo de Conhecimento (KG) a partir de abstracts de artigos científicos é investigada para representar as entidades nomeadas e seus relacionamentos. Este trabalho tem como objetivo estudar a extração de entidades nomeadas e seus relacionamentos a partir de abstracts de artigos científicos sobre diabetes mellitus. Utilizando técnicas e modelos de NLP, visa-se a construção de um Grafo de Conhecimento para facilitar a extração e organização do conhecimento, assim como facilitar a aplicação de técnicas que, por meio de inferências, gerem novos conhecimentos na área biomédica. Para a seleção de artigos, utilizou-se a plataforma PubMed, resultando em um total inicial de 518.432 registros, posteriormente filtrados para 361.688 registros. Abstracts foram utilizados para a análise. Inicialmente, foi realizada uma Análise Exploratória de Dados (EDA) para entender e preparar os dados. Em seguida, técnicas de NLP, como o reconhecimento de entidades nomeadas (NER) e a extração de relacionamentos entre entidades (ERE), foram aplicadas para identificar e extrair entidades e seus relacionamentos. Este processo permitiu a construção de triplas que compõem o Grafo de Conhecimento. A abordagem foi inspirada pelo Design Science Research Methodology (DSRM), que oferece uma estrutura sistemática para a criação e avaliação de artefatos na pesquisa científica. Os resultados demonstraram que o uso de modelos de Machine Learning e Deep Learning é eficaz na identificação de entidades relevantes em textos acadêmicos, permitindo a construção de um Grafo de Conhecimento robusto. O grafo resultante forneceu uma estrutura rica para análise e visualização das relações entre entidades biomédicas. No entanto, a pesquisa também identificou desafios significativos, como a necessidade de maior poder computacional e melhorias nas técnicas de pré-processamento e extração de relações.

Palavras-chave: Mineração de Dados; Grafo de Conhecimento; Processamento de Linguagem Natural; Reconhecimento de Entidades Nomeadas; Diabetes Mellitus.

ABSTRACT

The constant growth of scientific publications presents significant challenges in the extraction and organization of the generated knowledge. This work addresses these challenges in the context of diabetes mellitus by exploring Natural Language Processing (NLP) techniques for data mining. The construction of a Knowledge Graph (KG) from scientific article abstracts is investigated to map named entities and their relationships. The objective of this work is to study the extraction of named entities and their relationships from scientific article abstracts on diabetes mellitus. By using NLP techniques and models, the aim is to construct a Knowledge Graph to facilitate the extraction and organization of knowledge, as well as to enable the application of techniques that, through inferences, generate new knowledge in the biomedical field. For the selection of articles, the PubMed platform was used, resulting in an initial total of 518,432 records, which were subsequently filtered to 361,688 records. Abstracts were used for the analysis. Initially, Exploratory Data Analysis (EDA) was performed to understand and prepare the data. Next, NLP techniques such as Named Entity Recognition (NER) and Entity Relation Extraction (ERE) were applied to identify and extract entities and their relationships. This process allowed the construction of triples that make up the Knowledge Graph. The approach was inspired by the Design Science Research Methodology (DSRM), which provides a systematic framework for creating and evaluating artifacts in scientific research. The results demonstrated that the use of Machine Learning and Deep Learning models is effective in identifying relevant entities in academic texts, enabling the construction of a robust Knowledge Graph. The resulting graph provided a rich structure for analyzing and visualizing the relationships between biomedical entities. However, the research also identified significant challenges, such as the need for greater computational power and improvements in preprocessing and relation extraction techniques.

Keywords: Data Mining; Knowledge Graph; Natural Language Processing; Named Entity Recognition; Diabetes Mellitus.

LISTA DE FIGURAS

Figura 1 - NER aplicado em texto de Biomedicina.....55
Figura 2 - Arquitetura de Rede Neural *Tranformer*.....60
Figura 3 - Exemplo de um Grafo simples direcionado.....66
Figura 4 - Grafo de artigos relacionados a Rossanez *et al*. (2020).....82
Figura 5 - Capítulos referentes ao desenvolvimento desta dissertação.....84
Figura 6 - Pseudocódigo do algoritmo para a consulta ao PubMed.....87
Figura 7 - Extrato de metadados de um artigo do PubMed em formato JSon.....88
Figura 8 - Texto completo de Abstract utilizado no Processamento.....92
Figura 9 - Sentença extraída, do abstract, para posterior processamento.....93
Figura 10 - Etapas do método proposto.....97
Figura 11 - Estrutura resultante da primeira seleção de features.....104
Figura 12 - Quantidade de papers por idioma.....105
Figura 13 - Visualização da aplicação do LDA em Topic Modelling.....106
Figura 14 - Resultado da aplicação do modelo ner_bionlp13cg_md.....107
Figura 15 - Quantidade e entidades identificadas pelo modelo ner_bionlp13cg_md 107
Figura 16 - Relatório da execução do modelo BERT-base-SRL.....108
Figura 17 - Amostra de triplas geradas pelo modelo BERT-base-SRL.....108
Figura 18 - Zoom realizado no vértice hyperglycemia.....109
Figura 19 - Métricas geradas a partir do KG criado.....110
Figura 20 - Evolução da produção anual na língua inglesa e outras.....114
Figura 21 - Produção de papers relativos ao tema diabetes mellitus na América Latina.....115
Figura 22 - Produção de artigos per capita relacionados ao tema diabetes mellitus na América Latina.....115
Figura 23 - Produção por idioma no Brasil.....116
Figura 24 - Evolução por ano da quantidade de palavras usadas nos abstracts.....117
Figura 25 - BoW aplicado aos abstracts.....118
Figura 26 - Termos, em duplas, mais utilizados nos papers.....119
Figura 27 - Grafo criado utilizando relacionamento entre palavras-chave.....120
Figura 28 - Gráfico para a escolha da melhor métrica de coherence.....122

Figura 29 - Resultado do LDA visualizado utilizando o LDAvis.....124
Figura 30 - Pares de termo e a probabilidade que compõe cada tópico.....125
Figura 31 - Distribuição de probabilidade de cada tópico.....127
Figura 32 - Análise das distribuições de densidade de probabilidade.....129
Figura 33 - Pares de termo e a probabilidade que compõe cada tópico NMF.....131
Figura 34 - Escore de palavras de tópico do BERTopic.....133
Figura 35 - Probabilidade dos termos para o tópico 423.....135
Figura 36 - Visualização hierárquica de tópicos - BERTopic.....136
Figura 37 - Matriz de similaridade intertópicos - BERTopic.....137
Figura 38 - Escore de palavras por tópico - BERTopic (k = 7).....138
Figura 39 - Fluxo da extração das entidades nomeadas.....141
Figura 40 - Distribuição de entidades reconhecidas pelo modelo en_ner_bionlp13cg_md.....146
Figura 41 - Entidades identificadas em A0S6 - en_ner_bionlp13cg_md.....147
Figura 42 - Entidades identificadas em A0S7 - en_ner_bionlp13cg_md.....147
Figura 43 - Entidades identificadas em A0S6 - en_core_sci_sm.....149
Figura 44 - Entidades identificadas em A0S7 - en_core_sci_sm.....149
Figura 45 - Entidades identificadas em A0S6 - en_core_sci_scibert.....151
Figura 46 - Entidades identificadas em A0S7 - en_core_sci_scibert.....151
Figura 47 - Entidades identificadas em A0S6 - en_core_sci_lg.....153
Figura 48 - Entidades identificadas em A0S7 - en_core_sci_lg.....153
Figura 49 - Distribuição de entidades reconhecidas pelo modelo en_ner_bc5cdr_md 155
Figura 50 - Entidades identificadas em A0S6 - en_ner_bc5cdr_md.....156
Figura 51 - Entidades identificadas em A0S7 - en_ner_bc5cdr_md.....156
Figura 52 - Distribuição de entidades reconhecidas pelo modelo biomedical-ner-all 158
Figura 53 - Amostra do registro da classificação de entidades - biomedicall-ner-all 160
Figura 54 - Entidades identificadas em A0S6 - biomedical-ner-all.....160
Figura 55 - Entidades identificadas em A0S7 - biomedical-ner-all.....161
Figura 56 - Amostra de entidades não reconhecidas.....163
Figura 57 - Visualização do diagrama da ontologia DMTO.....163
Figura 58 - Distribuição de entidades reconhecidas por todos os modelos.....165

Figura 59 - Algoritmo para identificação de triplas.....167
Figura 60 - Estrutura do dataframe que armazena as triplas.....168
Figura 61 - Amostra de triplas descobertas.....168
Figura 62 - Amostra do RDF gerado ao salvar as triplas.....169
Figura 63 - Visão do vértice Hiperglicemia.....171
Figura 64 - Zoom dado no Grafo gerado, no centro o vértice da diabetes.....171
Figura 65 - Métricas geradas para o KG construído.....173
Figura 66 - Fluxo do uso do ChatGPT.....176
Figura 67 - ChatGPT - Etapa de definição de papel.....177
Figura 68 - ChatGPT - Definindo o escopo e a saída.....177
Figura 69 - ChatGPT - Informando o abstract de entrada.....178
Figura 70 - ChatGPT - Primeiras linhas da lista de entidades identificadas.....179
Figura 71 - ChatGPT - Últimas linhas da lista de entidades identificadas.....179
Figura 72 - ChatGPT - Parte final da resposta, contendo a lista de triplas.....180

LISTA DE QUADROS

Quadro 1 - Modelos que fazem parte da arquitetura Transformers e ano de lançamento.....61
Quadro 2 - Artigos selecionados com suas principais informações. (continua).....80
Quadro 3 - Mapeamento atividades de pesquisa e etapas DSRM.....83
Quadro 4 - Síntese das atividades desenvolvidas nesta pesquisa.....95
Quadro 5 - Modelos de NER da área de Biomedicina utilizados.....141
Quadro 6 - Sentença a serem utilizadas nas análises.....143
Quadro 7 - Categorias de entidades reconhecidas pelo modelo en_ner_bionlp13cg_md.....144
Quadro 8 - Categorias existentes no modelo en_core_sci_sm.....148
Quadro 9 - Categorias existentes no modelo en_ner_bc5cdr_md.....154
Quadro 10 - Ontologias e bases de dados de doenças utilizadas na comparação. 162
Quadro 11 - Quantidade de entidades não reconhecidas pelo modelo ner_bc5cdr_md.....162
Quadro 12 - Modelos de RE identificados e analizados.....166
Quadro 13 - Quantidade de triplas identificadas para a amostra de 1000 abstracts 167

LISTA DE TABELAS

Tabela 1 - Quantidade de entidades reconhecidas pelo modelo en_ner_bionlp13cg_md.....145
Tabela 2 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_sm..148
Tabela 3 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_scibert 150
Tabela 4 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_lg....152
Tabela 5 - Quantidade de entidades reconhecidas pelo modelo en_ner_bc5cdr_md 154
Tabela 6 - Quantidade de entidades reconhecidas pelo modelo biomedical-ner-all. Primeiros 16 registros.....158
Tabela 7 - Identificação do termo diabetes como sujeito e como objeto na tripla....169

LISTA DE ABREVIATURAS E SIGLAS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BERTopic	Bidirectional Encoder Representations from Transformers for Topic Modeling
BioBERT	Biomedical BERT
BLURB	Biomedical Language Understanding and Reasoning Benchmark
BoW	Bag of Words
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBOW	Continuous Bag of Words
CRF	Conditional Random Field
CRFs	Conditional Random Fields
DL	Deep Learning
DSRM	Design Science Research Methodology
DS	Design Science
ELMo	Embeddings from Language Models
ER	Entity Relations
ERIC	Education Resources Information Center
GNNs	Graph Neural Networks
GPT	Generative Pre-trained Transformer
HDP	Hierarchical Dirichlet Process
HMMs	Hidden Markov Models
ICD-9-CM	International Classification of Diseases, Ninth Revision, Clinical Modification
IDF	International Diabetes Federation
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery from Text
KG	Knowledge Graph
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LLMs	Large Language Models
LSA	Latent Semantic Analysis

LSTM	Long Short-Term Memory
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
pLSA	Probabilistic Latent Semantic Analysis
POS	Part of Speech
POS-tagging	Part of Speech Tagging
RAG	Retrieval-Augmented Generation
RC	Relation Classification
RE	Relation Extraction
RNN	Recurrent Neural Network
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System

SUMÁRIO

1 INTRODUÇÃO 20
1.1 DEFINIÇÃO DO PROBLEMA 22
1.2 PERGUNTA DE PESQUISA 25
1.3 OBJETIVOS 25
1.3.1 Objetivo Geral 25
1.3.2 Objetivo Específico 25
1.4 JUSTIFICATIVA E RELEVÂNCIA DO TEMA 25
1.5 ESCOPO DA PESQUISA 28
1.6 ADERÊNCIA AO PPGEGC 29
1.6.1 Argumentação 29
1.6.2 Referências Factuais 31
1.6.2.1 Mineração de Textos 32
1.6.2.2 Mineração de dados 32
1.6.2.3 Descoberta de conhecimento 32
1.6.2.4 Aprendizado de Máquina 33
1.6.2.5 Grafos de Conhecimento 33
1.7 ESTRUTURA DA DISSERTAÇÃO 34
2 FUNDAMENTAÇÃO TEÓRICA 35
2.1 ÁREAS DE CONHECIMENTO ENVOLVIDAS 35
2.2 ANÁLISE EXPLORATÓRIA DE DADOS E MINERAÇÃO DE DADOS 39
2.3 INTELIGÊNCIA ARTIFICIAL 41
2.4 PROCESSAMENTO DE LINGUAGEM NATURAL 43
2.4.1 Modelagem de Tópicos 48
2.4.2 Reconhecimento de Entidade Nomeada e Extração de Relação 51
2.4.3 Modelos de Linguagem de Grande Escala 57
2.4.3.1 Transformers 59
2.4.3.2 BERT 62
2.4.3.3 GPT 63
2.4.4 Retrieval-Augmented Generation 64
2.5 GRAFOS DE CONHECIMENTO 65
2.6 CONSIDERAÇÕES FINAIS 72

3 METODOLOGIA DE PESQUISA....74
3.1 ENQUADRAMENTO METODOLÓGICO....74
3.2 DESIGN SCIENCE RESEARCH METHODOLOGY....76
3.3 REVISÃO INTEGRATIVA DA LITERATURA....78
3.4 DESENVOLVIMENTO DA PESQUISA....82
3.4.1 Definição do Problema e Motivação....84
3.4.2 Definição dos Objetivos....85
3.4.3 Projeto e Desenvolvimento....85
3.4.3.1 Coleta dos Dados....85
3.4.3.2 Análise Exploratória de Dados....88
3.4.3.3 Pré-processamento....90
3.4.3.4 Transformação dos dados....91
3.4.3.5 Representação de Conhecimento....91
3.4.3.6 Explicitação do Conhecimento....93
3.4.4 Demonstração....93
3.4.5 Avaliação....93
3.4.6 Comunicação....95
3.5 SÍNTESE DA METODOLOGIA DE PESQUISA....95
4 MÉTODO PROPOSTO....97
4.1 APRESENTAÇÃO DO MÉTODO....97
4.2 COMPOSIÇÃO DO MÉTODO....99
4.2.1 Etapa 1: Coleta de Textos....99
4.2.2 Etapa 2: *Exploratory Data Analysis*....100
4.2.3 Etapa 3: Representação de Conhecimento....100
4.2.4 Etapa 4: Avaliação....101
4.2.5 Etapa 5: Apresentação dos Resultados....102
4.3 INSTANCIAÇÃO DO MODELO....102
4.3.1 Etapa 1: Coleta de Textos....102
4.3.2 Etapa 2: Exploratory Data Analysis....103
4.3.3 Etapa 3: Representação de Conhecimento....106
4.3.4 Etapa 4: Avaliação....109
4.3.5 Etapa 5: Apresentação dos Resultados....111
4.4 CONSIDERAÇÕES FINAIS....111
5 ANÁLISE E DISCUSSÃO DOS RESULTADOS....113

5.1 AVALIAÇÃO DO CONJUNTO DE DADOS....113
5.2 AVALIAÇÃO DA ETAPA 2: EDA....113
5.2.1 Análises de Distribuição dos Papers....113
5.2.2 Análises de Distribuição das Palavras....116
5.2.3 Topic Modelling – Latent Dirichlet Allocation....121
5.2.4 Topic Modelling – Non-negative Matrix Factorization....130
5.2.5 Topic Modelling – BERTopic....132
5.2.6 EDA – Avaliação de Resultados....139
5.3 AVALIAÇÃO DA ETAPA 3: REPRESENTAÇÃO DE CONHECIMENTO140
5.3.1 Named Entity Extraction....140
5.3.2 Modelo - en_ner_bionlp13cg_md....143
5.3.3 Modelo - en_core_sci_sm....147
5.3.4 Modelo – en_core_sci_lg....152
5.3.5 Modelo - en_ner_bc5cdr_md....153
5.3.6 Modelo - biomedical-ner-all....156
5.3.7 Entidades NER e Ontologias: Um Estudo Experimental....161
5.3.8 NER – Avaliação de Resultados....164
5.3.9 Entity Relation Extraction....165
5.3.10 Avaliação e visulização do Knowledge Graph....170
5.3.11 Experimento utilizando Generative Pre-trained Transformer....176
6 CONCLUSÕES E TRABALHOS FUTUROS....182
6.1 LIMITAÇÕES....185
6.2 PERSPECTIVAS E TRABALHOS FUTUROS....185
REFERÊNCIAS....187
APÊNDICE A – Protocolo para Revisão Integrativa....199

1 INTRODUÇÃO

Os dados biomédicos são complexos e heterogêneos, o que aumenta os desafios de compreensão e exploração. Com o crescimento do volume desses dados nos últimos anos, surgiram mais oportunidades para explorar suas interações e efeitos. A preparação dos dados é uma tarefa extensa e que consome muito tempo no campo biomédico, envolvendo a coleta de dados de fontes múltiplas e confiáveis, análise de suas estatísticas e processamento em formatos específicos que atendam aos requisitos do problema biomédico relacionado. Problemas cruciais nesse campo incluem descoberta de medicamentos, reconhecimento de entidades, extração de relações e detecção de reações adversas a medicamentos. Assim, a utilização de ferramentas apropriadas para investigar esses dados volumosos se faz necessária, especialmente no domínio biomédico (SAHA *et al.*, 2019).

O número de periódicos científicos cresceu exponencialmente, de 10 no final do século XVII para 100.000 no final do século XX, com a publicação evoluindo de impressa para online. Em 2020, estimava-se que 3,4 milhões de artigos foram publicados, alinhando-se com o relatório de 2018 da Scientific, Technical, and Medical (STM), que apontava mais de três milhões de artigos científicos, técnicos e médicos publicados anualmente. O número médio de artigos por periódico aumentou de 74,2 em 1999 para 99,6 em 2016 (GHASEMI *et al.*, 2022).

Esta vasta quantidade de documentos de textos não estruturados possui um grande valor científico, contendo o conhecimento de inúmeras pesquisas, o qual se torna de difícil acesso se pensarmos nisso como uma base de conhecimento onde possamos obter informações significativas e acionáveis que podem afetar positivamente o cuidado com os pacientes e permitir diagnósticos mais precisos, prevenção de doenças, tratamento personalizado e melhor tomada de decisões (ABU-SALIH *et al.*, 2023).

Nesse sentido, o uso de técnicas de Machine Learning (ML) e Deep Learning (DL) aplicadas ao Natural Language Processing (NLP) na construção de um Knowledge Graph (KG), que também é uma forma de representação de conhecimento, tem grande valor na extração e mapeamento de grandes volumes de dados, facilitando seu acesso e análise (NICHOLSON; GREENE, 2020).

Dentro do contexto da Biomedicina, e para este estudo, escolheu-se a diabetes mellitus como o principal assunto na seleção dos artigos que serviram como insumo para a aplicação desta pesquisa. A diabetes mellitus é uma

enfermidade crônica que acarreta severas consequências ao indivíduo e à sociedade, resultando em um impacto social e financeiro expressivo. Conhecida como ‘‘doença silenciosa’’, ela tende a evoluir sem sintomas evidentes, manifestando-se de forma grave em estágios irreversíveis.

Segundo a nona edição do IDF DIABETES ATLAS (2019), a prevalência global de diabetes alcançou 9,3% da população, com aproximadamente 50% dos casos diagnosticados desconhecendo a existência da doença. Na versão mais recente do IDF DIABETES ATLAS (2021), observa-se um aumento na prevalência global, agora estimada em 10,5%, afetando um total de 536,6 milhões de pessoas. No Brasil, a prevalência é estimada em 7,6%, afetando desde crianças até adultos mais velhos. Em 2019, a população mundial afetada pela doença era estimada em 463 milhões de pessoas, resultando em gastos em saúde de USD 760 bilhões.

Conforme o IDF DIABETES ATLAS (2021), o impacto do diabetes é enorme, tanto em termos de saúde pública quanto financeiros. Em 2021, os gastos globais com diabetes atingiram pelo menos USD 966 bilhões, representando um aumento de 316% nos últimos 15 anos. O Brasil se destaca como o terceiro país com maior gasto em saúde relacionado ao diabetes, com USD 42,9 bilhões. Além disso, 81% dos adultos com diabetes vivem em países de baixa e média renda, evidenciando a necessidade urgente de ações globais para prevenir e tratar a doença. O quadro nacional e mundial da doença tende a se agravar devido a diversos fatores, como sedentarismo, alimentação inadequada, obesidade e envelhecimento populacional (FLOR; CAMPOS, 2017). Atualmente, é amplamente reconhecido que, além dos fatores genéticos e da idade, outros fatores, como comportamentais, alimentares, socioeconômicos, escolaridade e ambientais, também influenciam o desenvolvimento da doença (MALTA *et al.*, 2019), aumentando assim a complexidade dos fatores envolvidos.

Os artigos científicos sobre “Diabetes Mellitus” foram obtidos da base de dados PubMed, totalizando 518.432 artigos em uma pesquisa realizada em 20 de maio de 2021. Desses, 361.688 foram selecionados com base em critérios como a disponibilidade de resumos, sendo visível que, mesmo aplicando limitações ao conjunto de dados, o volume de dados é considerável.

Diante da complexidade dos fatores que influenciam o desenvolvimento da diabetes mellitus, e do grande volume de documentos científicos em texto não estruturado, repletos de sentenças e conceitos complexos inerentes à linguagem natural científica, se torna visível a necessidade do uso de métodos alinhados com a

Engenharia do Conhecimento, que utilizem algoritmos, modelos e técnicas advindas das Ciências da Computação de forma automatizada. Nesse contexto, o uso de técnicas de mineração de dados com algoritmos de Natural Language Processing (NLP), como os modelos baseados em Transformers, incluindo BERT e suas variantes, bem como GPT-2 e GPT-3, apresenta-se como uma solução atual e viável (LAI; LU, 2021). Diversos modelos já foram citados no passado, mas fazer uso de textos com grande diversidade, como enciclopédias (BOSSELUT *et al.*, 2019), visa criar uma solução de Compreensão da Linguagem Natural (NLU, sigla em inglês) que busca o entendimento do texto processado. A etapa de mineração dos dados tem foco na extração das chamadas ‘entidades’ e seus relacionamentos, fazendo uso de técnicas como Named Entity Recognition (NER) e Relation Extraction (RE) com o objetivo de criar o Grafo de Conhecimento (KG).

O KG será um banco de dados contendo as informações extraídas dos textos, que poderão ser utilizadas por uma Rede Neural em Grafos (GNN) para análise prescritiva. O KG desempenha um papel fundamental para garantir uma análise adequada, pois a complexidade das relações entre as diferentes entidades é representada por vértices no grafo, tornando-o um modelo ideal para sistemas complexos (ZHOU *et al.*, 2018). A GNN é uma Neural Network (NN) aplicada a um conjunto de dados organizados na forma de um grafo, e desta forma, em um futuro, ser aplicada sobre o KG resultante desta pesquisa., Esse modelo tem sido explorado pela indústria biomédica, especialmente pela indústria farmacêutica, devido à facilidade de visualização dos fatores que levam ao resultado da NN, afastando-se da ideia comum de que uma NN é uma ‘caixa-preta’. Essa maior transparência no motivo que leva ao resultado é extremamente importante na área biomédica e de saúde em geral.

1.1 DEFINIÇÃO DO PROBLEMA

A diabetes mellitus é descrita como uma doença metabólica caracterizada por hiperglicemia crônica, devido a defeitos na secreção de insulina, na ação da insulina ou em ambos os processos. A doença, frequentemente encontrada na comunidade, é causada por distúrbios metabólicos que ocorrem no órgão do pâncreas. Pacientes diabéticos frequentemente enfrentam frustração, raiva, desesperança, estresse, ansiedade e instabilidade emocional, condições que são

denominadas como estresse diabético. Essas alterações psicológicas são resultado do tratamento contínuo que esses pacientes precisam realizar, adicionando uma camada de complexidade ao manejo da doença (AGUSTIN *et al.*, 2024).

Considerada a epidemia do século, a diabetes tem um impacto significativo na saúde pública e nos sistemas de saúde globalmente. Estima-se que, em 2019, cerca de 463 milhões de adultos viviam com diabetes, número que pode chegar a 700 milhões até 2045 (KHARROUBI, 2015). Comparando com os dados mais atuais do International Diabetes Federation, contidos na sua publicação IDF DIABETES ATLAS de 2021, o número de adultos com diabetes aumentou para 537 milhões e a previsão é que alcance 783 milhões até 2045. A doença está associada a complicações crônicas que afetam a qualidade de vida dos pacientes e aumentam os custos dos serviços de saúde. Isso representa um aumento de aproximadamente 16% na previsão de casos para 2045, o que sublinha ainda mais a urgência de intervenções eficazes para prevenção e tratamento, bem como a necessidade de pesquisas contínuas para enfrentar este crescente desafio global.

Conforme Pulugu *et al.* (2024) O crescimento exponencial da literatura biomédica nos últimos anos lançou a comunidade científica em uma era desafiadora, caracterizada por um influxo esmagador de dados textuais. Em 1980, foram registrados 1.825 artigos, crescendo para 50.213 em 2022, representando um aumento de aproximadamente 2.652%. Entre 2000 e 2010, o número de publicações aumentou cerca de 177%, e entre 2010 e 2020, houve um incremento de aproximadamente 108%. No entanto, observa-se um pequeno declínio após 2022, com 47.133 publicações em 2023. Os dados de 2024 são incompletos devido a estarmos na metade do ano. Este crescimento exponencial destaca a importância crescente da pesquisa sobre diabetes ao longo das últimas décadas.

O número de publicações científicas sobre diabetes mellitus cresce exponencialmente, tornando-se desafiador acompanhar todas as novas informações. Em 2020, estimava-se que 3,4 milhões de artigos foram publicados (GHASEMI *et al.*, 2022). O processo tradicional de pesquisa envolve buscas por palavras-chave em bases de dados e a leitura manual para seleção dos artigos relevantes, um método que consome tempo e esforço significativos.

Segundo Desai *et al.* (2018), enquanto ocorre uma explosão da literatura científica disponível em muitos campos, no crescente campo da biomedicina, um impressionante número de 3.000 a 5.000 artigos é publicado diariamente. Não sendo prático navegar por tantos artigos para identificar aqueles que podem ser

relevantes. Para os recém-chegados ao campo, é difícil estabelecer uma compreensão básica dos trabalhos fundamentais.

Além da grande produção de documentos e da complexidade da linguagem natural humana e científica na área da Biomedicina, como já descrito, há, conforme Malta *et al.* (2019), a própria complexidade da doença, que implica o envolvimento de diversos campos de conhecimento, desde nutrição e genética até ciências sociais. Desta forma, não temos apenas um problema de indexação ou busca de documentos, mas sim um problema maior na acessibilidade do próprio conhecimento gerado, o que pode acarretar, e possivelmente acarreta, que conhecimento importante fique oculto entre milhões de páginas de documentos científicos. Conforme Nicholson e Greene (2020), há um problema significativo na explicitação e disponibilização do conhecimento contido nos artigos científicos, muitas vezes não acessível de forma organizada e prática. Melhorar esses processos é essencial para tornar o conhecimento mais eficiente e eficaz.

Neste sentido a mineração de dados de textos tem se tornado importante na Biomedicina, pois pode fornecer insights valiosos que melhoram o cuidado com os pacientes, diagnósticos, tratamentos e decisões clínicas. No entanto, a mineração de dados de textos biomédicos é desafiadora devido à complexidade e heterogeneidade dos dados, à linguagem técnica e especializada, e ao volume crescente de publicações científicas (SAHA *et al.*, 2019).

Goldberg (2017) afirma que a linguagem humana, por sua vez, é ambígua e variável. As pessoas são excelentes em produzir e entender a linguagem, sendo capazes de expressar, perceber e interpretar significados muito elaborados e sutis. Portanto, o uso da computação para a compreensão e o tratamento da linguagem é altamente desafiador.

Para Zhou *et al.* (2018), a integração de tecnologias avançadas, como Processamento de Linguagem Natural (NLP) e Machine Learning (ML), para o mapeamento do conhecimento contido em artigos científicos sobre diabetes mellitus é fundamental para avançar na compreensão e tratamento dessa condição complexa. Isso permite que os profissionais de saúde obtenham insights mais profundos e precisos, melhorando os cuidados com os pacientes e as estratégias de intervenção.

1.2 PERGUNTA DE PESQUISA

Com base no contexto fornecido, apresenta-se a seguinte pergunta de pesquisa que orienta esta dissertação: Como extrair e representar conhecimento contido em grandes quantidades de textos, mais especificamente textos científicos sobre diabetes mellitus?

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Propor e desenvolver um método voltado à representação do conhecimento a partir de textos científicos sobre diabetes mellitus.

1.3.2 Objetivo Específico

- Identificar e catalogar fontes relevantes de literatura científica em biomedicina, com foco específico em estudos e publicações sobre diabetes mellitus;
- Selecionar técnicas para a aplicação de Exploratory Data Analysis (EDA), visando compreender a natureza dos documentos obtidos, e obter insights sobre a área;
- Identificar métodos e técnicas para a extração de informações relevantes no contexto da diabetes mellitus a partir de conteúdo não estruturado;
- Elaborar um cenário de estudo que permita avaliar o método proposto.

1.4 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

A justificativa para a presente pesquisa reside na intersecção de várias áreas de importância crescente na sociedade contemporânea. A Diabetes Mellitus, como já exposto anteriormente, é uma doença crônica que afeta milhões de pessoas em todo o mundo, com implicações significativas tanto para os indivíduos afetados

quanto para a sociedade em geral, resultando em um impacto social e financeiro considerável (IDF DIABETES ATLAS, 2021). O aumento contínuo no número de pacientes afetados por esta doença ressalta a urgência em aprofundar nosso entendimento sobre a condição e desenvolver estratégias eficazes para seu manejo.

Além dos desafios psicológicos, a diabetes mellitus representa uma ameaça significativa à saúde pública devido às suas complicações graves. Entre os problemas que a diabetes pode causar estão a cegueira, insuficiência renal e gangrena nos pacientes. A diabetes mellitus, ou simplesmente diabetes, é uma condição crônica que pode durar a vida toda. Essa doença é causada por um distúrbio metabólico que ocorre no pâncreas, caracterizado pelo aumento do nível de açúcar no sangue, uma condição frequentemente referida como hiperglicemia, resultante da diminuição da quantidade de insulina produzida pelo pâncreas (AGUSTIN *et al*., 2024).

Alguns pontos-chave sobre a diabetes são: a diabetes é uma condição crônica e séria que ocorre quando o corpo não pode produzir insulina suficiente ou não pode usar efetivamente a insulina que produz. A diabetes tipo 1 é a principal forma de diabetes na infância, mas pode ocorrer em qualquer idade e não pode ser prevenida. Pessoas com diabetes tipo 1 precisam de insulina para sobreviver. A diabetes tipo 2 representa a grande maioria (mais de 90%) dos casos de diabetes no mundo. Evidências indicam que a diabetes tipo 2 pode ser prevenida ou adiada, e há evidências acumuladas de que a remissão da diabetes tipo 2 pode, às vezes, ser possível. O termo ‘pré-diabetes’ é usado cada vez mais para descrever pessoas com tolerância à glicose prejudicada e/ou glicemia de jejum alterada, indicando um risco maior de desenvolver diabetes tipo 2 e complicações relacionadas. Mulheres grávidas com diabetes gestacional podem ter bebês com grande tamanho para a idade gestacional, aumentando o risco de complicações durante a gravidez e o parto tanto para a mãe quanto para o bebê (IDF DIABETES ATLAS, 2021).

Conforme Rao e Tejomurtula (2024), mais de 10% dos gastos anuais do Serviço Nacional de Saúde (NHS) do Reino Unido são destinados ao manejo da população de pacientes com diabetes, com mais da metade desse valor sendo direcionado ao tratamento de pessoas com complicações graves da doença. Se os custos de tratamento não forem drasticamente reduzidos, o crescente número de pessoas com diabetes no Reino Unido terá um impacto substancial nos gastos do NHS e poderá comprometer o cuidado dos pacientes. O fardo de tratar e prevenir a diabetes e suas consequências aumentará com a considerável elevação da

prevalência da doença, gerando um impacto financeiro significativo nos serviços de saúde. A adoção de diretrizes nacionais na prática clínica padrão tem sido associada ao aumento da prescrição de medicamentos e do uso de tecnologias, bem como à melhora no controle glicêmico ao nível recomendado de HbA1c. O empoderamento e a educação dos pacientes são cruciais para a implementação bem-sucedida de tais estratégias, além do sistema de cuidados de equipes especializadas diversificadas.

A incorporação de análises de dados na indústria de saúde tem avançado significativamente, impulsionada pela demanda por soluções eficientes e eficazes de big data. Knowledge Graphs (KGs) têm demonstrado sua utilidade neste campo, proporcionando uma melhor representação de dados e inferência de conhecimento. No entanto, muitas abordagens existentes carecem de uma taxonomia representativa para a construção de KGs, tornando-as inadequadas. A criação de KGs a partir de grandes volumes de dados textuais, como literatura médica e redes sociais, é crucial para otimizar o cuidado com a saúde, diagnósticos precisos, prevenção de doenças, tratamentos personalizados e melhores tomadas de decisão (ABU-SALIH *et al*., 2023).

O surgimento de big data abriu novas possibilidades e trouxe mudanças significativas em diversas disciplinas, especialmente na área de saúde, que requer análises de dados avançadas e sofisticadas para entender o crescente volume de informações. Embora os dados de saúde sejam frequentemente subutilizados, a extração de conhecimento significativo e acionável desses dados pode ter um impacto positivo no cuidado com o paciente. KGs evoluíram como uma nova forma de representação de conhecimento, integrando entidades do mundo real com relações semânticas, facilitando a conceituação do domínio e a gestão de dados. No entanto, a construção de KGs de qualidade a partir de fontes de dados não estruturados permanece um desafio, exigindo metodologias robustas e avaliações abrangentes para garantir sua eficácia (ABU-SALIH *et al*., 2023).

No campo da computação, a Inteligência Artificial (AI) e, mais especificamente, o Processamento de Linguagem Natural (NLP) e a mineração de dados, estão se tornando cada vez mais relevantes. Essas tecnologias têm o potencial de transformar a maneira como lidamos com grandes volumes de dados e extraímos informações valiosas deles. No contexto da medicina, a aplicação dessas tecnologias pode permitir uma melhor compreensão das doenças e suas causas, levando a melhores estratégias de tratamento e prevenção (YE *et al*., 2020).

A mineração de dados, em particular, está alinhada com a Engenharia do Conhecimento, pois ambas buscam extrair informações úteis de grandes volumes de dados. No caso da Diabetes Mellitus, a mineração de dados pode permitir a identificação de padrões e relações que podem não ser imediatamente aparentes, mas que podem ter implicações significativas para o entendimento e o tratamento da doença (NANDY *et al*., 2021). É fundamental lembrar que o vasto volume de produção científica resultante de pesquisas demanda projetos bem fundamentados para gerenciar e extrair a imensa quantidade de informações disponíveis, as quais precisam ser acessíveis para os pesquisadores em Biomedicina. Nesse contexto, o papel do engenheiro do conhecimento torna-se essencial.

A pesquisa proposta tem o potencial de beneficiar a sociedade de várias maneiras. Primeiro, pode contribuir para a compreensão da Diabetes Mellitus, uma doença que afeta um grande número de pessoas e tem implicações significativas para a saúde pública. Segundo, pode contribuir para o desenvolvimento de modelos de AI e da mineração de dados, campos que têm o potencial de transformar muitos aspectos da sociedade. Finalmente, pode contribuir para a Engenharia do Conhecimento, um campo que busca desenvolver métodos eficazes para a gestão e utilização do conhecimento.

1.5 ESCOPO DA PESQUISA

O escopo desta pesquisa concentra-se na utilização de abstracts de artigos científicos sobre Diabetes Mellitus como fonte primária para a mineração de dados. Serão coletados e processados no mínimo 50.000 artigos, utilizando modelos de NER e RE com o objetivo de formar triplas, fundamentais na construção de KGs. Para a implementação, será empregada a linguagem de programação Python, destacando-se o uso de algoritmos e modelos baseados na arquitetura Transformer e suas variantes, embora a pesquisa não esteja restrita exclusivamente a essas tecnologias.

Optou-se por limitar o corpus a artigos em língua inglesa, considerando que grande parte da literatura médica global é publicada neste idioma, mesmo que existam versões em outros idiomas, como o português. A escolha do inglês também se deve à disponibilidade e à abundância de textos nesse idioma, evitando as dificuldades associadas à coleta de um volume comparável de artigos em português.

Esta decisão assegura a acessibilidade e a relevância dos dados, essenciais para alcançar os objetivos delineados nesta investigação.

1.6 ADERÊNCIA AO PPGEGC

O propósito desta seção é elucidar a conexão e relevância do estudo em questão para o Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC). Esta análise se desdobra em duas vertentes: inicialmente, explora-se a congruência da pesquisa com os princípios e objetivos do PPGEGC, por meio de uma argumentação sólida; subsequente, apresentam-se referências a investigações anteriores que compartilham afinidades temáticas ou metodológicas com o presente trabalho, fortalecendo a evidência da sua pertinência ao programa.

1.6.1 Argumentação

Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC) o qual “tem como objeto de pesquisa e de formação o conhecimento, percebido como fator gerador de valor para a sociedade, e seus processos de criação, explicitação, gestão e disseminação.” (PPGEGC, 2023)

A pesquisa desta dissertação, focada na aplicação de técnicas de machine learning para mineração de dados, foi desenvolvida dentro da área de concentração da Engenharia do Conhecimento (EC), especificamente na linha de pesquisa voltada para a Teoria e Prática em Engenharia do Conhecimento. Esta linha de pesquisa concentra-se nas metodologias e tecnologias da Engenharia do Conhecimento e da Inteligência Computacional, explorando suas interações com a Gestão e a Mídia do Conhecimento (EGC, 2019).

O conhecimento no PPGEGC é definido de forma interdisciplinar, reconhecendo a multiplicidade de visões de mundo e escolas científicas presentes nas diferentes áreas de conhecimento do programa. Na Engenharia de Conhecimento (EC), a visão cognitivista predomina, concebendo o conhecimento como uma entidade armazenável e compartilhável em computadores, bases de dados, arquivos, manuais ou rotinas (PACHECO, 2016).

Para abordar a interdisciplinaridade, foi adotada a definição presente no documento da área interdisciplinar da CAPES (CAPES, 2019):

> Entende-se por Interdisciplinaridade a convergência de duas ou mais áreas do conhecimento, não pertencentes à mesma classe, que contribua para o avanço das fronteiras da ciência e tecnologia, transfira métodos de uma área para outra, gerando novos conhecimentos ou disciplinas e faça surgir um novo profissional, com um perfil distinto dos existentes, com formação básica sólida e integradora, capaz de compreender e solucionar os problemas cada vez mais complexos das sociedades modernas. (CAPES, 2019, p. 9).

A aderência aos EGC se dá pela conexão realizada entre duas áreas distintas que no caso são Saúde e Ciências da Computação, assim como pela diversidade das informações e conhecimento que influenciam na visão do desenvolvimento da doença e consequentemente as variáveis consideradas no momento da análise descritiva e prescritiva. Essas informações e conhecimento devem ser extraídos com a finalidade do seu armazenamento e posterior uso.

Uma das principais atividades da Engenharia do Conhecimento (EC) é a conversão de dados em conhecimento. De acordo com Pacheco (2014), a EC é definida como a “disciplina que se dedica à modelagem de conhecimento e à criação e implementação de sistemas de conhecimento nas organizações”, dito isto, consideremos os seguintes ponto:

- Mineração de Dados: A Engenharia do Conhecimento está fortemente relacionada à mineração de dados, que é uma das etapas fundamentais deste trabalho. Através da aplicação de técnicas avançadas de processamento de linguagem natural e algoritmos de mineração de dados, se busca extrair informações relevantes dos textos científicos sobre a Diabetes Mellitus, identificando padrões e relações úteis para a compreensão e tratamento da doença.
- Natural Language Processing (NLP): O uso de NLP é uma das principais abordagens da Engenharia do Conhecimento para lidar com dados não estruturados, como textos científicos. O NLP permite que computadores possam entender e interpretar a linguagem humana, o que é essencial para a

extração de conhecimento dos textos sobre a Diabetes Mellitus e a construção do Grafo de Conhecimento.

- Representação do Conhecimento: A Engenharia do Conhecimento também se preocupa em como representar o conhecimento adquirido de forma a ser útil e facilmente acessível para os sistemas. A construção do Grafo de Conhecimento é uma representação eficiente para lidar com as complexas relações entre as entidades relacionadas à Diabetes Mellitus, tornando o conhecimento extraído dos textos mais estruturado e organizado.
- Sistemas Inteligentes: O objetivo final da Engenharia do Conhecimento é criar sistemas inteligentes que possam utilizar o conhecimento adquirido para tomar decisões e realizar análises de forma automatizada. Ao utilizar técnicas de mineração de dados e NLP para construir um Grafo de Conhecimento, este trabalho se alinha com a criação de um sistema inteligente capaz de realizar análises e diagnósticos avançados sobre a Diabetes Mellitus.
- Contribuição para a Saúde: A aplicação da Engenharia do Conhecimento para a mineração de dados sobre a Diabetes Mellitus tem grande relevância na área da saúde. Ao extrair conhecimento valioso dos textos científicos, este trabalho pode contribuir para avanços no diagnóstico precoce, tratamentos mais eficazes e uma melhor compreensão da doença, impactando positivamente a qualidade de vida dos pacientes afetados pela Diabetes Mellitus.

1.6.2 Referências Factuais

No contexto da extração de conhecimento a partir de artigos científicos, particularmente no que tange à extração de conhecimento em textos biomédicos e à identificação de entidades e suas inter-relações visando a construção de Grafos de Conhecimento, não se identificaram registros pertinentes no Banco de Teses e Dissertações do PPGEGC. Dessa forma, foram selecionados tópicos que mais se alinham aos objetivos desta dissertação, englobando áreas como mineração de dados, mineração de textos, aprendizado de máquina, descoberta de conhecimento e grafos de conhecimento. A partir desta seleção, 10 trabalhos foram identificados como relevantes para o estudo proposto. Esta seleção é constituída por:

1.6.2.1 Mineração de Textos

- Andrade, Rafael. Um Modelo para recuperação e comunicação do conhecimento em documentos médicos. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011.
- Bovo, Alessandro Botelho. Um Modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011.

1.6.2.2 Mineração de dados

- Souza, Luiz Fernando Spillere de. Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.
- Ghisi, Fernando Benedet. Um método para geração semiautomática de sumários textuais para apoio à disseminação de conhecimento e ao processo decisório em projetos de business intelligence. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2013.

1.6.2.3 Descoberta de conhecimento

- Ribeiro, Alessandro Costa. Modelo de reconhecimento de padrões em ideias usando técnicas de descoberta de conhecimento em textos. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2018.
- Welter, Márcio. Método de identificação de padrões em discurso político a partir da descoberta de conhecimento. (Mestrado) - Programa de Pós-

graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2021.

1.6.2.4 Aprendizado de Máquina

- Piana, Valerio Júnior. Método voltado à recomendação de tratamentos fisioterapêuticos para pacientes com lesão na coluna espinhal por meio de técnicas de aprendizado de máquina. Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2023.
- Spillere de Souza, Luiz Fernando. Modelo de Mineração de Ideias utilizando técnicas de Engenharia de Conhecimento. Tese - Programa de Pós-graduação em Engenharia, Gestão e Mídia do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2022.

1.6.2.5 Grafos de Conhecimento

- Trauer, Eduardo. k-SCAS: framework do sistema de agronegócios de cafés especiais orientados ao conhecimento.: framework do sistema de agronegócios de cafés especiais orientados ao conhecimento. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.
- Silva, Thales do Nascimento da. Um modelo de recomendação de trabalhadores voltado à execução de tarefas no cenário de crowdsourcing. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2022.

Assim, com base nas referências citadas, esta dissertação se alinha ao PPGEGC ao propor e desenvolver um método voltado à representação do conhecimento a partir de documentos sobre diabetes mellitus com a finalidade de construir um grafo de conhecimento..

1.7 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está estruturada em cinco capítulos principais, abrangendo os seguintes tópicos:

- Capítulo 1: Apresenta o tema, os objetivos, as delimitações desta pesquisa, o escopo e a aderência ao PPGEGC.
- Capítulo 2: Apresenta o referencial teórico, abordando os principais tópicos como Diabetes Mellitus, NLP, Topic Modelling, NER, ER, ML, DL, Transformers, BERT, GPT, LLMs e KG.
- Capítulo 3: Apresenta a metodologia de pesquisa adotada e a revisão da literatura.
- Capítulo 4: Apresenta o desenvolvimento da pesquisa, relatando os resultados obtidos a partir da aplicação das metodologias descritas nos capítulos anteriores. Inclui a construção e análise do Grafo de Conhecimento, identificando padrões e relações relevantes relacionados à Diabetes Mellitus, contribuindo para a compreensão e tratamento da doença. Cada seção contém uma breve introdução sobre o modelo, algoritmo ou técnica utilizada, seguida de resultados e uma análise dos principais pontos.
- Capítulo 5: Apresenta as considerações finais da dissertação, incluindo uma síntese das análises de cada experimento e discussão dos principais resultados encontrados, suas implicações para a Engenharia e Gestão do Conhecimento, e para a área biomédica. Também destaca as limitações do estudo e sugestões para futuras pesquisas.
- Capítulo 6: Apresenta as conclusões, limitações e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão introduzidas as principais áreas, tecnologias e abordagens que fundamentam este estudo, destacando a relevância da pesquisa em Diabetes Mellitus e em outros campos da Medicina. Abordaremos a contribuição vital das Ciências da Computação no desenvolvimento de algoritmos e modelos de Inteligência Artificial, essenciais para avanços significativos na área. Discutiremos também o papel crucial da Engenharia do Conhecimento na criação de processos e ferramentas para a extração, armazenamento e publicação do conhecimento, enfatizando sua importância na manipulação de ontologias. Em seguida, apresentaremos uma visão geral sobre o Processamento de Linguagem Natural (NLP), Deep Learning e outras técnicas relevantes para este trabalho. Este capítulo serve como uma fundamentação teórica que prepara o terreno para as investigações e aplicações subsequentes.

2.1 ÁREAS DE CONHECIMENTO ENVOLVIDAS

Biomedicina é a convergência entre a pesquisa biológica e a medicina dentro das ciências da vida. Desde 1970, com o fim da Guerra Fria, a biomedicina passou a ocupar um papel central nos debates públicos. Através dela, sociedades, governos e estados renegociaram os papéis das profissões médicas e científicas, bem como as oportunidades, riscos e limites associados à pesquisa científica e tecnologias médicas. A biomedicina tornou-se um prisma para discutir responsabilidades individuais e profissionais em decisões de vida ou morte e normas que moldam a vida de indivíduos, famílias e comunidades. Este enfoque é essencial para entender os desenvolvimentos recentes na história contemporânea, que poderiam ser distintos de narrativas evolutivas focadas apenas em política, economia ou direito (JANSEN; ROESCH, 2022).

A história da biomedicina é um domínio híbrido que intersecta com muitas outras disciplinas acadêmicas. Desde a década de 1970, historiadores que investigam desenvolvimentos recentes na medicina têm compartilhado abordagens, pressupostos e métodos de investigação com historiadores e sociólogos da ciência e tecnologia. Isso se deve em parte à crescente dependência da medicina em tecnologias, instrumentos e medicamentos, o que torna a demarcação entre

“medicina”, “ciência” e “indústria” mais difícil. Além disso, a “virada prática” na história da ciência deu maior atenção às formas como cientistas e médicos trabalham. Embora esses novos enfoques tenham trazido realizações impressionantes, eles também resultaram em um custo: o afastamento de gerações anteriores de historiadores da medicina pode ter limitado as ambições mais recentes de entender a saúde e a doença na sociedade. Laços mais estreitos com historiadores da ciência e tecnologia e sociólogos da ciência podem ter obscurecido a especificidade da medicina como um domínio baseado na distinção entre o normal e o patológico, diminuindo o interesse dos estudiosos na “clínica” como um local único de produção de conhecimento (LOWY, 2011).

Como objeto para a aplicação desta pesquisa, assunto do qual tratam os papers que são insumo para os experimentos, a diabetes mellitus possui pesquisa que envolve várias áreas interdisciplinares que colaboram para uma compreensão abrangente da doença e o desenvolvimento de tratamentos eficazes. A genética e a biologia molecular são cruciais para identificar predisposições genéticas e os mecanismos celulares subjacentes à diabetes. A imunologia desempenha um papel importante, especialmente no estudo da diabetes tipo 1, que é uma doença autoimune. A farmacologia está envolvida no desenvolvimento e teste de novos medicamentos para controlar os níveis de glicose no sangue e tratar complicações associadas. A neurociência investiga as complicações neurológicas da diabetes, enquanto a nutrição foca em dietas terapêuticas para a gestão da doença. A bioinformática é essencial para analisar grandes volumes de dados genômicos e clínicos, auxiliando na identificação de novos alvos terapêuticos e no desenvolvimento de medicina personalizada. A combinação dessas áreas de pesquisa contribui significativamente para o avanço do conhecimento sobre diabetes e a melhoria do cuidado com os pacientes.

Segundo a Organização Mundial da Saúde (OMS), a Diabetes Mellitus é uma condição metabólica caracterizada pela elevação persistente dos níveis de glicose no sangue, podendo causar graves consequências para a saúde do indivíduo e para a sociedade como um todo (WHO, 2021). Essa doença “silenciosa” tende a evoluir com sintomas leves e se manifestar apenas em estágios mais avançados e irreversíveis (ADA, 2019).

Existem diferentes tipos de diabetes, sendo os mais comuns o tipo 1 e o tipo 2. A diabetes tipo 1 é uma doença autoimune que geralmente se desenvolve na infância ou adolescência e ocorre quando o sistema imunológico ataca e destrói as

células produtoras de insulina no pâncreas (ADA, 2019). Por outro lado, a diabetes tipo 2 é mais comum em adultos e está associada a fatores de risco como obesidade, sedentarismo e histórico familiar da doença (ADA, 2019). A doença progride em diferentes estágios, desde a fase pré-diabetes, onde os níveis de glicose estão acima do normal, mas ainda não atingem o limiar para o diagnóstico de diabetes, até a fase avançada da doença, onde ocorre uma falência completa da produção de insulina e os sintomas se tornam mais graves e irreversíveis (IDF, 2019).

Com a prevalência global estimada em 9,3% da população, sendo ainda maior no Brasil com 7,6%, a Diabetes Mellitus representa um importante desafio para os profissionais de saúde e pesquisadores na busca por soluções mais eficazes de prevenção e tratamento (IDF Diabetes Atlas, 2019). Os sintomas da diabetes podem variar de pessoa para pessoa, mas os mais comuns incluem aumento da sede e da fome, vontade frequente de urinar, fadiga, perda de peso inexplicável e visão turva (IDF, 2019). Além disso, a diabetes pode levar a complicações graves, como doenças cardiovasculares, danos nos rins, neuropatias, problemas nos olhos e dificuldades de cicatrização (OMS, 2021).

O tratamento da diabetes envolve uma combinação de mudanças no estilo de vida, como dieta saudável e prática regular de atividade física, além do uso de medicamentos, como a insulina, para controlar os níveis de glicose no sangue (ADA, 2019). No entanto, apesar dos avanços na medicina, o tratamento e prevenção da diabetes ainda representam desafios significativos. A adesão ao tratamento por parte dos pacientes e a educação sobre a doença são aspectos fundamentais para o controle efetivo da diabetes e para a redução de suas complicações (OMS, 2021).

A partir da análise de textos científicos sobre a doença, torna-se possível obter insights relevantes para melhor compreensão e abordagem terapêutica, contribuindo assim para a melhoria da assistência médica e o enfrentamento desse problema de saúde pública. Essas informações são pertinentes para o entendimento do objeto de aplicação desta pesquisa, que envolve a extração e representação de conhecimento de artigos científicos sobre diabetes mellitus utilizando técnicas de NER (*Named Entity Recognition*) e ER (*Entity Relation*) para a criação de triplas que representam um KG (*Knowledge Graph*).

As **Ciências da Computação** são dedicadas ao estudo dos fundamentos teóricos e práticos da computação e das tecnologias da informação, abrangendo áreas como processamento de dados, algoritmos, sistemas de informação, e

inteligência artificial. Segundo Harel e Feldman (2004), a Ciência da Computação é uma ciência em constante evolução, impulsionada pelo rápido avanço da tecnologia e pela crescente demanda por soluções computacionais inovadoras em diversas áreas.

Dentro da Ciência da Computação, a Inteligência Artificial (AI) tem desempenhado um papel fundamental nas últimas décadas. A AI busca desenvolver sistemas computacionais capazes de realizar tarefas que antes eram exclusivas de seres humanos, como aprendizado, raciocínio, planejamento e tomada de decisão (RUSSELL; NORVIG, 2020). Ela se baseia em algoritmos e modelos que permitem que os computadores aprendam e melhorem seu desempenho ao longo do tempo, a partir da análise de grandes volumes de dados (LECUN *et al*., 2015).

A AI tem sido amplamente aplicada em diversas áreas, incluindo a medicina, onde contribui para o desenvolvimento de sistemas de diagnóstico mais precisos e eficientes (ESTEVA *et al*., 2017). No contexto desta pesquisa sobre Diabetes Mellitus, a AI tem um papel relevante na análise de textos científicos e na extração de conhecimento a partir desses dados não estruturados. Segundo Rothman (2021), a relevância do Processamento de Linguagem Natural (NLP) e dos transformers reside na capacidade de entender e processar a complexidade do pensamento humano expresso na linguagem escrita. A linguagem humana é o meio mais preciso para transmitir uma grande quantidade de conhecimento, e o NLP, juntamente com os transformers, possibilitam a conversão do discurso em texto e a extração significativa de informações de grandes conjuntos de dados não estruturados, como os textos científicos sobre Diabetes Mellitus. Essa abordagem é essencial para uma compreensão aprofundada da doença, seus fatores de risco e possíveis estratégias de tratamento e prevenção.

Segundo Studer, Benjamins e Fensel (1998), a **Engenharia do Conhecimento (KE)** é uma disciplina voltada para o desenvolvimento e manutenção de processos automatizados que geram e aplicam conhecimento. Pacheco e Todesco (2012) corroboram essa visão, afirmando que a Engenharia do Conhecimento se dedica à explicitação, formalização, representação e operacionalização do conhecimento em atividades intensivas em conhecimento.

O grande volume de informações e a necessidade de automação na Mineração de Dados na extração de conhecimento de papers na área da Biomedicina só é viável com o uso de técnicas e planejamento da Engenharia do Conhecimento, além dos algoritmos e modelos da Ciência da Computação.

A Engenharia do Conhecimento destaca-se especialmente no uso de ontologias, essenciais para o sucesso em projetos de Mineração de Dados que envolvem grandes volumes de informação científica. Segundo Allen, Stork e Groth (2023), as ontologias são cruciais na organização e estruturação do conhecimento, definindo termos e relações entre conceitos em um domínio específico. Elas promovem a interoperabilidade e integração de dados heterogêneos, estabelecendo uma taxonomia formal e semântica que possibilita uma compreensão mais aprofundada e precisa dos dados. Dessa forma, são instrumentos valiosos na descoberta de padrões, extração de informações relevantes e facilitação de tomadas de decisão informadas em contextos de Mineração de Dados.

2.2 ANÁLISE EXPLORATÓRIA DE DADOS E MINERAÇÃO DE DADOS

Análise Exploratória de Dados (do inglês Exploratory Data Analysis - EDA) é uma abordagem inicial de análise de dados que utiliza principalmente técnicas gráficas e estatísticas para resumir suas principais características. EDA é fundamental para compreender melhor os dados, identificar padrões, detectar anomalias, testar hipóteses e verificar pressupostos. Ao aplicar EDA, os analistas podem interpretar os resultados e tomar decisões informadas sobre quais modelos e técnicas de mineração de dados devem ser aplicados posteriormente. Através da visualização dos dados e das estatísticas descritivas, EDA facilita a compreensão dos dados antes da aplicação de técnicas mais avançadas (PURI *et al*., 2023).

Mineração de Dados (do inglês Data mining - DM), no seu termo genérico, também conhecido como text mining quando aplicado a textos, ou abstract-data mining ao ser aplicado a abstracts de papers, é uma área de pesquisa amplamente reconhecida entre diversas disciplinas acadêmicas. Seu objetivo principal é a identificação de padrões e tendências em grandes volumes de dados, transformando dados brutos em informações úteis e acionáveis. Um exemplo prático é a análise de publicações científicas altamente citadas, onde técnicas de data mining são aplicadas para extrair e categorizar informações relevantes. A análise revelou que áreas de pesquisa como GIS, análise bibliométrica e ciência da informação são frequentemente associadas ao data mining (JAYASEKARA; ABU, 2018).

O DM, no caso, Text Mining, envolve a descoberta automática de novas informações previamente desconhecidas, extraídas de diversos recursos escritos. Este processo abrange a identificação, extração e análise de padrões textuais para a

descoberta de conhecimento. Por exemplo, ao analisar publicações científicas, o Text Mining pode revelar áreas de pesquisa inexploradas e ajudar provedores de bancos de dados a melhorar a indexação de artigos. Assim, Text Mining permite que pesquisadores analisem grandes volumes de literatura científica de maneira eficiente, identificando comportamentos e padrões que podem ser úteis em diversas aplicações, desde a previsão de comportamentos em redes sociais até a melhoria das condições de saúde pública por meio da informática médica (JAYASEKARA; ABU, 2018).

O Text Mining, também conhecido como KDT (Knowledge Discovery in Textual Databases) ou mineração de dados textuais, é definido como o processo de extração de padrões ou conhecimento previamente desconhecidos, compreensíveis, potenciais e práticos de grandes coleções de dados textuais não estruturados. Este campo de pesquisa é composto por três etapas principais: pré-processamento do texto, operações de mineração de texto e pós-processamento. O pré-processamento envolve tarefas como seleção de dados, classificação e extração de características, preparando os documentos para diferentes propósitos de mineração. As operações de mineração de texto incluem clustering, descoberta de regras de associação, análise de tendências, descoberta de padrões e outros algoritmos de descoberta de conhecimento. O pós-processamento manipula os dados ou conhecimentos obtidos, incluindo a avaliação, seleção, interpretação e visualização do conhecimento (ZHANG; CHEN; LIU, 2015).

Com o desenvolvimento rápido da tecnologia da informação e a aplicação extensiva da internet, a mineração de texto tornou-se uma área de pesquisa essencial. A internet gera grandes quantidades de dados textuais não estruturados, como blogs, postagens em fóruns e documentação técnica. Esses dados contêm muita informação valiosa, mas são extremamente difíceis de processar devido ao seu volume e diversidade. A demanda por análise de dados textuais está crescendo, e a mineração de texto surgiu como uma solução para adquirir a informação necessária a partir desses dados não estruturados. A mineração de texto é considerada mais valiosa comercialmente do que a mineração de dados porque uma grande parte das informações de uma empresa está contida em documentos textuais. No entanto, a mineração de texto é mais complexa devido à natureza não estruturada dos dados textuais. Ela envolve áreas como inteligência artificial, aprendizado de máquina, estatísticas matemáticas e sistemas de banco de dados (ZHANG; CHEN; LIU, 2015).

2.3 INTELIGÊNCIA ARTIFICIAL

Conforme Russel e Norvig (2021), a Inteligência Artificial (do inglês Artificial Intelligence - AI) é um campo que busca construir máquinas capazes de comportamento inteligente. Envolve a criação de sistemas que podem perceber, entender, prever e manipular o mundo. A AI é um campo em rápido crescimento, já gerando impacto econômico significativo, com vastos horizontes intelectuais ainda a serem explorados. Quatro Abordagens para AI:

- Agir como Humano. A Abordagem do Teste de Turing: Proposto por Alan Turing em 1950, o teste de Turing avalia a capacidade de uma máquina em exibir comportamento inteligente equivalente ao de um humano. Para passar esse teste, uma máquina precisa de capacidades em processamento de linguagem natural, representação de conhecimento, raciocínio automatizado e aprendizado de máquina. Uma versão estendida, o teste de Turing total, inclui visão computacional, reconhecimento de fala e robótica.
- Pensar como Humano. A Abordagem de Modelagem Cognitiva: Esta abordagem foca em entender como os humanos pensam. Envolve introspecção, experimentos psicológicos e imagens cerebrais para desenvolver programas que possam replicar processos de pensamento humano. Modelando a cognição humana, a AI busca imitar as operações mentais dos humanos.
- Pensar Racionalmente. A Abordagem das “Leis do Pensamento”: Baseada nas leis do pensamento idealizadas por filósofos, essa abordagem envolve a criação de algoritmos que possam simular o pensamento racional. O objetivo é criar sistemas que tomem decisões baseadas em raciocínio lógico.
- Agir Racionalmente. A Abordagem do Agente Racional: Um agente racional age para alcançar o melhor resultado ou, quando há incerteza, o melhor resultado esperado. Esta abordagem enfatiza o design de agentes que possam tomar decisões e agir autonomamente para alcançar objetivos específicos, utilizando campos como estatística, teoria de controle e economia.

Ao perseguir essas abordagens diversas, os pesquisadores de AI visam desenvolver sistemas que possam realizar uma ampla gama de tarefas, desde jogar

xadrez até diagnosticar doenças, ampliando nossa compreensão da inteligência e expandindo suas aplicações práticas em vários domínios.

Aprendizado de Máquina (do inglês Machine Learning - ML) é uma subárea da AI focada no desenvolvimento de algoritmos e modelos que permitem aos sistemas aprender a partir de dados. Esses modelos fazem previsões ou tomam decisões com base em dados históricos, sem programação explícita para essas tarefas. A construção de modelos que podem se adaptar e melhorar com a experiência é essencial para criar sistemas inteligentes, tornando o aprendizado de máquina uma parte fundamental da computação inteligente na era atual (SARKER, 2021).

Em outros termos ML é uma abordagem científica que utiliza a probabilidade em vez da lógica booliana, aprendizado automático em vez de codificação manual, e resultados experimentais em vez de alegações filosóficas. Em vez de criar teorias totalmente novas, os pesquisadores de ML costumam construir sobre teorias existentes, baseando suas alegações em teoremas rigorosos ou metodologias experimentais sólidas (COHEN, 1995). Este campo promoveu o uso de conjuntos de problemas de benchmark compartilhados para demonstrar progresso, como o repositório da UC Irvine para conjuntos de dados de aprendizado de máquina e a competição de planejamento internacional para algoritmos de planejamento (RUSSEL; NORVIG, 2021).

A AI, inicialmente separada de outros campos como a teoria de controle e a estatística, começou a adotar os resultados positivos dessas áreas. McAllester (1998) destacou que o aprendizado de máquina não deve ser isolado da teoria da informação, e o raciocínio incerto não deve ser separado da modelagem estocástica. Um exemplo ilustrativo é o campo de reconhecimento de fala, que nos anos 1980 passou a dominar com modelos ocultos de Markov (do inglês Hidden Markov Models - HMMs). Esses modelos são baseados em uma teoria matemática rigorosa e são treinados em grandes corpora de dados reais, melhorando constantemente seu desempenho em testes cegos rigorosos (RUSSEL; NORVIG, 2021).

A aceitação de dados, modelagem estatística, otimização e aprendizado de máquina pela AI resultou na reunificação gradual de subcampos como visão computacional, robótica, reconhecimento de fala, sistemas multiagente e processamento de linguagem natural. Essa reintegração trouxe benefícios significativos tanto em termos de aplicações práticas, como a expansão do uso de robôs, quanto em uma melhor compreensão teórica dos problemas centrais da AI.

Aprendizado Profundo (do inglês Deep Learning – DL) é uma ampla família de técnicas de aprendizado de máquina onde as hipóteses assumem a forma de circuitos algébricos complexos com forças de conexão ajustáveis. O termo “deep” se refere ao fato de que esses circuitos são tipicamente organizados em muitas camadas, significando que os caminhos de computação dos inputs aos outputs possuem muitos passos. DL é amplamente utilizado em aplicações como reconhecimento de objetos visuais, tradução automática, reconhecimento e síntese de fala, e síntese de imagens, além de desempenhar um papel significativo em aplicações de aprendizado por reforço (RUSSEL; NORVIG, 2021).

As origens do DL estão no trabalho inicial que tentou modelar redes de neurônios no cérebro (McCulloch e Pitts, 1943) com “circuitos computacionais”, que são representações matemáticas. Por essa razão, as redes treinadas por métodos de deep learning são frequentemente chamadas de redes neurais, apesar da semelhança com células e estruturas neurais reais ser superficial (RUSSEL; NORVIG, 2021).

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (do inglês Natural Language Processing - NLP) é definido como a capacidade computacional de reconhecer e processar a linguagem utilizada pelos seres humanos. Uma aplicação básica dessa tecnologia pode ser exemplificada pela identificação e remoção de palavras específicas em um texto, uma entre as diversas funcionalidades possíveis dentro deste campo (SRINIVASA-DESIKAN, 2023). De acordo com Allhyari *et al*. (2017), o NLP é um subcampo da ciência da computação, inteligência artificial (IA) e linguística, amplamente aplicado na mineração de texto. Lauriola, Lavelli e Aiolli (2021) definem NLP como um ramo da IA repleto de tarefas complexas e desafiadoras relacionadas à linguagem, como tradução automática, respostas a perguntas e resumo de textos.

O campo do processamento de linguagem natural é vasto e diversificado, englobando desde teorias e modelos formais não baseados em IA, como autômatos de estado finito e sistemas de regras, até algoritmos complexos de aprendizado de máquina e IA, como redes de Markov ocultas e campos aleatórios condicionais. Essas abordagens são cruciais para o tratamento de ambiguidades e a análise profunda de dados em NLP, refletindo a evolução da área em direção a métodos

que podem aprender e adaptar-se dinamicamente a partir dos dados disponíveis (JURAFSKY; MARTIN, 2008).

O NLP é indispensável na era da IA, sendo utilizado em aplicações de interação homem-computador (HCI) em softwares de última geração. Esta tecnologia permite estabelecer interfaces entre máquinas e humanos por meio do entendimento de linguagens humanas, facilitando diálogos mais naturais (CHAO *et al*., 2021). A linguagem natural é uma fonte rica de informações para muitas aplicações, mas é discreta e esparsa, tornando-se uma fonte de dados desafiadora. Para que o texto seja utilizável como dado de entrada, deve-se transformá-lo em uma representação adequada, geralmente um vetor de números. As representações vetoriais de texto podem ser construídas de várias maneiras diferentes (Babić; Martinčić-Ipšić; Meštrović, 2020).

Dentro do contexto desta pesquisa, o NLP é uma tecnologia essencial para o avanço da assistência à saúde, pois permite converter informações relevantes contidas em textos em dados estruturados, utilizáveis por processos computacionais que visam aprimorar o atendimento ao paciente e a prática médica (FRIEDMAN; RINDFLESCH; CORN, 2013). A importância do NLP na área da saúde motivou a National Library of Medicine (NLM) a patrocinar um workshop para revisar o estado da arte do NLP em biomedicina e no domínio geral da linguagem.

De modo geral, antes de qualquer processamento de linguagem natural de um texto, este deve ser normalizado (JURAFSKY; MARTIN, 2021). Entre as funções importantes no tratamento do texto, citam-se:

- Tokenization: Separa o texto em palavras e símbolos, incluindo pontuação.
- Token Frequency/Count: Realiza a contagem de ocorrência de tokens.
- Stopword Removal: Remove tokens irrelevantes.
- N-grams: Sequência de palavras onde n indica o número de palavras. Um bigrama é a sequência de duas palavras, trigrama é uma sequência de três palavras e assim por diante.
- Stemming: Consiste em reduzir a palavra ao seu radical.
- Lemmatization: Permite reduzir a palavra à sua forma canônica, levando em conta sua classe gramatical.
- Part-of-Speech Tagging: Classifica cada ocorrência de uma palavra em uma frase como, por exemplo, um substantivo, adjetivo ou um verbo.

- Named Entity Recognition (NER): Busca extrair e classificar as entidades mencionadas em um texto escrito em linguagem natural.
- Relation Extraction (RE): Identifica e classifica os relacionamentos entre entidades no texto.

A aplicação de cada função depende do objetivo desejado, pois nem toda atividade de NLP requer todas as funções ou mesmo parte delas. Cada projeto de NLP seleciona as etapas mais relevantes para alcançar resultados específicos, otimizando o processamento e a análise dos dados textuais.

Dentro do NLP com abordagem de AI, o copus assume um papel de grande importância devido a ser um grande e estruturado conjunto de textos, essencial para o início da análise de texto. Exemplos gratuitos de corpora incluem o Open American National Corpus e o British National Corpus. A Wikipedia oferece uma lista útil dos maiores corpora disponíveis. Estes não se limitam ao idioma inglês, existindo vários corpora em línguas europeias e asiáticas, com esforços contínuos em todo o mundo para criar corpora para a maioria das línguas. Laboratórios de pesquisa universitários também são uma fonte valiosa para obter corpora; um dos mais icônicos corpora da língua inglesa, o Brown Corpus, foi elaborado na Brown University. Isso ressalta que, além de modelos, algoritmos e técnicas, a qualidade dos dados de treinamento é crucial para o NLP baseado em AI (SRINIVASA-DESIKAN, 2023).

A construção de corpora é fundamental para o sucesso do Processamento de Linguagem Natural (NLP) na biomedicina. Um corpora bem alinhado a uma determinada área de conhecimento, como a biomedicina, facilita a estruturação de informações clínicas que, de outra forma, estariam dispersas em textos não estruturados. Desde a década de 1960, os sistemas de NLP têm mostrado que é viável transformar informações clínicas textuais em dados estruturados que podem melhorar o cuidado com o paciente e a medicina (FRIEDMAN; RINDFLESCH; CORN, 2013).

Na biomedicina, recursos como o SNOMED, o Sistema de Linguagem Médica Unificada (UMLS) e bases de dados como PubMed desempenham um papel crucial ao fornecer terminologias e dados estruturados para o desenvolvimento de sistemas de NLP. Esses sistemas são projetados para reconhecer entidades específicas no texto, como procedimentos médicos, medicamentos, doenças e substâncias do corpo, facilitando a extração de informações pertinentes e a

descoberta de relações entre essas entidades, essencial para a criação de Knowledge Graphs (KGS) (PULUGU *et al*., 2024).

A utilização de corpora específicos e bem anotados é indispensável para o treinamento e avaliação de modelos de NLP, especialmente na biomedicina, onde a precisão e a especificidade das informações são críticas. O alinhamento do corpora ao domínio de conhecimento garante que os modelos aprendam as nuances e particularidades do vocabulário médico, aumentando a eficiência na extração de conhecimento e contribuindo para uma melhor compreensão e gestão de condições de saúde complexas.

O corpus de texto é um grande conjunto estruturado de textos essencial para treinar modelos de NLP. Corpora como o Open American National Corpus e o British National Corpus fornecem recursos valiosos para a análise de texto. Na biomedicina, corpora especializados alinhados ao conhecimento do domínio, como SNOMED e o Unified Medical Language System (UMLS), são cruciais. Esses corpora permitem o treinamento preciso de modelos de NLP para tarefas como reconhecimento de entidades e extração de informações, melhorando significativamente a qualidade e relevância dos modelos em aplicações médicas (FRIEDMAN; RINDFLESCH; CORN, 2013).

O NLP possui grande aplicabilidade em várias áreas, sendo uma ferramenta essencial para a análise e compreensão de textos. Algumas das principais aplicações do NLP incluem:

- Tradução Automática: Permite traduzir frases de uma língua para outra utilizando motores estatísticos, como o Google Translate, garantindo que o significado e a gramática sejam mantidos. Técnicas modernas envolvem redes neurais artificiais e Deep Learning para melhorar a qualidade da tradução (Tillmann *et al*., 1997).
- Categorização de Textos: Utilizada para classificar grandes volumes de dados, como documentos oficiais e notícias, em categorias predefinidas. Este processo é fundamental para filtros de spam de e-mail, que categorizam mensagens para identificar spam (Hayes, 1992).
- Filtragem de Spam: Baseada em técnicas de categorização de texto, a filtragem de spam utiliza aprendizado de máquina para identificar e bloquear e-mails indesejados, utilizando modelos como Naive Bayes e máquinas de vetor de suporte (Androutsopoulos *et al*., 2000).

- Extração de Informações: Envolve identificar e extrair frases ou entidades relevantes de dados textuais, como nomes, locais, eventos e datas. Técnicas como modelos ocultos de Markov (HMMs) são usadas para melhorar a precisão e eficiência das buscas específicas (McCallum e Nigam, 1998).
- Sumarização de Textos: Essencial na era da sobrecarga de informações, permite criar resumos de grandes volumes de texto, mantendo seu significado original. Isso é útil para entender rapidamente informações importantes de um conjunto de dados extenso (Zajic *et al*., 2008).
- Sistemas de Diálogo: Implementados em assistentes virtuais como Siri, Alexa e Google Assistant, esses sistemas utilizam todos os níveis de linguagem para interagir com os usuários de maneira natural e eficiente (Liddy, 2001).
- Aplicações Médicas: O NLP é amplamente utilizado na medicina para extrair e resumir informações de registros médicos, facilitando a identificação de efeitos colaterais de medicamentos e outras informações críticas para a saúde (Glasgow *et al*., 1998).

Os **embeddings** são uma forma avançada de representação de palavras que conecta o conhecimento humano de maneira significativa à compreensão da máquina (BIRUNDA; DEVI, 2022). Técnicas de embeddings representam palavras como vetores numéricos em um espaço n-dimensional, onde a posição dos vetores transmite a semântica das palavras. Palavras com significados semelhantes possuem vetores próximos uns dos outros (RIZKALLAH; ATIYA; SHAHEEN, 2022). Isso significa que os vetores codificam o significado das palavras e suas relações semânticas com outras palavras do corpus.

Representações tradicionais, como Bag-of-Words (BOW) e TF-IDF, têm limitações, pois ignoram a ordem das palavras e suas semelhanças sintáticas e semânticas (LI *et al*., 2018; KALYAN; SANGEETHA, 2020). Para superar essas limitações, Mikolov *et al*. (2013a) propuseram o uso de embeddings para capturar relações sintáticas e semânticas, melhorando a qualidade dos vetores e acelerando o treinamento. Os embeddings mapeiam palavras em representações vetoriais densas, minimizando problemas como a dimensionalidade alta e a falta de informações nas representações tradicionais (KALYAN; SANGEETHA, 2020).

Os embeddings podem ser divididos em várias categorias: embeddings de texto e de conceito, além de embeddings de caracteres, palavras, frases, sentenças e documentos. Word embeddings, como Word2Vec, GloVe e FastText, são vetores

densos aprendidos a partir de grandes coleções de texto, observando o contexto de cada palavra (CAMACHO-COLLADOS; PILEHVAR, 2018). Métodos mais recentes, como BERT e ELMo, geram representações altamente específicas para cada contexto, tornando a análise semântica mais precisa em casos de ambiguidade (AKBIK *et al*., 2019; HAJ-YAHIA; SIEG; DELERIS, 2019).

Essas técnicas têm uma ampla aplicação em processamento de linguagem natural, aprendizado de máquina e sistemas de recomendação, proporcionando um espaço vetorial de baixa dimensionalidade que contém informações úteis para várias tarefas, como classificação de texto, agrupamento de documentos e recomendação de itens (BIRUNDA; DEVI, 2022; RIZKALLAH; ATIYA; SHAHEEN, 2022). Em suma, os embeddings são fundamentais para capturar relações semânticas e sintáticas, melhorando significativamente a qualidade e a eficiência das aplicações de NLP.

Entendimento de Linguagem Natural (do inglês Natural Language Understanding NLU), para Dahl (2023), é uma subárea do NLP que estrutura a linguagem para que sistemas computacionais possam processá-la e executar aplicações úteis. NLU é valiosa para aplicações práticas, onde os benefícios justificam os custos de desenvolvimento e manutenção. Envolve avaliar o desempenho do sistema, melhorar os resultados e implantar aplicações, sendo adequada para projetos acadêmicos, demonstrações, provas de conceito ou questões de pesquisa avançada. A diferença entre NLU e NLP está na especificidade de suas funções: enquanto NLP abrange uma variedade de técnicas para manipular a linguagem natural, como tradução automática e análise de sentimentos, NLU foca especificamente na compreensão e interpretação da linguagem, estruturando os dados textuais de forma que possam ser utilizados de maneira significativa por sistemas computacionais.

2.4.1 Modelagem de Tópicos

A Modelagem de Tópicos (do inglês Topic Modelling) é uma técnica de aprendizado não supervisionado em NLP que identifica padrões e agrupa palavras em tópicos dentro de grandes conjuntos de dados textuais. Esses modelos ajudam a descobrir temas latentes nos documentos, fornecendo uma visão geral dos principais tópicos abordados (IBM, 2024). Williams *et al*. (2024) definem Topic Modeling como uma técnica de mineração de texto que identifica temas significativos em um conjunto de documentos. O resultado geralmente é um conjunto de tópicos

composto por tokens isolados que frequentemente coocorrem nos documentos. Embora esses tokens ajudem a inferir o significado dos tópicos, sua interpretação pode ser difícil para os humanos.

O ***Latent Dirichlet Allocation*** (LDA) é um modelo probabilístico generativo para coleções de dados discretos, como corpora de textos, baseado em um modelo bayesiano hierárquico de três níveis, no qual cada item de uma coleção é modelado como uma mistura finita sobre um conjunto subjacente de tópicos. Cada tópico, por sua vez, é modelado como uma mistura infinita sobre um conjunto subjacente de probabilidades de tópicos. No contexto da modelagem de textos, as probabilidades dos tópicos fornecem uma representação explícita de um documento, fazendo com que o LDA seja visto como um método do tipo *Bag of Word* (CBOW) para modelagem de linguagem e redução de dimensionalidade (HOLZINGER *et al*., 2014).

O princípio básico do LDA é que os documentos podem ser representados como uma mistura de tópicos latentes, que são representados por uma distribuição de palavras (BLEI *et al*., 2003). Similar a ***Latent Semantic Analysis*** (LSA) e ***Probabilistic Latent Semantic Analysis*** (pLSA), o número de tópicos latentes usados no modelo deve ser fixado *a priori*. No entanto, em contraste com métodos como o LSA, que utilizam métodos de álgebra linear, o LDA utiliza métodos probabilísticos para inferir a mistura de tópicos e palavras. O processo gerativo assumido para cada palavra em um documento dentro de um corpus envolve os seguintes passos: escolher N, o número de palavras de um documento, estimado por uma distribuição de Poisson; escolher uma mistura de tópicos θ de acordo com uma distribuição de Dirichlet sobre α, um prior de Dirichlet sobre todos os documentos; e cada uma das N palavras é selecionada primeiro escolhendo um tópico representado como uma variável aleatória multinomial z, e segundo escolhendo uma palavra de uma probabilidade multinomial condicionada ao tópico z (BLEI *et al*., 2003).

A meta da inferência no modelo LDA é encontrar os valores de φ, a probabilidade de uma palavra w ocorrer em um tópico z, e θ, a distribuição de tópicos sobre um documento. Existem vários algoritmos propostos para resolver o problema de inferência no LDA, incluindo um algoritmo de Expectation-Maximization variacional, um algoritmo de Expectation-Propagation, e um algoritmo de amostragem de Gibbs colapsada (GRIFFITHS; STEYVERS, 2004).

A **Non-negative Matrix Factorization** (NMF) é outra técnica comumente utilizada, e é definda por Liu (2019) como sendo uma técnica baseada em álgebra linear que decompõe uma matriz de entrada V em duas matrizes menores, W e H, todas com valores não-negativos. Em NLP, a matriz de entrada V representa a contagem de termos ou a matriz TF-IDF de documentos. A matriz W é a matriz de características, onde cada linha representa um tópico e o peso de cada termo nesse tópico. A matriz H é a matriz de coeficientes, indicando a relevância de cada tópico para cada documento. Já para Barber (2012), a NMF é uma técnica de análise que se diferencia por impor a restrição de não negatividade nas entradas das matrizes base e de peso. Essa característica faz com que o NMF seja uma variação da Análise Fatorial, particularmente útil quando se deseja garantir que os componentes gerados sejam interpretáveis em termos de contribuições positivas. O NMF também é considerado uma generalização do pLSA, mas sem a necessidade de normalização das bases, o que pode resultar em convergência mais lenta durante o treinamento.

Por sua parte, **BERTopic** é um modelo de última geração para modelagem de tópicos que utiliza embeddings de palavras para representar os segmentos de texto, possibilitando a localização de palavras, frases ou documentos semanticamente semelhantes em proximidade espacial. Diferente de métodos tradicionais como a LDA, que requer a especificação do número de tópicos antecipadamente, BERTopic emprega o *Hierarchical Dirichlet Process* (HDP) para agrupar vetores em tópicos, oferecendo uma descrição mais detalhada e contextualizada dos tópicos extraídos. Este método é particularmente eficaz em capturar sutilezas e complexidades dentro dos conjuntos de dados, proporcionando uma compreensão mais rica e matizada dos dados textuais. BERTopic destaca-se por sua capacidade de capturar relações semânticas profundas entre textos, tornando-o altamente eficaz em identificar e agrupar tópicos de maneira contextualizada e detalhada. A implementação do BERTopic apresenta uma visão multidimensional dos tópicos, refletindo uma camada mais profunda de entendimento temático e melhorando a interpretabilidade das descrições de tópicos em comparação com abordagens tradicionais (WILLIAMS *et al*., 2024).

Existem vários métodos para distribuir textos em tópicos, sendo os mais tradicionais a pLSA, LSA e a LDA. Recentemente, algoritmos avançados como BERTopic e Top2Vec têm atraído atenção no campo do NLP. Esses métodos modernos oferecem vantagens adicionais e são avaliados quanto à sua capacidade

de melhorar a interpretação dos modelos de tópicos. A aplicabilidade de tais métodos na biomedicina, por exemplo, pode ajudar na descoberta de padrões em grandes volumes de literatura científica, auxiliando na identificação de novas áreas de pesquisa e aprimorando a análise de dados biomédicos (WILLIAMS *et al*., 2024).

A modelagem de tópicos possui grande valor no contexto desta pesquisa, não apenas como um método para a descoberta dos tópicos abrangidos pela literatura científica que serve de insumo para a criação do KG, mas também para a extração de informações importantes que podem enriquecer cada uma das entidades que fazem parte do KG em questão. Ao identificar temas recorrentes e relevantes nos textos científicos, a modelagem de tópicos permite a integração de informações adicionais às entidades do grafo, proporcionando uma visão mais completa e detalhada. Esse processo de enriquecimento das entidades pode levar à inferência de novos conhecimentos, facilitando a descoberta de relações e padrões que, de outra forma, poderiam passar despercebidos. Assim, a aplicação de técnicas de modelagem de tópicos contribui significativamente para o representação e a descoberta de conhecimento dentro do domínio da pesquisa biomédica, especialmente no estudo da diabetes mellitus.

2.4.2 Reconhecimento de Entidade Nomeada e Extração de Relação

Reconhecimento de Entidade Nomeada (do inglês Named Entity Recognition - NER) é uma tarefa de extração de texto que envolve a identificação e classificação de entidades nomeadas, como lugares, pessoas, datas e horas. NER é uma das principais tarefas de NLP, assim como a etiquetagem de partes do discurso (POS-tagging). Entidades nomeadas são objetos do mundo real com nomes próprios, como “França” (GPE - Entidade Geopolítica), “Donald Trump” (PER - Pessoa) e “Twitter” (ORG - Organização). A ferramenta spaCy, uma biblioteca Python, utiliza modelos estatísticos para reconhecer diferentes tipos de entidades em documentos, embora possa necessitar de ajustes específicos conforme o caso de uso. Além disso, a Named Entity Disambiguation (NED) é importante para diferenciar entre entidades com o mesmo nome, mas em contextos diferentes, como “Rome” a cidade versus “Rome” o artista (SRINIVASA-DESIKAN, 2018).

NER apresenta desafios adicionais em aplicações de extração de informação, como identificar referências a pessoas reais, organizações e locais em textos longos, como artigos de jornal, onde um mesmo indivíduo pode ser referido

de várias formas, como Joe Biden sendo chamado de presidente, Sr. Biden, ele, ou ex-vice-presidente. É crucial evitar interpretações erradas, como confundir Dr. Biden com Joe Biden, sendo na verdade uma referência à sua esposa. Modelos BERT, especialmente os “cased”, são úteis para NER, pois consideram a capitalização dos nomes próprios (DAHL, 2023).

Para Devarakonda, Raja e Xu (2024a), existem alguns desafios envolvidos na tarefa de NER, centrados principalmente na detecção e classificação das entidades:

- Detecção de Intervalo (Span Detection): Este problema se refere à identificação da localização exata no texto e à sequência de caracteres que compõem a entidade. Isso geralmente é feito determinando o índice do primeiro caractere da entidade e seu comprimento. Por exemplo, se a entidade for “câncer de mama”, o sistema deve identificar onde essa frase começa e quantos caracteres ela contém.
- Identificação do Tipo de Entidade (Entity Type Identification): Após detectar a entidade, é necessário classificá-la em uma categoria específica, como problema médico, teste ou tratamento. Por exemplo, “câncer de mama” seria classificado como um problema médico. Além disso, uma vez que a entidade é reconhecida, pode ser necessário mapeá-la para uma entrada padronizada em um sistema de codificação, como o ICD-9-CM (Classificação Internacional de Doenças) ou em um tesauro como o UMLS (Sistema de Linguagem Médica Unificada). Esse processo de mapeamento é conhecido como “entity linking” ou normalização de conceitos.

Devarakonda, Raja e Xu (2024a) também descrevem algumas complexidades adicionais, focadas na variabilidade da comunicação humana e na complexidade da escrita, que podem surgir durante a aplicação de NER:

- Entidades Aninhadas: Às vezes, uma entidade pode estar completamente dentro do intervalo de outra. Por exemplo, na frase “com uma história de três anos de dormência bilateral das mãos”, “dormência bilateral das mãos” é um problema médico, enquanto “mão” é uma entidade interna do tipo parte do corpo.

- Entidades Disjuntas: Em alguns casos, uma entidade pode não ser contígua, ou seja, pode ser separada por outras palavras. Por exemplo, na frase “O átrio esquerdo está moderadamente dilatado”, a expressão “átrio esquerdo dilatado” não é uma sequência contínua.
- Entidades Distribuídas: Às vezes, duas ou mais entidades podem compartilhar um prefixo comum. Por exemplo, na frase “A ressonância magnética mostrou processo neoplásico no fígado e no cólon”, existem duas entidades distintas: “processo neoplásico no fígado” e “processo neoplásico no cólon”, que compartilham o prefixo “processo neoplásico”.

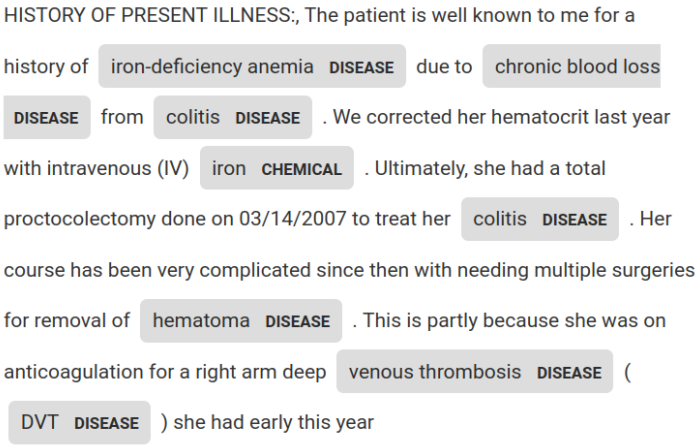
A aplicação de NER pode seguir cinco abordagens principais, conforme descrito por Devarakonda, Raja e Xu (2024a):

- Métodos Baseados em Dicionário: Esta abordagem inicial envolve o uso de um dicionário que contém conceitos médicos e seus sinônimos, como o UMLS. O texto de entrada é segmentado em tokens, que são então normalizados (caso, variações ortográficas, etc.) e comparados com as entradas do dicionário. Se houver uma correspondência, a entidade é reconhecida e classificada. Uma variação desse método usa um PoS (Part of Speech) tagger e um parser superficial para identificar frases nominais, aumentando a precisão do reconhecimento.
- Métodos Supervisionados: Nesta abordagem, um corpus de treinamento anotado com entidades é usado para treinar modelos de aprendizado de máquina ou aprendizado profundo. Modelos supervisionados como CRFs (Conditional Random Fields) e HMMs (Hidden Markov Models) foram utilizados antes do advento das redes neurais. Atualmente, redes neurais como LSTM bidirecionais com uma camada CRF no topo são comuns. Estes modelos utilizam embeddings de métodos como Word2Vec, GloVe e ELMo para representar tokens.
- Abordagem de Rotulagem de Sequências: Esta é a abordagem padrão mais comum para NER, onde o texto é tokenizado e cada token é rotulado como parte de uma entidade (B para início, I para interior) ou fora de qualquer entidade (O). Esta rotulagem, conhecida como BIO tagging, pode ser estendida para suportar múltiplos tipos de entidades. Modelos como RNNs, LSTMs e especialmente BERT são utilizados para esta tarefa.

- Aprendizado Baseado em Prompt e Grandes Modelos de Linguagem (LLMs): Recentemente, o aprendizado baseado em prompt, habilitado por LLMs como GPT-3, tem sido uma abordagem eficiente para NER. Ao invés de treinar modelos com dados anotados, prompts são usados para guiar os LLMs a identificar entidades em texto. Esta abordagem pode ser usada com zero ou poucas amostras de treinamento, tornando-se uma solução prática e poderosa para várias tarefas de NLP, incluindo NER.
- Aprendizado Semi-Supervisionado e Não Supervisionado: Devido aos altos custos e atrasos na anotação de dados no domínio biomédico, métodos semi-supervisionados e não supervisionados são valiosos. Um exemplo é o uso de regras “denoised” por redes neurais para gerar um grande conjunto de dados anotados a partir de um pequeno conjunto de sementes. Outra abordagem é reformular a extração de entidades como uma tarefa de compreensão de máquina, onde entidades são extraídas como respostas a perguntas específicas, abordando naturalmente entidades sobrepostas.

O NER não é um problema solucionado, a pesquisa sobre este assunto ainda é muito ativa, seja na avaliação de novas arquiteturas ou modelos, seja na solução de problemas específicos. Para esta pesquisa e a pesquisa na área biomédica, NER é fundamental, pois facilita a identificação e classificação de entidades como doenças, genes, proteínas e outros termos biomédicos em artigos científicos. Esta capacidade é essencial para a construção de KGs, permitindo o processamento de grandes volumes de dados e a descoberta de relações entre entidades, o que facilita insights e avanços no campo da biomedicina. A aplicação do NER em um extrato de texto pode ser visualizada na Figura 1.

Figura 1 - NER aplicado em texto de Biomedicina



Fonte: Elaborado pelo autor (2024)

Extração de Relações *(do inglês Relation Extraction* - RE) é um processo que envolve a identificação e categorização das relações entre diferentes entidades mencionadas em textos. Essas entidades podem ser genes, proteínas, doenças, drogas, entre outras. A RE é crucial em várias aplicações de NLP, especialmente na área biomédica, onde é fundamental para entender interações biológicas e farmacológicas, entre outros aspectos (DEVARAKONDA; RAJA; XU, 2024b).

Conforme BRUCHES *et al.* (2020) RE é uma tarefa frequentemente resolvida acompanhada de tarefa de RE, que busca encontrar pares de entidades que podem ser ligados por uma relação semântica. Quando um conjunto de relações é predefinido, trata-se da tarefa de *Relation Classification* (RC), que consiste em associar cada par de entidades com uma relação semântica específica. É comum a suposição de que as entidades estejam na mesma sentença para facilitar essa associação.

RE possui importância reconhecida na mineração de literatura, onde textos de fontes como artigos do PubMed, resumos médicos, e redes sociais são analisados para extrair informações estruturadas que podem ser armazenadas em bases de conhecimento. Esses dados estruturados são essenciais para a construção de gráficos de conhecimento que apoiam descobertas científicas e melhoram a precisão de modelos de NLP(DEVARAKONDA; RAJA; XU, 2024b).

Para Devarakonda, Raja e Xu (2024b) a RE pode ser realizada através de diversas abordagens. São estas:

- Abordagens Baseadas em Regras ou Padrões: As abordagens baseadas em regras e padrões foram utilizadas inicialmente no NLP biomédico com bons

resultados. Hearst (1992) identificou padrões típicos na linguagem inglesa que sugerem relações ISA (hiperônimo). Por exemplo, a frase “An NSAID such as Aspirin relieves simple body pains” indica uma relação hiperônimo-hipônimo entre “NSAID” e “Aspirin”. A ferramenta SemRep é um exemplo prático que extrai proposições de três partes (sujeito, objeto e relação nomeada) de textos biomédicos, mapeando e normalizando os argumentos para conceitos UMLS.

- Aprendizado Supervisionado: O aprendizado supervisionado é a abordagem mais comum para a extração de relações. Envolve o uso de um corpus de treinamento anotado com entidades e relações entre elas. As relações podem ser definidas dentro de uma estrutura sintática como uma sentença ou podem cruzar sentenças e até parágrafos. A preparação dos dados é crítica, e o problema de extração de relações é frequentemente mapeado para uma tarefa de classificação de texto.
- Aprendizado Não Supervisionado: Quando não há um conjunto de treinamento disponível, o aprendizado não supervisionado pode ser uma alternativa. Essas abordagens dependem de características dos pares de entidades e/ou das sentenças nas quais aparecem para categorizá-los em classes de relações. Elementos sintáticos e semânticos das sentenças são usados tradicionalmente.
- Supervisão Distante: A supervisão distante é usada quando há uma quantidade limitada de dados anotados. Neste método, utiliza-se um conjunto de sementes de pares de entidades conhecidas com relações desejadas para buscar passagens relevantes no corpus. As passagens encontradas tornam-se amostras positivas para treinamento, e as amostras negativas são criadas a partir de pares que não existem na base de dados utilizada como referência.
- Aprendizado Semi-Supervisionado: No domínio biomédico, onde a anotação manual de dados é cara e demorada, métodos semi-supervisionados são promissores. Uma abordagem é o bootstrapping, onde um pequeno número de pares de entidades de alta precisão é usado para identificar passagens no corpus que contenham padrões de relações, que então são usados para encontrar um grande número de amostras positivas para treinamento.
- Métodos Baseados em Aprendizado Profundo: Com o avanço das técnicas de aprendizado profundo, redes neurais começaram a ser amplamente utilizadas

para a extração de relações. Modelos como CNNs (Redes Neurais Convolucionais) e RNNs (Redes Neurais Recorrentes), incluindo variações como LSTMs (Long Short-Term Memory), são treinados para reconhecer padrões complexos em dados textuais. Mais recentemente, modelos baseados em transformers, como BERT, têm demonstrado excelentes resultados na tarefa de extração de relações, beneficiando-se de sua capacidade de capturar dependências de longo alcance no texto. Para BRUCHES *et al.* (2020) métodos baseados na arquitetura Transformer são considerados os mais promissores. Os Transformers geralmente passam por um aprendizado semi-supervisionado, que envolve um pré-treinamento não supervisionado seguido de um ajuste fino supervisionado para a tarefa em questão.

Modelos neurais grandes como o BERT, quando pré-treinados com textos específicos do domínio biomédico, demonstram ganhos significativos em tarefas de NLP em comparação aos modelos de domínio geral. Estudos mostram que a compilação de conjuntos de dados biomédicos públicos e o pré-treinamento desde o início melhoram o desempenho em tarefas como reconhecimento de entidades nomeadas. Esses avanços aceleram a pesquisa em NLP biomédico, proporcionando uma base sólida para a compreensão e raciocínio em linguagem biomédica, destacada no benchmark BLURB (GUNTURU *et al*., 2024).

2.4.3 Modelos de Linguagem de Grande Escala

Modelos de Linguagem de Grande Escala (do inglês Large Language Models – LLMs) são, assim como BERT e GPT, modelos de linguagem pré-treinados baseados na arquitetura transformers, porém sempre contendo um grande número de parâmetros, e treinados para realizar várias tarefas de processamento de texto, como NER e RE (ZHAO *et al*., 2023). A abordagem de LLM envolve o treinamento de modelos em grandes conjuntos de dados não rotulados para adquirir representações de linguagem altamente generalizáveis, adaptáveis para tarefas específicas de NLP (ZHAO *et al*., 2023).

Cabe salientar que um dos principais atributos de uma LLM é a grande quantidade de parâmetros que são gerados a partir de uma imensa quantidade de textos para o treinamento. Devido a isso, tanto modelos BERT quanto GPT podem

ou não ser considerados LLMs, dependendo do número de parâmetros que possuem. Neste sentido, também existe o termo SLM (*Small Language Models*), que se aplica a modelos com um número relativamente menor de parâmetros. Estes modelos menores podem ser adequados para aplicações específicas onde recursos computacionais e dados são limitados.

LLMs têm a capacidade de aprender e prever a próxima palavra ou caractere em uma sequência de texto com base em probabilidades, além de gerar texto coerente e executar tarefas de linguagem natural, como tradução automática, respostas a perguntas e geração de embeddings. O GPT-3, por exemplo, possui 175 bilhões de parâmetros e pode realizar tarefas de linguagem natural com poucas amostras ou instruções simples (BROWN *et al*., 2020). Recentemente, a OpenAI apresentou o GPT-4, que adicionou melhorias na capacidade e segurança em comparação com versões anteriores, resolvendo problemas complexos de maneira significativamente superior (ZHAO *et al*., 2023).

Entretanto, os LLMs apresentam desafios e limitações significativas, como a necessidade de grandes volumes de dados e recursos computacionais para treinamento, além da complexidade na interpretação dos resultados gerados. Eles podem ter dificuldade em lidar com nuances e ambiguidades na linguagem natural e podem gerar respostas que não fazem sentido ou são inadequadas para determinadas tarefas (ZHAO *et al*., 2023).

Um dos principais desafios no uso de modelos de linguagem de larga escala (LLMs) para a geração de linguagem natural é garantir a qualidade e a confiabilidade dos textos gerados, apreciando ao mesmo tempo as capacidades de raciocínio dos LLMs (KOJIMA *et al*., 2022). Estudos recentes sobre técnicas de prompting mostraram que as habilidades de raciocínio dos LLMs podem ser aprimoradas ao encadear pensamentos em uma estrutura semelhante a um fluxograma (WEI *et al*., 2022). Embora os LLMs possam produzir textos fluentes e coerentes, sem uma verificação adequada, podem faltar consistência factual e introduzir informações irrelevantes ou falsas que não são suportadas por evidências (JI *et al*., 2023).

Uma abordagem comum para lidar com a falta de consistência factual nos textos gerados por LLMs é aumentar os LLMs com fontes de conhecimento externas, como grafos de conhecimento ou bancos de dados relacionais, que podem fornecer fatos relevantes e contexto para a tarefa de geração (LEWIS *et al*., 2020). Essa abordagem é conhecida como Geração Aumentada por Recuperação (RAG) e tem mostrado melhorar o desempenho dos LLMs em várias tarefas intensivas em

conhecimento (WANG *et al*., 2019; BRATE *et al*., 2022). No entanto, o RAG também apresenta limitações, como a dificuldade de selecionar os pontos de conhecimento mais apropriados de uma fonte de conhecimento grande e ruidosa e a falta de transparência e explicabilidade no processo de raciocínio dos LLMs.

Em relação a esta pesquisa, LLMs podem ser extremamente úteis na extração e representação de conhecimento a partir de abstracts de artigos científicos para a criação de um KG. Trabalhar com LLMs pode se tornar mais simples ao não exigir tantas atividades de programação, muitas vezes se limitando à criação de prompts que geram a saída desejada. Para isso, a LLM deve ser treinada com textos da área desejada, podendo ser realizado um fine-tuning de um modelo previamente treinado ao adicionar mais textos específicos da área. Essas técnicas permitirão identificar e relacionar entidades de maneira eficiente, facilitando a construção de uma representação estruturada do conhecimento disponível na literatura biomédica, porém, é importante frizar que LLMs são frequentemente consideradas caixas-pretas devido à sua complexidade e à dificuldade de interpretar os processos internos que levam às suas previsões. Esses modelos têm bilhões de parâmetros, e as interações entre esses parâmetros não são facilmente compreensíveis, o que dificulta a rastreabilidade e a explicação dos resultados gerados.

Na área da Biomedicina, avalar o funcionamento dos LLMs é especialmente difícil porque a precisão e a confiabilidade são cruciais, talvez a Biomedicina seja das áreas onde mais se exige transparência ou explicabilidade dos processos que levam a determinados resultados.

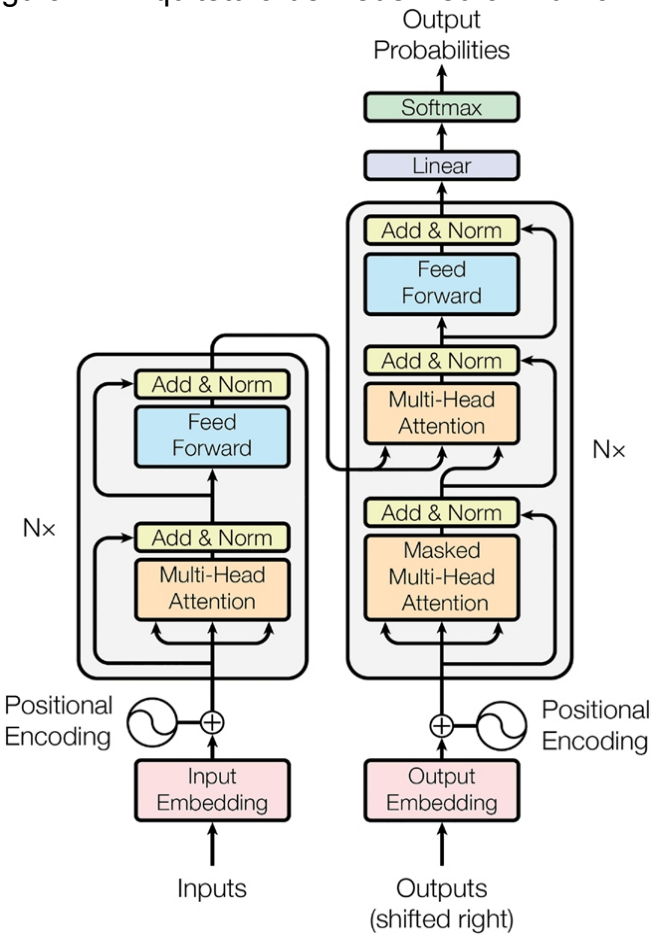
2.4.3.1 Transformers

A arquitetura conhecida como Transformers foi apresentada por Vaswani *et al*. (2017) em seu trabalho intitulado “Attention Is All You Need”, a equipe de pesquisa da Google® desenvolveu uma inovadora arquitetura de rede neural denominada transformer. Este modelo de Deep Learning adota uma estrutura de codificador-decodificador e é fundamentado no mecanismo de autoatenção, destacando-se pela sua capacidade de processar sequências de dados de maneira eficiente. Na arquitetura de transformers, o mecanismo de autoatenção é essencial para entender como as palavras em uma sentença se relacionam entre si. Essa técnica ajuda o modelo a captar a dependência entre as palavras, contribuindo para uma compreensão mais completa do contexto geral do texto. Isso permite que o modelo processe informações complexas e desempenhe tarefas como tradução e

geração de texto de maneira eficaz, ajustando seu foco nas partes relevantes do texto conforme necessário.

Na visão de Vaswani *et al*. (2017), a arquitetura do transformer é descrita através de uma série de camadas de autoatenção e camadas densas, configuradas tanto no codificador quanto no decodificador. A estrutura inicia com o codificador, que consiste em seis camadas idênticas; cada uma dessas camadas possui uma subcamada de autoatenção seguida por uma subcamada de rede feed-forward. O processo começa com o codificador recebendo uma sequência de vetores, organizados em uma matriz, que são primeiramente processados pela subcamada de autoatenção e, posteriormente, encaminhados para a rede feed-forward. Esta, então, passa os vetores processados para a próxima camada dentro do próprio codificador, facilitando uma integração e processamento eficientes da informação. Esta arquitetura é representada pela Figura 2 , apresentada no seu artigo original.

Figura 2 - Arquitetura de Rede Neural *Tranformer*



Fonte: Vaswani *et al*. (2017)

No Quadro 1 são apresentados alguns modelos que fazem uso da arquitetura Transformers, todos eles com aplicação em NLP.

Quadro 1 - Modelos que fazem parte da arquitetura Transformers e ano de lançamento

Mês e Ano	Modelo	Descrição
Junho 2018	GPT	Primeiro modelo Transformer pré-treinado, NLP.
Outubro 2018	BERT	Pré-treinado projetado para resumos de sentenças.
Fevereiro 2019	GPT-2	Versão aprimorada e maior do GPT.
Outubro 2019	DistilBERT	Versão destilada do BERT, é mais rápida e leve.
Outubro 2019	BART, T5	Modelos pré-treinados grandes que usam a mesma arquitetura do Transformer original.
Maio 2020	GPT-3	Versão maior do GPT-2.
Julho 2019	RoBERTa	Otimização do BERT que foi ajustada com hiperparâmetros.
Junho 2019	XLNet	Combina de linguagem autorregressiva e autoatenção.
Abril 2019	ERNIE	Incorpora conhecimento do mundo real.
Setembro 2019	ALBERT	Versão aprimorada do BERT com menos parâmetros.
Março 2020	Megatron	Expande arquiteturas GPT-2, BERT e T5
Fevereiro 2023	LLaMA	Raciocínio de senso comum com zero e poucos shot.

Fonte: Elaborado pelo autor (2024)

Transformers têm sido uma inovação revolucionária no campo do NLP, permitindo a criação de modelos avançados como os já citados, e suas variantes. O mecanismo de autoatenção para processar textos de forma eficiente e paralela, melhorando significativamente a precisão e a capacidade de generalização dos modelos, têm se demonstrado importante para a biomedicina, e outras áreas, já que

esses modelos facilitam tarefas relacionadas ao DM, aprimorando a análise de literatura científica e dados clínicos, exemplificado por modelos como BioBERT.

2.4.3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers), um modelo de aprendizado profundo pré-treinado para NLP, foi criado em 2018 utilizando bases de texto como o BookCorpus (800 milhões de palavras de livros) e a versão em inglês da Wikipedia (2,5 bilhões de palavras). Segundo Devlin *et al*. (2018), o BERT foi criado para pré-treinar representações bidirecionais profundas a partir de um contexto não rotulado, incorporando conjuntamente tanto o contexto esquerdo quanto o direito em todas as camadas. Como resultado, o modelo BERT pré-treinado pode ser ajustado com apenas uma camada de saída adicional para criar modelos de última geração que podem ser utilizados para resolver várias tarefas.

A estrutura do BERT possui duas etapas principais: o pré-treinamento e o ajuste fino. No pré-treinamento, o modelo é treinado com dados não rotulados em diferentes tarefas. No ajuste fino, todos os parâmetros do modelo pré-treinado são ajustados utilizando dados rotulados da tarefa downstream, criando assim modelos específicos para cada tarefa a partir da mesma inicialização pré-treinada (Devlin *et al*., 2018).

A arquitetura transformer do BERT é usada tanto no pré-treinamento quanto no ajuste fino, alterando apenas a camada de saída conforme a tarefa. Tokens especiais como [CLS] e [SEP] são inseridos nas entradas do modelo durante o pré-treinamento para demarcar sentenças e separar pares de sentenças. O BERT primeiro pré-treina uma representação textual genérica antes de ajustá-la para tarefas específicas downstream por meio do ajuste fino (Devlin *et al*., 2018).

O BERT foi inicialmente construído com foco em duas tarefas: Masked Language Modeling (MLM) e Next Sentence Prediction (NSP) (Rothman, 2021). No MLM, o objetivo é prever palavras ausentes em uma frase, utilizando tanto palavras anteriores quanto subsequentes. Já na NSP, o objetivo é prever se duas sentenças estão sequencialmente conectadas no texto original, ajudando o modelo BERT a entender os relacionamentos entre sentenças, crucial para diversas tarefas downstream como perguntas e respostas e inferência de linguagem natural (Devlin *et al*., 2018; Ekman, 2021).

BERT é fundamental para a pesquisa em NLP, especialmente em tarefas de NER, ER e Q&A (Question and Answer), devido à sua capacidade de pré-treinar representações bidirecionais profundas de texto. No contexto da biomedicina, o BioBERT, uma variante do BERT treinada em textos biomédicos, demonstra eficácia superior na identificação de entidades biomédicas e na extração de relações entre elas. Outro exemplo significativo é o ClinicalBERT, que é ajustado especificamente para textos clínicos, melhorando a precisão em tarefas de NLP dentro do domínio médico. Esses modelos permitem uma análise mais precisa e contextualizada de grandes volumes de literatura científica, auxiliando na construção de grafos de conhecimento robustos e na extração de informações relevantes para avanços na pesquisa biomédica.

2.4.3.3 GPT

O GPT (Generative Pre-trained Transformer) é um modelo de DL pré-treinado, desenvolvido pela equipe de inteligência artificial da OpenAI® em 2018, utilizando grandes quantidades de dados textuais, como o BookCorpus (800 milhões de palavras) e a versão em inglês do Wikipedia (2,5 bilhões de palavras) (RADFORD *et al*., 2018). Para Ekman (2021), o GPT é um modelo de linguagem natural treinado para prever a próxima palavra, ou seja, sua tarefa de pré-treinamento é a geração de texto.

O pré-treinamento do GPT utiliza a arquitetura do decodificador do transformer, um modelo de autoatenção que calcula a atenção utilizando apenas as palavras que precedem determinada palavra na sentença, de acordo com a ordem de passagem, seja da esquerda para a direita ou vice-versa. Esse mecanismo de autoatenção permite ao modelo aprender como diferentes tokens se relacionam durante o pré-treinamento (RADFORD *et al*., 2018).

Durante o pré-treinamento, o modelo recebe uma sentença arbitrária como entrada e utiliza funções de ativação softmax em suas camadas para gerar uma distribuição de probabilidade sobre todo o vocabulário. A palavra com maior probabilidade é prevista como a próxima palavra na sentença de entrada. Por exemplo, para a sentença “gpt is pre trained on an lm task”, o modelo tentaria prever a palavra “pre” após observar “gpt is” (EKMAN, 2021).

O GPT se destaca por ser um modelo com bom desempenho para a geração de texto e outras tarefas relacionadas à linguagem, utilizando uma arquitetura de transformer decodificador para aprender representações bidirecionais profundas a partir de contextos não rotulados. Sua importância para a minha pesquisa na biomedicina reside na capacidade de gerar textos coerentes e realizar tarefas complexas de NLP, como a extração e representação de conhecimento de artigos científicos para a criação de KGs.

2.4.4 Retrieval-Augmented Generation

A *Retrieval-Augmented Generation* (RAG) é uma abordagem avançada em NLP que integra Modelos de LLMs com fontes de conhecimento externas para melhorar a relevância e precisão das respostas geradas. Ao combinar módulos não-paramétricos e paramétricos, o RAG recupera documentos ou conhecimentos relevantes de um grande corpus e os utiliza como contexto adicional no processo de geração. Essa técnica aprimora a profundidade e a fidelidade das respostas, superando modelos que dependem apenas de parâmetros internos (OMRANI *et al*., 2024; PAN *et al*., 2024).

Em aplicações práticas, o RAG emprega técnicas como *Maximum Inner Product Search* (MIPS) para buscar conhecimento relevante e utiliza esses documentos como variáveis ocultas dentro de uma estrutura Seq2Seq (sequence-to-sequence) de LLM. O processo iterativo permite o refinamento contínuo das respostas, melhorando tarefas como perguntas e respostas em domínio aberto e geração de histórias. Essa abordagem melhora significativamente a especificidade, diversidade e precisão factual do texto gerado em comparação com métodos tradicionais (PAN *et al.*, 2024).

Além disso, o RAG mostrou benefícios substanciais em *Spoken Language Understanding* (SLU) ao integrar a recuperação de fala com um codificador de reconhecimento automático de fala pré-treinado. Esse processo iterativo de recuperação e geração permite o refinamento progressivo das previsões, melhorando significativamente a previsão de intenção a partir da fala (YANG *et al.*, 2024). A natureza iterativa do RAG não só aumenta o desempenho, mas também permite que o modelo melhore suas respostas de forma adaptativa, tornando-o uma ferramenta poderosa em várias aplicações de NLP.

A utilização de RAG pode trazer várias vantagens para a extração e representação de conhecimento de textos científicos em Biomedicina. Primeiramente, o RAG permite incorporar informações contextuais externas, aumentando a precisão e relevância das entidades e relações extraídas. Além disso, a capacidade de iterar sobre a recuperação e geração de conhecimento permite refinar continuamente as respostas, garantindo uma análise mais detalhada e contextual. No caso de um software final para uso em pesquisa ou comercial, essas melhorias poderiam facilitar a criação de gráficos de conhecimento mais precisos e completos, aprimorar a capacidade de resposta a consultas específicas e suportar descobertas mais rápidas e eficazes em textos científicos.

2.5 GRAFOS DE CONHECIMENTO

Grafos (do inglês Graphs) têm uma longa história na matemática e, ao longo do tempo, evoluíram para desempenhar um papel fundamental em várias áreas da ciência da computação. De acordo com Needham e Hodler (2019), os grafos surgiram no século XVIII, com o matemático Leonhard Euler resolvendo o famoso problema das Pontes de Königsberg. Desde então, os grafos têm sido amplamente estudados e aplicados em diversos campos da ciência e tecnologia.

Com o advento da ciência da computação, os grafos ganharam ainda mais destaque. Eles se tornaram uma estrutura de dados essencial para representar e modelar uma ampla variedade de problemas computacionais. Conforme apontado por Cormen *et al*. (2022), os grafos são usados em algoritmos de busca, sistemas de recomendação, otimização, análise de redes sociais e muito mais.

Uma das aplicações mais notáveis dos grafos na ciência da computação é na representação de redes complexas, como redes de computadores, redes sociais e redes biológicas. De acordo com Albert and Barabási (2002), as redes complexas são caracterizadas por uma estrutura não trivial e propriedades emergentes que não podem ser compreendidas apenas através do estudo de seus componentes individuais. Os grafos oferecem uma maneira poderosa de modelar e analisar essas redes, permitindo a identificação de padrões, comunidades e relações relevantes.

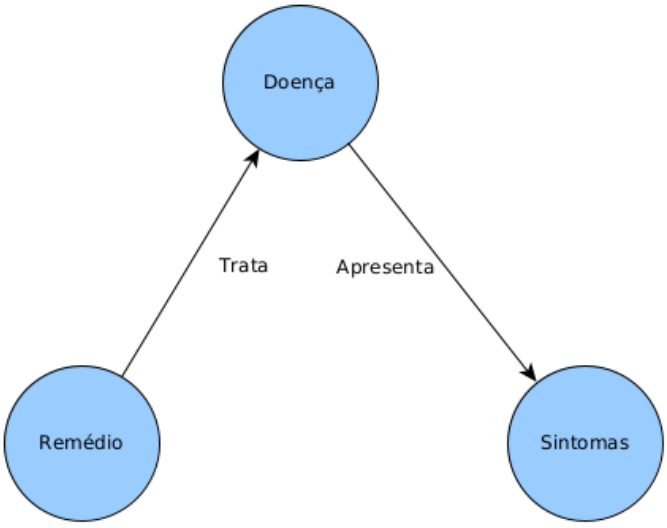
Além disso, os grafos têm sido amplamente utilizados em técnicas de aprendizado de máquina e inteligência artificial. Como mencionado por Goyal and Ferrara (2018), os grafos são essenciais para modelar relações complexas entre

entidades em problemas de aprendizado supervisionado e não supervisionado. Algoritmos de aprendizado de máquina baseados em grafos, como as Redes Neurais Convolucionais em Grafos (GCNs), tornaram-se amplamente usados em tarefas como classificação, detecção de anomalias e recomendação.

Nesta dissertação, será explorada a importância dos grafos na representação e análise de dados não estruturados, como os textos científicos sobre Diabetes Mellitus. Por meio de técnicas avançadas de NLP e algoritmos de mineração de dados, buscamos extrair informações relevantes dos textos e construir um KG que possa contribuir para uma compreensão mais aprofundada da doença e suas implicações na área biomédica.

Conforme Bratanic (2024) um grafo é composto por “vértices”, também conhecidos por “nós”, que representam entidades como pessoas, organizações, locais ou conceitos biomédicos, e por relações, que descrevem como essas entidades estão conectadas, relacionadas, montadas, entre outras interações. Informações apresentadas em formatos de tabelas ou documentos podem ser modeladas e transformadas em um grafo, possibilitando uma visualização mais dinâmica e interativa das conexões e relações entre os dados. Esta transformação permite explorar e analisar complexas redes de informações de maneira mais eficiente e intuitiva. Sobre esse assunto irei me aprofundar um pouco mais na próxima seção. Na Figura 3 pode-se visualizar um grafo simples.

Figura 3 - Exemplo de um Grafo simples direcionado



Fonte: Elaborado pelo autor (2024)

No grafo ilustrado, os vértices são representados por círculos e cada um é rotulado com textos que identificam diferentes entidades. As arestas, por outro lado, simbolizam os relacionamentos entre essas entidades e são igualmente etiquetadas, porém com textos que são verbos ou atributos que descrevem a natureza dessas conexões. Então podemos afirmar que o vértice “Remédio” possui o relacionamento (aresta) “Trata” com o vértice “Doença”, indicando que “os remédios tratam doenças”.

Cabe salientar que a teoria de Maturana e Varela, conhecida como “autopoiese”, apresenta conceitos que podem ser relacionados com a estrutura de grafos. Segundo Maturana (1970), um sistema autônomo é definido pela sua capacidade de auto-organização, ou seja, a habilidade de se autogerir e manter sua identidade. Esse conceito pode ser equiparado ao funcionamento de um grafo, em que os nós representam entidades autônomas e as arestas simbolizam as relações entre elas.

Além disso, a teoria da autopoiese também destaca a importância da circularidade das interações no sistema. Nesse sentido, Maturana e Varela (1980) enfatizam que a auto-organização ocorre por meio de recursões múltiplas e entrelaçadas. Analogamente, em um grafo, a exploração das conexões e iterações entre os nós pode revelar informações importantes sobre o comportamento do sistema representado.

Ao aplicar esses conceitos da teoria de Maturana e Varela ao estudo de grafos, podemos perceber como as estruturas de redes complexas e a interação entre seus elementos são fundamentais para a compreensão de sistemas dinâmicos e auto-organizados. De fato, é possível encontrar trabalhos que utilizam a teoria da autopoiese em conjunto com a análise de grafos para modelar sistemas adaptativos e emergentes Urrestarazu (2011).

Assim, a relação entre grafos e as teorias de Maturana abre caminhos para uma abordagem mais profunda na compreensão de sistemas complexos e sua dinâmica, contribuindo para a construção de modelos e estratégias que possam ser aplicados em diversas áreas do conhecimento.

Conforme Stamile, Marzullo e Deusebio (2021), em matemática e ciência da computação, um grafo é uma estrutura composta por dois elementos fundamentais: os vértices (V) e as arestas (E). Essa representação gráfica é utilizada para descrever as relações entre diferentes entidades, objetos ou elementos de um determinado conjunto.

Os vértices (V) são os pontos ou nós do grafo, e cada um deles representa uma entidade ou objeto específico. Essas entidades podem ser qualquer coisa, desde cidades em um mapa, pessoas em uma rede social, até palavras em um texto.

As arestas (E), por sua vez, são as conexões ou relações que existem entre os vértices. Elas representam os vínculos, interações ou associações entre as entidades representadas pelos vértices. As arestas podem ser direcionadas, indicando uma relação unidirecional entre os vértices, ou não-direcionadas, denotando uma relação bidirecional.

Essas estruturas, os vértices (V) e as arestas (E), são essenciais para a definição dos diversos tipos de grafos, conforme segue:

- Grafo não direcionado: Um grafo não direcionado é aquele em que as arestas não possuem direção, ou seja, a conexão entre dois vértices é bidirecional. Essa é a forma mais simples de grafo, e sua representação gráfica consiste em linhas que conectam os vértices.
- Grafo direcionado: Em contraste com o grafo não direcionado, o grafo direcionado possui arestas com direção, indicando uma relação unidirecional entre os vértices conectados. Nesse tipo de grafo, a aresta é representada por uma seta que aponta do vértice de origem para o vértice de destino.
- Grafo ponderado: Em um grafo ponderado, cada aresta é associada a um valor numérico conhecido como peso. Esse peso pode representar a distância, o custo ou qualquer outra medida relevante da relação entre os vértices conectados.
- Grafo cíclico: Um grafo cíclico é aquele que contém, pelo menos, um ciclo, ou seja, uma sequência de arestas que permitem percorrer o grafo e retornar ao mesmo vértice.
- Grafo acíclico: Em contraste, um grafo acíclico é aquele que não possui ciclos. Isso significa que não é possível percorrer o grafo e voltar ao vértice inicial seguindo as arestas.
- Grafo completo: Um grafo completo é aquele em que todos os vértices estão conectados diretamente entre si por uma aresta. Ou seja, cada par de vértices possui uma aresta que os conecta.

Aplicando essas classificações no grafo apresentado na Figura 3, temos um grafo direcionado simples. Os grafos possuem diversos níveis de complexidade, porque são capazes de representar dados de alta complexidade.

Esses são alguns dos tipos de grafos mais comuns, e cada um deles possui características únicas que os tornam relevantes em diferentes contextos e aplicações. A estrutura dos vértices e arestas é fundamental para a análise e modelagem dessas relações, proporcionando uma visão clara e abrangente dos sistemas e fenômenos representados por essas estruturas complexas. (Stamile, Marzullo & Deusebio, 2021)

Grafos de Conhecimento (do inglês Knowledge Graph - KG) é um tópico de pesquisa cada vez mais popular na área de representação de conhecimento (CHEN *et al*., 2022). Conforme observado por Zhang *et al*. (2020), não há uma definição unificada de grafo de conhecimento na literatura. Devido à sua relevância no cenário atual e crescente popularidade, há uma quantidade significativa de estudos recentes que exploram este tema.

Bratanic (2024) expõe que um KG é uma representação de dados que utiliza a estrutura de grafos para organizar informações e suas interconexões, tornando-o uma poderosa ferramenta para armazenar conhecimento. Por sua vez, para Serles e Fensel (2024), um grafo de conhecimento é uma grande rede semântica que integra fontes de dados heterogêneas em uma estrutura de grafo. Essa definição de mais alto nível diferencia um grafo de conhecimento de uma simples estrutura de dados (Grafo), enfatizando sua capacidade de interconectar diferentes tipos de informações.

Os atributos de um KG, conforme Bratanic (2024), incluem:

- Vértices (Nós): Os vértices do KG representam entidades, conceitos ou objetos do mundo real. Por exemplo, em um KG sobre doenças, os vértices podem ser diferentes tipos de doenças, como “Diabetes Mellitus tipo 1” e “Diabetes Mellitus tipo 2”.
- Arestas (Relações): As arestas do KG representam as relações ou conexões entre os vértices. Em um KG sobre doenças, as arestas podem representar diferentes tipos de relacionamentos entre as doenças, como “é uma subclasse de”, “tem como causa” e “é tratada por”.
- Direcionamento das Arestas: Algumas vezes, as arestas podem ser direcionadas ou não-direcionadas. A direção indica a natureza da relação entre os vértices. Por exemplo, uma aresta direcionada pode representar a

relação “é uma subclasse de”, enquanto uma aresta não-direcionada pode representar a relação “tem uma conexão com”.

- Peso das Arestas: Em um KG ponderado, as arestas podem ter um peso associado, que representa a importância ou a intensidade da relação entre os vértices. Isso é útil para fornecer informações adicionais sobre a natureza da conexão.
- Ciclicidade: Alguns KGs podem ser cíclicos, o que significa que possuem ciclos ou loops, permitindo que um vértice seja alcançado seguindo um caminho de arestas e voltando ao mesmo vértice. Outros KGs são acíclicos, o que significa que não possuem ciclos.
- Completude: Em um KG completo, cada par de vértices está conectado por uma aresta, formando um grafo completamente conectado. No entanto, nem todos os KGs precisam ser completos; eles podem ser esparsos, com conexões seletivas entre os vértices.
- Atributos dos Vértices e Arestas: Além das relações entre os vértices, os KGs também podem armazenar informações adicionais nos próprios vértices e arestas. Por exemplo, um vértice que representa uma doença pode conter atributos como “sintomas”, “prevalência” e “tratamento”.

Esses são alguns dos atributos importantes que podem ser encontrados em um Grafo de Conhecimento. A estrutura flexível e rica do KG permite representar e explorar informações complexas de diversas áreas, como a medicina, a ciência da computação, engenharia do conhecimento, redes sociais, biologia, dentre outras.

Para Serles e Fensel (2024), os grafos de conhecimento diferenciam-se dos sistemas tradicionais baseados em conhecimento por conterem muitos fatos explícitos em vez de um grande conjunto de regras para deduzir respostas a perguntas. Um grafo é uma estrutura matemática na qual alguns pares em um conjunto de objetos estão de alguma forma relacionados. No exemplo de um grafo de conhecimento sobre eventos, representado como um grafo rotulado com arestas direcionadas, os nós representam entidades e as arestas representam os relacionamentos entre elas. Uma característica central dos grafos de conhecimento é que não há uma distinção rígida entre esquema e dados: tudo é apenas nós e arestas. Isso torna os grafos de conhecimento muito flexíveis na integração de fontes heterogêneas, sem a necessidade de seguir ou atualizar um esquema rígido como em bancos de dados relacionais, facilitando seu crescimento em tamanho. No

entanto, algumas arestas e nós podem vir de um vocabulário controlado e ter um significado especial, utilizando uma ontologia.

O conhecimento dentro de um KG é representado por meio de triplas (sujeito-predicado-objeto), em que este conjunto representa uma unidade de informação ou uma declaração verdadeira dentro do grafo de conhecimento. Em outras palavras, uma tripla captura um relacionamento específico entre duas entidades, e essa declaração é tratada como um “fato” no contexto do grafo de conhecimento (CHEN *et al*., 2022; CHEN; JIA; XIANG, 2020; ZHANG *et al*., 2020).

Infelizmente, os KGs são frequentemente desenvolvidos por diversas instituições de forma independente, cada uma utilizando estruturas, linguagens e métodos próprios para representar as mesmas entidades e relações. Essa falta de padronização resulta em incompatibilidades e dificuldades na integração e interoperabilidade dos grafos de conhecimento. Essas disparidades podem complicar a combinação de dados e a utilização conjunta de informações provenientes de diferentes fontes, dificultando a criação de um sistema unificado e abrangente de conhecimento (YU *et al*., 2022).

Os grafos de conhecimento são projetados para ser em grande escala, capazes de lidar com bilhões a trilhões de dados. Eles podem ser descritos como uma enorme coleção de fatos que descrevem o mundo a partir de múltiplos pontos de vista.

KGs são uma ferramenta poderosa não apenas para representar conhecimento, mas, quando combinados com técnicas como GNNs, RAG e LLMs, podem tornar as consultas mais assertivas. Estas combinações têm se tornado cada vez mais comuns em pesquisas. Um exemplo é o trabalho de Matsumoto *et al*. (2024), que propõe uma ferramenta chamada KRAGEN.

Matsumoto *et al*. (2024) apresentam a aplicação de código aberto *Knowledge Retrieval Augmented Generation Engine* (KRAGEN), que visa superar as limitações do RAG usando a técnica de *graph-of-thoughts* (GoT) para modelar e executar problemas complexos com LLMs (Besta *et al*., 2023). O GoT permite decompor um problema em subproblemas e representar a informação gerada pelo LLM como um grafo, onde os vértices são unidades de informação e as arestas são dependências. Essa técnica possibilita realizar várias operações semelhantes a grafos, como criar arestas e unir pensamentos para melhorar a qualidade dos vértices.

Como um GoT é necessariamente um grafo, o KRAGEN pode fornecer uma interface visual que mostra o processo de pensamento do LLM, revelando como ele constrói e valida suas respostas usando lógica transitiva e evidências factuais. A novidade em combinar a estrutura RAG com GoT, utilizando uma estrutura de grafo para navegar pelos pensamentos de qualidade de conhecimento, permite alcançar uma geração de linguagem natural mais explicável e confiável (Lv *et al*., 2019).

A construção de um KG a partir de artigos científicos na área de Biomedicina, especificamente sobre diabetes mellitus, utilizando técnicas de NER e ER, possui uma importância significativa para o mapeamento e organização do conhecimento. Um KG bem estruturado permite a integração de dados heterogêneos em uma estrutura coerente e inter-relacionada, facilitando a descoberta de novas conexões e insights que poderiam passar despercebidos em uma análise convencional de texto. Este processo de transformação de informações não estruturadas em um formato estruturado e interligado é crucial para avançar na compreensão de doenças complexas como a diabetes mellitus, promovendo uma visão mais ampla e detalhada das relações entre genes, proteínas, tratamentos e resultados clínicos.

2.6 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os conceitos que fundamentam a proposição e o desenvolvimento do modelo desta pesquisa. Os conceitos de Knowledge Graphs (KGs), Named Entity Recognition (NER) e Relation Extraction (ER) foram abordados em relação aos objetivos geral e específicos do estudo, compondo o núcleo principal da investigação. A Inteligência Artificial, especialmente através das técnicas de NLP, NER, ER, e ML/DL representada por modelos pré-treinados como BERT e GPT, forneceu a base tecnológica essencial para a construção de representações vetoriais semânticas e a criação do KG.

Além disso, apresentou-se a importância das técnicas de embeddings, tanto estáticos quanto contextualizados, na representação do conhecimento extraído de textos científicos. Essas técnicas são cruciais para a realização de tarefas de NER e ER, permitindo a identificação precisa de entidades e suas relações, fundamentais para a construção de um KG robusto e informativo.

Por fim, exploramos as vantagens do uso de tecnologias como Retrieval-Augmented Generation (RAG) em conjunto com KGs e LLMs, destacando seu

potencial para tornar as consultas mais assertivas e promover a descoberta de novos insights na área de biomedicina, especificamente sobre diabetes mellitus. A síntese dos trabalhos correlatos foi apresentada, destacando as contribuições deste estudo em diferentes dimensões, e reforçando a relevância da criação de um KG validado e integrado para a pesquisa científica e aplicações práticas na biomedicina.

3 METODOLOGIA DE PESQUISA

Para assegurar o entendimento e a correta interpretação dos resultados, é fundamental compreender os procedimentos metodológicos adotados (GRAY, 2012). Este capítulo detalha a metodologia de pesquisa utilizada nesta dissertação, abordando os procedimentos e etapas realizadas. Serão apresentadas a classificação da pesquisa quanto à natureza, objetivos e procedimentos, assim como a *Design Science Research Methodology* (DSRM) aplicada. Por fim, será exposto o desenvolvimento da pesquisa e uma síntese da metodologia aplicada.

3.1 ENQUADRAMENTO METODOLÓGICO

Desde Aristóteles, que no século IV a.C. estabeleceu os alicerces do pensamento científico ocidental com o uso da lógica e da observação empírica, o método científico tem evoluído consideravelmente. De acordo com Barnes (1995), Aristóteles via a ciência como um processo de dedução a partir de princípios claros e incontestáveis. Esta abordagem foi expandida por Francis Bacon no século XVII, que introduziu um método indutivo enfatizando a observação e a experimentação sistemática. A evolução continuou com René Descartes e Isaac Newton, que integraram o empirismo de Bacon com dedução lógica e matemática, solidificando as bases da ciência moderna. No século XX, Karl Popper adicionou a falsificação como critério de validade científica, sublinhando a natureza evolutiva e provisória do conhecimento científico, que se desenvolve através de um processo contínuo de conjecturas, refutações e aprimoramentos à luz de novas evidências.

Para Novikov e Novikov (2013) “Ciência é definida como um campo da atividade humana, cuja função consiste na geração e sistematização teórica do conhecimento objetivo a respeito da realidade.”

Para Bunge (2004) “O conhecimento científico é, por definição, o resultado da pesquisa científica, ou seja, da pesquisa realizada com o método e o objetivo da ciência.”. Já para Marconi e Lakatos (2003) o conhecimento científico “É sistemático, já que se trata de um saber ordenado logicamente, formando um sistema de ideias (teoria) e não conhecimentos dispersos e desconexos.”.

Marconi e Lakatos (2003) definem método com o “conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o

objetivo - conhecimentos válidos e verdadeiros –, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista".

Por sua parte a metodologia científica pode ser descrita como:

> O estudo sistemático e lógico dos métodos empregados nas ciências, seus fundamentos, sua validade e sua relação com as teorias científicas. Em geral, o método científico compreende basicamente um conjunto de dados iniciais e um sistema de operações ordenadas adequado para a formulação de conclusões, de acordo com certos objetivos predeterminados (GERHARDT; SILVEIRA, 2009).

Conforme Cupani (2006), a pesquisa tecnológica se dedica ao projeto e desenvolvimento de artefatos tecnológicos utilizando o conhecimento científico como base, sendo este o enfoque desta pesquisa. Complementando essa ideia, Freitas Júnior *et al*. (2014) afirmam que a pesquisa tecnológica vem "ganhando cada vez mais espaço na academia, especialmente em áreas como engenharia e computação, campos do saber humano que se ocupam principalmente do desenvolvimento de novos artefatos, nem sempre baseados no conhecimento científico clássico". De acordo com Bunge (1985), o conhecimento científico relativo ao projeto de artefatos envolve o planejamento de sua realização, operação, ajuste, manutenção e monitoramento.

Freitas Júnior *et al*. (2017) afirmam que a pesquisa tecnológica visa à solução de problemas específicos e pontuais, com foco no artefato a ser desenvolvido, que pode não ser necessariamente material, mas sim um projeto ou uma intervenção artificial sobre um sistema. Essas características possibilitam que o conhecimento tecnológico seja aplicado em problemas específicos, com o objetivo de construir novos artefatos (CUPANI, 2016).

Diante do exposto, ao analisar as características desta pesquisa, torna-se evidente que ela se enquadra como uma pesquisa de natureza Tecnológica e Aplicada. Nesse contexto, considerando as características da pesquisa tecnológica, optou-se pela aplicação dos princípios e fundamentos da *Design Science Research Methodology* (DSRM) (PEFFERS *et al*., 2007).

3.2 DESIGN SCIENCE RESEARCH METHODOLOGY

A *Design Science Research* (DSR) é um método empírico que foca na criação sistemática de soluções inovadoras, com ênfase na geração de artefatos (HEVNER *et al*., 2004, p. 4). Horita, Neto e Santos (2018) descrevem o DSR como um conjunto rigoroso de etapas destinadas a avanços no estado da arte, sendo o artefato o conceito central, materializado em modelos, métodos, processos e ferramentas. O DSR se fundamenta na premissa de que criar um artefato inovador para resolver um problema requer avanços no conhecimento do domínio. Os artefatos gerados podem ser categorizados em construtos, modelos, métodos e instâncias (HEVNER *et al*., 2004).

Nos últimos anos, o paradigma do DSR ganhou destaque especialmente nas áreas de sistemas de informação e engenharia. Hevner (2020) define o DSR como um método voltado para solucionar problemas, aprimorando o conhecimento humano por meio de artefatos inovadores. A *Design Science Research Methodology* (DSRM), derivada do DSR, incorpora princípios e práticas para conduzir pesquisas, objetivando consistência com a literatura, orientação para pesquisadores e um modelo para apresentação de resultados (PEFFERS *et al*., 2007). O DSRM cria e avalia artefatos para resolver problemas organizacionais, baseando-se em teorias existentes e produzindo soluções relevantes que são rigorosamente avaliadas quanto à utilidade, qualidade e eficácia (GREGÓRIO *et al*., 2021; HEVNER *et al*., 2004).

Horita, Neto e Santos (2018) descrevem os tipos de artefatos gerados pelo DSR da seguinte forma:

- Construtos: Vocabulários e símbolos essenciais para a descrição e análise de fenômenos de interesse. Por exemplo, taxonomias.
- Modelos: Representações de um problema e algumas das possíveis soluções. Por exemplo, arquiteturas conceituais de sistemas.
- Métodos: Conjunto de atividades para possibilitar a execução de uma tarefa.
- Instâncias: Realizações dos elementos (por exemplo, modelos, métodos, construtos) em um ambiente natural. Essas instâncias também são especializações de conhecimento adquirido para domínios com características similares ou diferentes, como sistemas para o gerenciamento de vendas.

Segundo Peffers *et al*. (2007), a DSRM é composta por seis etapas procedurais, que podem ser executadas de acordo com a necessidade do projeto. Essas etapas são:

1) Identificação do problema e motivação: Esta etapa é dedicada à definição do problema de pesquisa específico, apresentando-se uma justificativa para a sua investigação. É essencial que a definição do problema seja clara para que o artefato a ser desenvolvido possa efetivamente oferecer uma solução. Recursos necessários incluem o estado da arte do problema e a relevância da solução proposta.
2) Definição dos objetivos da solução: Partindo do conhecimento do problema e da viabilidade das soluções possíveis, delineiam-se os objetivos da solução a ser desenvolvida. É importante ter um entendimento claro das possíveis soluções já apresentadas anteriormente.
3) Projeto e desenvolvimento: Nesta etapa, cria-se o artefato, determinando-se suas funcionalidades desejadas e sua arquitetura. O desenvolvimento do artefato real é realizado com base no conhecimento teórico disponível.
4) Demonstração: Esta etapa envolve a demonstração do uso do artefato para resolver uma ou mais instâncias do problema por meio de experimentos, simulações, estudos de caso, provas formais ou outras atividades apropriadas. É crucial entender como usar o artefato para resolver o problema.
5) Avaliação: Observa-se e mensura-se como o artefato atende à solução do problema, comparando os objetivos propostos com os resultados obtidos. Caso necessário, pode-se retornar às etapas de projeto e desenvolvimento para aprimorar o artefato.
6) Comunicação: A etapa final é a divulgação do problema, a relevância da solução proposta e a apresentação do artefato desenvolvido. Esta comunicação pode ocorrer por meio de publicações acadêmicas, apresentações em conferências, entre outros meios.

Na pesquisa em questão, a DSRM será aplicada para propor um método voltado à tarefa de mapeamento de conhecimento a partir de fontes de dados não estruturados em forma de texto, com o objetivo de construir um KG, isto seguindo o método proposto por Peffers *et al*. (2007).

3.3 REVISÃO INTEGRATIVA DA LITERATURA

Nesta dissertação, propõe-se um modelo que integra técnicas avançadas de Processamento de Linguagem Natural (NLP), mineração de dados, representação do conhecimento e aprendizado profundo. O foco é apoiar a análise e extração de conhecimento em documentos científicos dentro do campo da biomedicina, com especial atenção ao diabetes mellitus. O principal objetivo é aprimorar o processo de descoberta de conhecimento, com base nas realizações de pesquisadores anteriores, e facilitar a tomada de decisões. Isso é feito promovendo a extração de informações valiosas de textos científicos e possibilitando inferências que podem revelar novos insights sobre a prevenção, diagnóstico e tratamento do diabetes.

O desafio de identificar problemas e propor soluções inovadoras demanda um entendimento aprofundado do conhecimento atual, obtido por meio de uma revisão integrativa da literatura. Esse método de pesquisa permitiu uma análise crítica dos estudos precedentes, ressaltando tanto as contribuições relevantes quanto as lacunas no domínio específico da biomedicina, o que reforça a importância e a inovação da abordagem proposta.

Diante do aumento constante e da complexidade das informações no setor de saúde, é fundamental o desenvolvimento de ferramentas no âmbito da pesquisa científica. Essas ferramentas devem ser capazes de estabelecer etapas metodológicas claras e permitir que os profissionais aproveitem melhor as evidências reveladas em diversos estudos (TAVARES DE SOUZA; DIAS DA SILVA; DE CARVALHO, 2010). Referente aos exposto, Tavares de Souza e Dias da Silva (2010) deixam claro que “Nesse cenário, a revisão integrativa emerge como uma metodologia que proporciona a síntese do conhecimento e a incorporação da aplicabilidade de resultados de estudos significativos na prática”.

Para a execução de uma pesquisa bibliográfica segundo Tavares de Souza, Dias da Silva e de Carvalho (2010), envolve etapas meticulosas para garantir a qualidade e relevância dos dados coletados. Inicia-se com a seleção cuidadosa de bases de dados relevantes. Segue-se a definição de descritores e suas combinações, e ao mesmo definindo o idioma desses descritores. Os critérios de inclusão são então estabelecidos, novamente podendo considerar os idiomas desejados, o alcance temático dos artigos e a temporalidade da publicação. A análise dos artigos selecionados é orientada por uma abordagem descritiva, baseada nos delineamentos de pesquisa recomendados por autores reconhecidos

na área, permitindo uma síntese eficaz das informações extraídas. Essa metodologia não apenas cria um direcionamento, mas também facilita a obtenção de bons resultados. Na visão deste autor, existe ainda a etapa de compilação do conhecimento extraído, uma organização de simples acesso às informações coletadas.

Assim sendo, procedeu-se à execução de cada uma das etapas como Segue:

- Escolha das bases de artigos: Para esta pesquisa foram escolhidas as seguintes bases de artigos científicos: ResearchGate; ScienceDirect; arXiv; PubMed. Adicionalmente a essas bases foi realizada pesquisa fazendo uso da ferramenta ConnectedPapers, a qual organiza os artigos científicos em um gráfico, baseando-se em sua semelhança. Essa semelhança é determinada por meio de métricas específicas, incluindo a co-citação e o acoplamento bibliográfico, conceitos que refletem como os documentos são citados conjuntamente ou compartilham referências bibliográficas em comum, facilitando assim a identificação de trabalhos relacionados dentro de um campo de estudo específico. Para a utilização do ConnectedPapers foram utilizadas como sementes os artigos com linhas de pesquisa mais próximos a esta dissertação.
- Definição dos descritores: Para esta pesquisa em questão foram definidos os seguintes descritores e álgebra booliana aplicada:
  - ResearchGate: (“data mining” or “knowledge graphs”) and (“Biomedical Scientific Literature” or “Literature” or “Papers”);
  - PubMed: (“data mining”[All Fields] AND “knowledge graphs”[All Fields]) AND ((y_10[Filter]) AND (ffrft[Filter]));
  - arXiv: “data mining” “knowledge graphs”;
  - ScienceDirect: “text mining” “data mining” “Biomedical Scientific Literature”
- Critérios de inclusão: A fim de concentrar os esforços na literatura mais pertinente foram aplicados os seguintes filtros: Anos entre 2010 e 2024; idioma inglês, e título próximo ao objetivo desejado. A execução resultou em conjuntos de 24 registros para PubMed; 100 para ResearchGate; 6 para ScienceDirect; 28 para arXiv.
- Análise dos artigos: A análise iniciou com a verificação do acesso ao artigo (download), passando a analisar o abstract, corpo do texto, o objetivo, a clareza do método, experimento prático, e a explicitação das ferramentas

utilizadas, aplicando assim uma última seleção que resultou em 7 artigos. Os artigos obtidos estão listados no Quadro 2.

Quadro 2 - Artigos selecionados com suas principais informações. (continua)

Título	**Autor**	**Área**	**Método, Técnica, Algoritmo**
KGen: a knowledge graph generator from biomedical scientific literature	(ROSSANEZ *et al*., 2020)	Bioinformatics, Data Mining, NLP	Semantic Role Labeling (SRL), Ontology, NER, ER
Mining a stroke knowledge graph from literature	(YANG *et al*., 2021)	Bioinformatics, Data Mining, Deep Learning, NLP	BioBERT, Co-occurrence, Rule-based, BiLSTM-CRF, NER, ER
Mining On Alzheimer's Diseases Related Knowledge Graph to Identity Potential AD-related Semantic Triples for Drug Repurposing	(NIAN *et al*., 2022)	Bioinformatics, Data Mining, Deep Learning, NLP	BERT-based, PubMedBERT, Rule-based, NER, ER
Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases	(ROSSANEZ; CESAR, 2019)	Bioinformatics, Data Mining, Deep Learning, NLP	Semantic Role Labeling (SRL), Ontology, NER, ER
Terminological resources for text mining over biomedical scientific literature	(RINALDI; KALJURAND; SÆTRE, 2011)	Bioinformatics, Data Mining, NLP	LingPipe, part-of-speech tagger, computational linguistics, Knowledgebase, Ontology

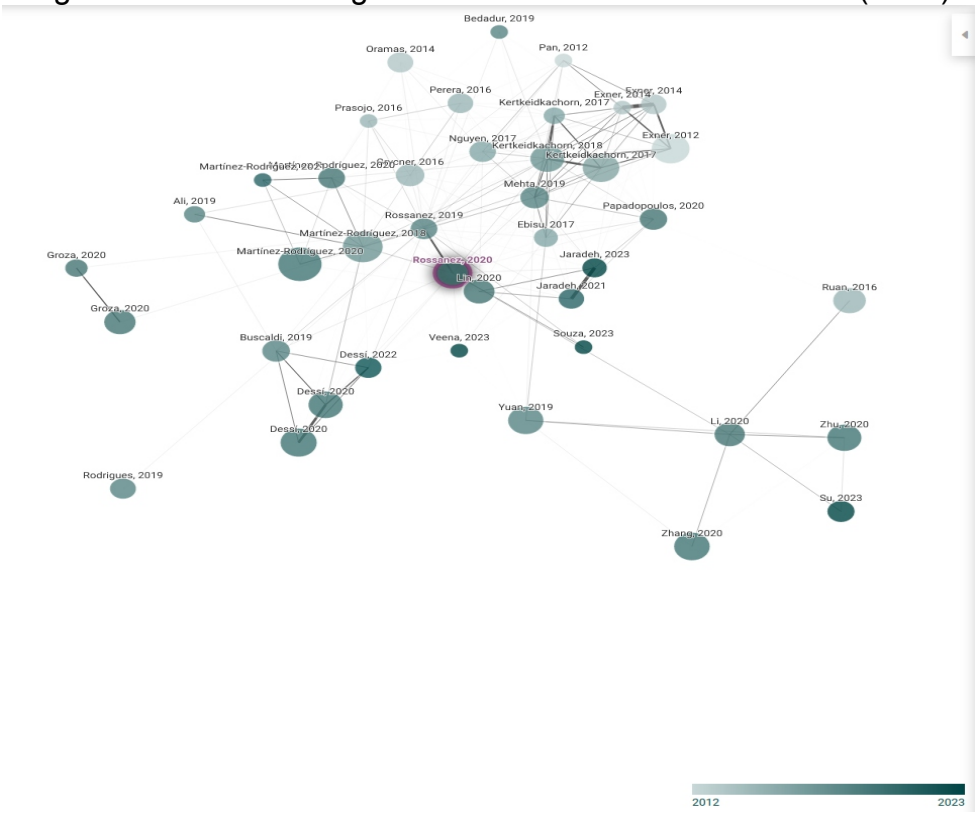
(Conclusão)

Título	**Autor**	**Área**	**Método, Técnica, Algoritmo**
Biomedical named entity recognition based on fusion multi-features embedding	(LI; YANG; LIU, 2023)	Bioinformatics, Data Mining, Deep Learning, NLP	Part-Of-Speech, BioBERT, BiLSTM-CRF, Multi-Feature Embedding, Biomedical Named Entity Recognition (BNER), Knowledgebase
Applying BioBERT to Extract Germline Gene-Disease Associations for Building a Knowledge Graph from the Biomedical Literature	(DIAZ GONZALEZ *et al*., 2023)	Bioinformatics, Data Mining, Deep Learning, NLP	BioBERT, NER, ER, NEN, Knowledgebase, Ontology, Semantic Relation

Fonte: Elaborado pelo autor (2024)

Adicionalmente se fez uso da ferramenta ConnectedPapers a fim de ratificar os artigos escolhidos ou realizar nova adição, para a pesquisa foram utilizados como “sementes” os artigos: “KGen: a knowledge graph generator from biomedical scientific literature” (ROSSANEZ *et al*., 2020) e “Mining a stroke knowledge graph from literature” (YANG *et al*., 2021). Apesar de ter obtido estudos interessantes, estes, ou já estavam na seleção prévia, ou não estavam de acordo com os requisitos, não aletarndo assim a seleção prévia a esta atividade. Na Figura 4 é apresentado o grafo gerado pela ferramenta para o artigo de Rossanez *et al*. (2020).

Figura 4 - Grafo de artigos relacionados a Rossanez *et al*. (2020)



Fonte: Elaborado pelo autor (2024)

A seleção dos artigos apresentados no Quadro 2 reflete a aplicação dos parâmetros desejados na elaboração da base teórica e metodológica desta dissertação. Os primeiros quatro títulos, focados na geração de KGs a partir da literatura científica da área biomédica, estão diretamente relacionados ao objetivo principal deste estudo: explorar técnicas de data mining e NLP para extrair conhecimento estruturado de textos científicos, com a finalidade da representação do conhecimento. Esta escolha reforça a relevância do KG como ferramenta de representação e análise de dados complexos na biomedicina, permitindo uma visão

3.4 DESENVOLVIMENTO DA PESQUISA

Esta pesquisa é de natureza tecnológica e aplicada, com objetivos exploratórios, focando principalmente na proposição e desenvolvimento de um método voltado à representação do conhecimento a partir de documentos sobre diabetes mellitus com a finalidade de construir um grafo de conhecimento. Para isso,

será utilizada a Design Science Research Methodology (DSRM) proposta por Peffers *et al*. (2007), evidenciando os passos utilizados para a solução do problema conforme Quadro 3.

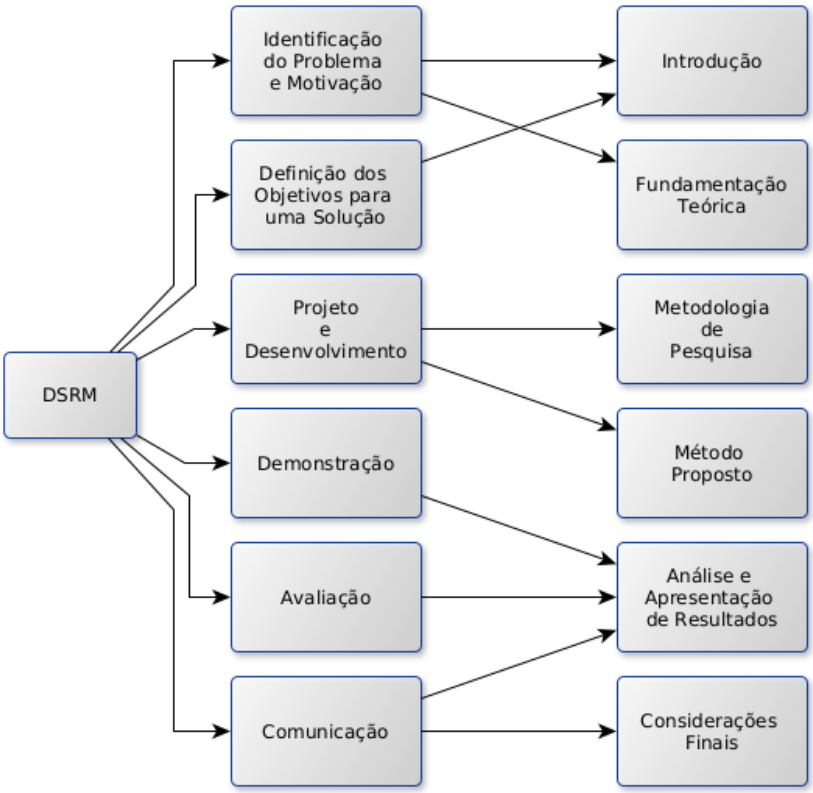
Quadro 3 - Mapeamento atividades de pesquisa e etapas DSRM

Atividade	Descrição
Identificação do problema e motivação	– Como extrair e representar conhecimento contido em grandes quantidades de textos, mais especificamente textos científicos não extruturados sobre diabetes mellitus?
Definição dos objetivos para uma solução	– Propor e desenvolver um método voltado à representação do conhecimento a partir de documentos sobre diabetes mellitus com a finalidade de construir um grafo de conhecimento. – Identificar e catalogar fontes relevantes de literatura científica em biomedicina, com foco específico em estudos e publicações sobre diabetes mellitus; – Identificar e selecionar técnicas para a aplicação de EDA, visando compreender a natureza dos documentos obtidos; – Identificar e listar técnicas e modelos para a extração de informações de textos não estruturados, com ênfase em NER e ER; – Identificar e listar técnicas para a avaliação automatizada das informações extraídas por meio de algoritmos.
Projetar e desenvolver	– Método voltado à representação do conhecimento a partir de documentos sobre diabetes mellitus com a finalidade de construir um KG.
Demonstração	– Demonstração do modelo utilizando o cenário da diabetes mellitus.
Avaliação	– Avaliar primeiramente por observação as triplas resultantes da execução do método; – Avaliar as métricas estruturais e de centralidade do KG (Grafo); – Testar e avaliar a possibilidade de utilizar, de forma automática, ontologia da área para a avaliação das triplas.
Comunicação	– Comunicação à comunidade acadêmica dos resultados obtidos.

Fonte: Elaborado pelo autor (2024)

Na Figura 5 é exposto o mapeamento entre os capítulos desta dissertação e as atividades definidas no DSRM.

Figura 5 - Capítulos referentes ao desenvolvimento desta dissertação



Fonte: Adaptado de Peffers *et al.*(2007)

3.4.1 Definição do Problema e Motivação

As primeiras etapas desta dissertação concentram-se na identificação clara do problema e na fundamentação da motivação que norteia esta pesquisa. Dessa forma, realizou-se uma revisão integrativa da literatura, descrita na Seção 3.3 e com protocolo completo disponível no Apêndice A. O objetivo desta revisão foi identificar uma lacuna significativa no campo da Engenharia do Conhecimento, particularmente em relação à mineração de dados e à extração de informações de artigos científicos sobre Diabetes Mellitus. O objetivo desta pesquisa é a criação de KGs que permitam representar e explicitar o conhecimento adquirido, fornecendo uma base sólida para futuras investigações focadas na prevenção e tratamento da doença.

Durante a revisão, foram analisados 158 documentos oriundos de bases de dados reconhecidas, tais como ResearchGate®, Science Direct®, PubMed® e arXiv®. Dentre os documentos selecionados, sete foram cuidadosamente

selecionados para integrar a matriz de síntese, devido à sua relevância direta ao tema e potencial contribuição para a resolução da lacuna identificada. Esta seleção criteriosa garante que a base de conhecimento desenvolvida está de acordo com as necessidades atuais da pesquisa em Diabetes Mellitus e abre espaço para inovações significativas no tratamento e manejo da condição.

3.4.2 Definição dos Objetivos

Com base na revisão integrativa da literatura para a definição do problema e da motivação, foram delineados os objetivos desta pesquisa, tanto gerais quanto específicos, que sustentam o modelo proposto na tese e sua avaliação através de um cenário de estudo, assim como a formulação da pergunta de pesquisa. A partir da revisão integrativa da literatura e dos estudos iniciais, foram definidas diversas etapas fundamentais para a elaboração do método proposto, que tem como objetivo a representação do conhecimento extraído de textos científicos não estruturados para a construção de um KG.

3.4.3 Projeto e Desenvolvimento

O principal produto desta atividade é o artefato, representado pelo método proposto nesta dissertação, que permite a representação de conhecimento a partir de fontes de dados não estruturados na forma de textos provenientes de artigos científicos. Este método visa a representação e a explicitação do conhecimento, bem como a possibilidade da inferência para a descoberta de novos conhecimentos. Para alcançar esses objetivos, esta dissertação utiliza um conjunto de métodos e técnicas de Engenharia do Conhecimento e Computacionais, buscando atender necessidades da comunidade científica e da sociedade, especialmente no cenário da Biomedicina. As atividades para o desenvolvimento do artefato (método) são apresentadas nas seções a seguir.

3.4.3.1 Coleta dos Dados

Na análise de bases de dados para a pesquisa em biomedicina, com foco em artigos sobre diabetes mellitus, o PubMed apresenta-se como a melhor opção por várias razões estratégicas:

- Amplitude de Cobertura: A cobertura do PubMed é vasta, com mais de 36 milhões de citações e resumos, superando em muito outras bases, como o BioRxiv.
- Acesso gratuito: oferece uma API que facilita a obtenção automática de metadados, o que é indispensável para o processo de ETL (Extração, Transformação e Carga)
- Consistência dos Metadados: A utilização de uma única base de dados torna o processo de ETL mais uniforme, simplificando as etapas de transformação e integração de dados.
- Prestígio Acadêmico: O NCBI, que faz parte da U.S. National Library of Medicine (NLM), assegura que os artigos sejam de elevado nível e reconhecidos internacionalmente.
- Integração de Recursos Diversos: A NLM integra diversos recursos da NLM, como MEDLINE—principal componente com artigos indexados com Medical Subject Headings (MeSH) e dados adicionais—, PubMed Central (PMC) e Bookshelf, que, juntos, oferecem uma grande variedade de textos completos e recursos relacionados às ciências biomédicas e da vida.
- Atualização e Histórico Completos: Apresenta um histórico abrangente de artigos, atualizados constantemente, permitindo análises longitudinais dos estudos em diabetes mellitus.

A escolha do PubMed como fonte primária para esta pesquisa é justificada pelo grande número de dados disponíveis, pela eficiência operacional e pela confiabilidade acadêmica. Esta opção assegura um processo metodológico sólido e confiável para a realização de uma pesquisa científica relevante na área da saúde.

A última consulta dos metadados ocorreu em 15 de abril de 2023, utilizando o pacote BioPython, especificamente através do módulo Entrez. Para operar corretamente o pacote, é necessário fornecer um endereço de e-mail, havendo a opção de realizar consultas anônimas ou identificadas por uma chave de acesso. A principal diferença entre essas modalidades reside no limite de consultas por minuto: três para consultas anônimas e dez para consultas identificadas.

O processo de consulta foi dividido em duas etapas principais. A primeira etapa envolveu a obtenção dos IDs dos artigos a partir de um termo de pesquisa,

neste caso, “diabetes mellitus”, que resultou em 518.432 IDs. Na segunda etapa, procedeu-se à obtenção dos metadados correspondentes a cada ID. Devido ao volume substancial de dados e ao tempo de resposta, optou-se por realizar essas consultas em lotes de 4.000 artigos cada.

A API permite a definição do formato da resposta através de parâmetros, tendo-se escolhido o formato JSON para facilitar a seleção de características. O pseudocódigo da função principal de obtenção de dados é ilustrado na Figura 6.

Figura 6 - Pseudocódigo do algoritmo para a consulta ao PubMed

```
SET step TO 4000
SET query TO 'diabetes mellitus'
SET DEST TO local_path
SET count TO pubmed_count(query)
SET results TO search(query, count)
save_pmids(DEST, results)
SET id_list TO results['IdList']

FOR bottom_bound IN range(0, count, step):
    SET top_bound TO (bottom_bound + step) - 1
    SET papers TO fetch_details(id_list, bottom_bound, top_bound)
    save_papers(DEST, papers, bottom_bound, top_bound)
    sleep(40)
```

Fonte: Elaborado pelo autor (2024)

A Figura 7 apresenta um exemplo de registro no formato JSON. É importante salientar que os atributos vazios são ignorados nos registros resultantes. A omissão é crucial para determinar as tarefas que serão executadas na etapa seguinte de Seleção de Features.

Figura 7 - Extrato de metadados de um artigo do PubMed em formato JSon


```
{"DB": "PubMed", "CLASS": "PubmedArticle", "ID": "33403891", "DateCompleted": null,
 "DateRevised": {"Year": "2021", "Month": "03", "Day": "05"}, "Language": ["eng"],
 "ELocationID": ["10.1080/21623945.2020.1870060"], "ArticleTitle": "Brown and beige
 adipose tissue: a novel therapeutic strategy for obesity and type 2 diabetes
 mellitus.", "KeywordList": [["Brown adipose tissue", "beige adipose tissue",
 "browning of white adipose tissue", "obesity", "type 2 diabetes mellitus", "white
 adipose tissue"]], "AbstractText": ["Mammalian adipose tissue can be divided into
 two major types, namely, white adipose tissue (WAT) and brown adipose tissue (BAT).
 According to classical view, the main function of WAT is to store excess energy in
 the form of triglycerides, while BAT is a thermogenic tissue that acts a pivotal
 part in maintaining the core body temperature. White adipocytes display high
 plasticity and can transdifferentiate into beige adipocytes which have many similar
 morphological and functional properties with brown adipocytes under the stimulations
 of exercise, cold exposure and other factors. This phenomenon is also known as
 'browning of WAT'. In addition to transdifferentiation, beige adipocytes can also
 come from de novo differentiation from tissue-resident progenitors. Activating BAT
 and inducing browning of WAT can accelerate the intake of glycolipids and reduce the
 insulin secretion requirement, which may be a new strategy to improve glycolipids
 metabolism and insulin resistance of obese and type 2 diabetes mellitus (T2DM)
 patients. This review mainly discusses the significance of brown and beige adipose
 tissues in the treatment of obesity and T2DM, and focuses on the effect of the
 browning agent on obesity and T2DM, which provides a brand-new theoretical reference
 for the prevention and treatment of obesity and T2DM."], "MedlineJournalInfo":
 {"Country": "United States", "MedlineTA": "Adipocyte", "NlmUniqueID": "101567863",
 "ISSNLinking": "2162-3945"}, "PublicationStatus": "ppublish", "ArticleIdList":
 ["33403891", "10.1080/21623945.2020.1870060", "PMC7801117"]}
```


Fonte: Elaborado pelo autor (2024)

3.4.3.2 Análise Exploratória de Dados

A EDA é uma etapa fundamental em qualquer projeto relacionado a dados, sendo a primeira fase crucial para a compreensão dos conceitos subjacentes presentes nos dados. O EDA permite descobrir características estatísticas e padrões nos dados, facilitando uma visão inicial que orienta as etapas subsequentes de análise. Por exemplo, ao analisar petições online submetidas ao Conselho da Cidade de Kyiv, a EDA foi utilizada para identificar padrões e características estatísticas das petições, mostrando a popularidade do serviço de petições e as palavras mais frequentes utilizadas. A aplicação do EDA é especialmente valiosa para dados textuais, pois ajuda a revelar insights importantes sobre a estrutura e o conteúdo dos textos, permitindo a identificação de tendências e anomalias. Em suma, o EDA é uma ferramenta indispensável para a exploração e compreensão preliminar dos dados, servindo como base para análises mais aprofundadas e decisões informadas (SAMVELYAN; SHAPTALA; KYSELOV, 2020).

Para OLULEYE (2023), além de sua aplicação geral, o EDA é frequentemente necessária para dados textuais, dada a quantidade significativa de dados digitais criados diariamente, incluindo e-mails, postagens em redes sociais e

mensagens de texto. Dados textuais são classificados como dados não estruturados, pois geralmente não aparecem em linhas e colunas. A seguir, apresento algumas técnicas comuns de EDA para dados textuais, destacando que esta lista é apenas um exemplo do que é possível, e que essas etapas não são obrigatórias e podem não ser necessárias em sua totalidade ou em parte, dependendo do objetivo a ser alcançado:

- *Preparing text data*: Preparação dos dados textuais para análise, incluindo limpeza e padronização.
- *Removing stop words*: Remoção de palavras comuns que geralmente não acrescentam valor analítico, como “e”, “de”, “o”.
- *Analyzing part of speech* (POS): Análise das partes do discurso (substantivos, verbos, etc.) para entender a estrutura gramatical do texto.
- *Performing stemming and lemmatization*: Redução das palavras às suas raízes ou formas base para simplificar a análise.
- *Analysing Ngrams*: Análise de sequências de N palavras para identificar padrões e combinações frequentes.
- *Creating word clouds*: Criação de nuvens de palavras para visualizar a frequência e a relevância das palavras no texto.
- *Checking term frequency*: Verificação da frequência dos termos para identificar as palavras mais comuns no corpus.
- *Checking sentiments*: Análise de sentimentos para determinar o tom emocional do texto (positivo, negativo, neutro).
- *Performing topic modeling*: Modelagem de tópicos para identificar temas latentes nos textos.
- *Choosing an optimal number of topics*: Seleção do número ideal de tópicos para modelagem, garantindo uma representação adequada dos dados.

Para este trabalho, utilizando a massa de dados obtida de artigos, foram geradas algumas estatísticas sobre a produção científica mundial e regional sobre o tema diabetes, tanto no que se refere à produção por país quanto por idioma. Foi feita a análise da evolução temporal dessa produção, considerando tanto a quantidade de artigos criados quanto a quantidade de palavras utilizadas nos abstracts. Também foram aplicadas técnicas como BoW e Topic Modelling para entender a abrangência das pesquisas. Todas as informações obtidas,

especialmente as relativas aos textos, podem de alguma forma contribuir para o enriquecimento de um KG, tornando possível um maior número de inferências para a descoberta de novos conhecimentos.

Conforme já mencionado, o EDA constitui uma etapa fundamental para a compreensão dos dados, aplicando-se tanto à sua estrutura e qualidade quanto à área a que pertencem. A aplicação de EDA foi essencial para que este pesquisador obtivesse uma visão mais abrangente e concisa dos textos analisados, bem como de suas áreas de conhecimento e complexidade. Isso é especialmente importante para alguém sem formação ou conhecimento aprofundado sobre o assunto tratado nos dados, permitindo avaliações e entendimentos valiosos, mesmo que não tão rigorosos quanto os de um especialista da área.

O EDA é uma tarefa circular que influencia outras ações e etapas, como o pré-processamento e a transformação de dados. Esta seção teve como objetivo dar ao leitor uma introdução ao EDA sem entrar em detalhes, que serão abordados nas próximas seções.

3.4.3.3 Pré-processamento

O pré-processamento de dados foi aplicado aos registros obtidos do PubMed e incluiu tarefas como limpeza, normalização, transformação e redução de dados, eliminando inconsistências e ruídos e harmonizando diferentes formatos. Este processo não só assegurou a confiabilidade dos resultados analíticos, mas também ajudou a identificar padrões ocultos e a revelar correlações essenciais para decisões informadas. A etapa foi dividida em duas fases devido à presença de textos com informações irrelevantes. Na primeira fase, foram realizadas conversões básicas, incluindo a transformação de formatos de data, essenciais para as análises subsequentes. As atividades de pré-processamento de dados incluíram:

- Remoção de substrings específicas
- Remoção de elementos HTML
- Substituição de múltiplas vírgulas por uma única vírgula
- Substituição de espaço vírgula por vírgula
- Substituição de múltiplos espaços por um único espaço
- Remoção de símbolos específicos

- Remoção de espaços no início e fim das strings
- Preenchimento de valores ausentes
- Substituição de strings vazias ou espaços em branco
- Remoção de stop words
- Expansão de contrações para formas completas
- Remoção de caracteres de pontuação

3.4.3.4 Transformação dos dados

No decorrer desta pesquisa, diversas atividades de transformação de dados foram realizadas para garantir a qualidade e a consistência dos dados utilizados nas análises. Entre as atividades de transformação, destacam-se a conversão de strings para minúsculas (*casefold*), a remoção de acentuação (*unidecode*), a conversão de datas para objetos *datetime*, e a criação de uma nova coluna “group” baseada na condição do idioma. Além disso, técnicas de stemming e lematização foram aplicadas utilizando o PorterStemmer e o WordNetLemmatizer, respectivamente, para reduzir as palavras às suas formas raiz ou base, facilitando a análise subsequente.

Adicionalmente a pesquisa fez uso extensivo de embeddings de palavras e texto, especialmente nas atividades de Topic Modelling, NER e RE; as duas últimas essenciais para a geração das triplas da representação de conhecimento.

3.4.3.5 Representação de Conhecimento

Esta etapa envolve a extração de informações utilizando técnicas de NER e RE para a criação das triplas (sujeito, predicado, objeto). As triplas são uma representação de entidade-relacionamento-entidade, que é a forma mais simples e compacta de representar um KG ou Grafo. A criação de um KG através deste método não requer treinamento prévio, sendo suficiente a disponibilidade de artigos científicos, sejam completos ou apenas seus abstracts, dos quais as entidades e seus relacionamentos são extraídos.

A utilização de NER permite identificar e classificar automaticamente as entidades mencionadas nos textos, como nomes de pessoas, organizações, locais, doenças, tratamentos, químicos, entre outros. Já a RE é responsável por determinar as relações existentes entre essas entidades, como “trata”, “causa”, “localizado em”, etc. Juntas, essas técnicas transformam textos não estruturados em um conjunto estruturado de triplas, que podem ser usadas para construir um KG.

A vantagem deste método é a sua eficiência e eficácia na construção de KGs a partir de grandes volumes de dados textuais. Não é necessário um processo de treinamento complexo, o que facilita a implementação e aplicação em diferentes domínios do conhecimento. O uso de técnicas avançadas de NER e RE garante que as entidades e suas relações sejam extraídas com alta precisão, contribuindo para a criação de um KG robusto e útil para diversas aplicações, como análise de tendências, descoberta de conhecimento e suporte à tomada de decisões.

Para esta etapa do estudo, o dado utilizado é o abstract, que é um texto resumido do artigo científico, contendo as principais ideias e entidades da pesquisa. Um exemplo de texto utilizado nesta pesquisa pode ser observado na Figura 8. É importante salientar que os textos são divididos em sentenças para que cada sentença passe pelo processamento dos modelos de NER e RE. Para o mesmo texto já apresentado como exemplo, temos a primeira sentença ilustrada na Figura 9.

Figura 8 - Texto completo de Abstract utilizado no Processamento

```
control of hyperglycemia and prevention of glucose reabsorption
(glucotoxicity) are important objectives in the management of type 2
diabetes. this study deals with an oral combined dosage form design for two
anti-diabetic drugs, sitagliptin and dapagliflozin using self-nanoemulsifying
drug delivery systems (snedds). the snedds were developed using naturally
obtained bioactive medium-chainlong-chain triglycerides oil, mixed glycerides
and nonionic surfactants, and droplet size was measured followed by the test
for antioxidant activities. equilibrium solubility and dynamic dispersion
experiments were conducted to achieve the maximum drug loading. the in vitro
digestion, in vivo bioavailability, and anti-diabetic effects were studied to
compare the representative snedds with marketed product dapazin (r). the
representative snedds containing black seed oil showed excellent
self-emulsification performance with transparent appearance. characterization
of the snedds showed nanodroplets of around 50-66.57 nm in size (confirmed by
tem analysis), in addition to the high drug loading capacity without causing
any precipitation in the gastro-intestinal tract. the snedds provided higher
antioxidant activity compared to the pure drugs. the in vivo pharmacokinetic
parameters of snedds showed significant increase in c max (1.99 +- 0.21 mg
ml -1 ), auc (17.94 +- 1.25 mg ml -1 ), and oral absorption (2-fold) of
dapagliflozin compared to the commercial product in the rat model. the
anti-diabetic studies showed the significant inhibition of glucose level in
treated diabetic mice by snedds combined dose compared to the single drug
therapy. the combined dose of sitagliptin-dapagliflozin using snedds could be
a potential oral pharmaceutical product for the improved treatment of type 2
diabetes mellitus.
```

Fonte: Elaborado pelo autor (2024)

Figura 9 - Sentença extraída, do abstract, para posterior processamento.

```
'control of hyperglycemia and prevention of glucose reabsorption
(glucotoxicity) are important objectives in the management of type 2
diabetes.'
```

Fonte: Elaborado pelo autor (2024)

3.4.3.6 Explicitação do Conhecimento

A explicitação do conhecimento ocorre através da criação das triplas, resultando na geração do KG. O KG estabelece relações entre as entidades extraídas, proporcionando uma forma de visualizar as diferentes entidades interconectadas. Dessa maneira, permite-se que os pesquisadores naveguem e analisem as diversas entidades identificadas e suas conexões, facilitando a realização de estudos.

Uma vez gerado, o KG pode ser armazenado como um arquivo HTML navegável, oferecendo uma interface interativa para exploração e análise das informações estruturadas.

3.4.4 Demonstração

A demonstração do método proposto será realizada através da experimentação com o conjunto de dados obtidos e descritos na Seção 3.4.3.1.

3.4.5 Avaliação

Para avaliar o experimento dentro da estrutura do DSRM, serão adotadas três abordagens complementares. Primeiramente, será realizada uma avaliação por observação das triplas resultantes da execução do método. Esta análise qualitativa inicial permitirá verificar a coerência e a relevância das entidades e relacionamentos extraídos, proporcionando uma visão preliminar da eficácia do método proposto.

Em seguida, serão avaliadas as métricas estruturais e de centralidade do KG. Utilizando a biblioteca NetworkX em Python, serão calculadas as seguintes métricas:

- Densidade do grafo: Mede o quão interconectadas estão as entidades no grafo. A densidade é a razão entre o número de arestas presentes no grafo e o número máximo possível de arestas. Um valor de densidade alto indica um grafo mais densamente conectado.
- Diâmetro do grafo: Representa a maior distância entre dois nós no grafo. O diâmetro é a maior distância geodésica entre quaisquer dois nós, mostrando a extensão do grafo. Em grafos desconectados, esta métrica não é aplicável.
- Coeficiente de agrupamento médio: Indica o grau de agrupamento entre as entidades. O coeficiente de agrupamento é a média dos coeficientes de todos os nós, onde cada coeficiente de nó mede a proporção de triângulos que envolvem esse nó.
- Centralidade de grau: Mede a importância de uma entidade com base no número de conexões diretas que ela possui. Um valor alto de centralidade de grau indica que a entidade está diretamente conectada a muitas outras entidades.
- Centralidade de proximidade: Avalia a importância de uma entidade com base na sua proximidade com todas as outras entidades no grafo. Entidades com alta centralidade de proximidade têm uma menor distância média para todas as outras entidades.
- Centralidade de intermediação: Mede a influência de uma entidade no controle do fluxo de informações entre outras entidades. Entidades com alta centralidade de intermediação frequentemente estão nos caminhos mais curtos que ligam pares de outras entidades.
- Estas métricas fornecerão uma avaliação detalhada da topologia do grafo, permitindo identificar as entidades mais influentes e os padrões de conexão predominantes.

Por fim, será testada e avaliada a possibilidade de utilizar ontologia específica da área de biomedicina para a avaliação automática das triplas. A integração de ontologias permitirá validar as entidades e relações extraídas com base em um conhecimento previamente estruturado, aumentando a precisão e a

confiabilidade das inferências realizadas pelo KG. Esta abordagem combinada de avaliação qualitativa e quantitativa possibilitará uma análise abrangente do desempenho e da utilidade do método proposto.

3.4.6 Comunicação

A comunicação das análises e discussão dos resultados desta pesquisa será realizada através de publicações científicas na forma de artigos em periódicos ou conferências, e após a conclusão da pesquisa, na forma deste documento final de dissertação. Pretende-se também solicitar o registro do protótipo desenvolvido para a avaliação do método, que consiste basicamente em um módulo de coleta de dados, um módulo de extração de texto de diversos tipos de mídias, um módulo de representação de conhecimento, um módulo de monitoramento e análise da qualidade do KG e um módulo de consulta do KG. Adicionalmente, deverá ser implementado pelo menos um módulo de inferência de dados e descoberta de novos conhecimentos.

3.5 SÍNTESE DA METODOLOGIA DE PESQUISA

Neste capítulo, foram apresentados os procedimentos metodológicos adotados nesta dissertação. Quanto à natureza, a pesquisa se caracteriza como tecnológica e aplicada, com objetivos exploratórios. A metodologia utilizada para a construção do método foi a Design Science Research Methodology (DSRM). Por fim, foram descritas as etapas de desenvolvimento da pesquisa de maneira geral, tendo como base a DSRM, sendo a síntese das atividades apresentada no Quadro 4.

Quadro 4 - Síntese das atividades desenvolvidas nesta pesquisa.

Atividade	Descrição
Identificação do problema e motivação	– Revisão integrativa da literatura para definir o problema e a motivação.
Definição dos objetivos para uma solução	– Revisão integrativa da literatura para delinear os objetivos da pesquisa.
Projetar e desenvolver	– Coleta de dados: obtenção de artigos científicos da base Pub-

	Med. – Exploratory Data Analysis (EDA): análise preliminar para compreensão da estrutura e características dos dados. – Pré-processamento: limpeza, normalização, transformação e redução de dados. – Transformação dos dados: conversão de strings para minúsculas, remoção de acentuação, conversão de datas para objetos datetime, criação de novas colunas e aplicação de técnicas de stemming e lematização. – Extração de informações: utilização de NER e RE para criar triplas (sujeito, predicado, objeto). – Construção do Knowledge Graph (KG): integração das triplas em um grafo de conhecimento.
Demonstração	– Cenário de estudo: aplicação do método nos abstracts dos artigos científicos sobre diabetes mellitus. – Visualização do KG gerado, permitindo que pesquisadores naveguem e analisem as entidades e suas conexões.
Avaliação	– Avaliação inicial por observação das triplas resultantes. – Avaliação das métricas estruturais e de centralidade do KG usando NetworkX (densidade, diâmetro, coeficiente de agrupamento, centralidade de grau, centralidade de proximidade, centralidade de intermediação). – Testar e avaliar a possibilidade de utilizar ontologias específicas da área para a validação automática das triplas.
Comunicação	– Apresentação dos resultados e considerações finais na dissertação. - Publicação de artigos científicos em periódicos e conferências. – Solicitação de registro de patente do protótipo desenvolvido.

Fonte: Elaborado pelo autor (2024)

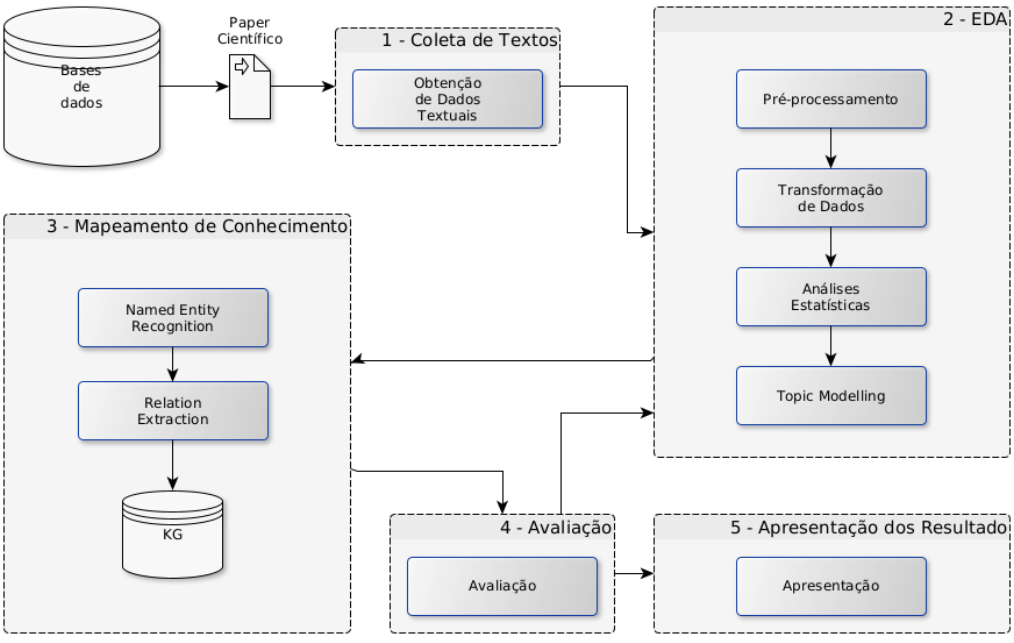
4 MÉTODO PROPOSTO

Este capítulo apresenta de forma detalhada o método proposto nesta dissertação, expondo as etapas e o funcionamento do mesmo. Ressalta-se que, a partir da revisão integrativa da literatura, não foram identificados trabalhos nesta linha que tenham como tema artigos sobre diabetes mellitus. Após a apresentação geral do modelo, é realizada uma instanciação, onde são apresentados os componentes técnicos e tecnológicos, visando clarificar todas as etapas e suas interconexões.

4.1 APRESENTAÇÃO DO MÉTODO

O método proposto é composto por uma série de etapas, elaboradas com base na revisão integrativa da literatura e na fundamentação teórica, com a finalidade de responder à pergunta de pesquisa e atingir os objetivos geral e específicos. A seguir, serão apresentadas as etapas do modelo com uma breve descrição, conforme ilustrado na Figura 10.

Figura 10 - Etapas do método proposto



Fonte: Elaborado pelo autor (2024)

A Etapa 1 (Coleta de Textos) se destina à coleta de textos científicos relevantes a partir de bases de dados, como a PubMed. Nesta fase, são obtidos os dados textuais que servirão como insumo para todas as etapas subsequentes. A coleta de textos é crucial para garantir que o conjunto de dados seja representativo e abrangente, permitindo uma análise aprofundada sobre o tema diabetes mellitus.

A Etapa 2 (*Exploratory Data Analysis* - EDA) se concentra na Análise Exploratória de Dados. Esta fase inclui o pré-processamento dos dados, onde são aplicadas técnicas de limpeza, normalização e transformação para eliminar inconsistências e ruídos. Em seguida, são realizadas análises estatísticas para compreender a distribuição e as características dos dados. Além disso, são aplicadas técnicas de modelagem de tópicos (*Topic Modelling*) para identificar os principais temas abordados nos textos científicos.

A Etapa 3 (Representação de Conhecimento) é dedicada à representação de conhecimento. Utilizando técnicas de *Named Entity Recognition* (NER) e *Relation Extraction* (RE), são extraídas as entidades e suas relações presentes nos textos. As informações extraídas são então organizadas em triplas (sujeito, predicado, objeto) e integradas em um *Knowledge Graph* (KG). O KG é uma representação estruturada do conhecimento, permitindo uma visualização clara das interconexões entre as entidades.

A Etapa 4 (Avaliação) envolve a análise contínua dos dados e do *Knowledge Graph* (KG) gerado. Esta fase inclui a aplicação de métricas estruturais e de centralidade fazendo uso de algoritmos específicos para avaliação da qualidade e a coerência do KG. A avaliação também pode incluir o uso de ontologias específicas da área para validar automaticamente as triplas extraídas, assegurando a precisão e a relevância das informações representadas. Com base nos resultados obtidos nas métricas, e considerando parâmetros que definem os valores mínimos aceitáveis para essas métricas, o fluxo pode exigir um retorno à Etapa 2. Esse retorno visa obter mais informações que possam levar a ajustes no pré-processamento dos dados e, consequentemente, a melhorias nas métricas de avaliação. Assim, o fluxo não é estritamente linear, mas sim cíclico, permitindo iterações que busquem continuamente otimizar os resultados obtidos.

A Etapa 5 (Apresentação dos Resultados) se destina à apresentação dos resultados. Os resultados obtidos são comunicados à comunidade acadêmica através de publicações científicas em periódicos e conferências. Além disso, os resultados e as considerações finais são detalhados na dissertação. Também é

prevista a solicitação de registro de patente do protótipo desenvolvido, incluindo os módulos de coleta de dados, extração de texto, representação de conhecimento, monitoramento e análise da qualidade do KG, e consulta do KG.

4.2 COMPOSIÇÃO DO MÉTODO

Nas subseções a seguir, são apresentados mais detalhes sobre cada uma das etapas do método proposto.

4.2.1 Etapa 1: Coleta de Textos

Esta se destina à coleta de textos científicos relevantes a partir de bases de dados, como a PubMed. Esta fase é crucial para garantir que o conjunto de dados seja representativo e abrangente, permitindo a análise sobre o tema diabetes mellitus. O processo de coleta envolve várias atividades, começando com a identificação das palavras-chave e termos de pesquisa específicos para a área de estudo. A seleção criteriosa das palavras-chave assegura que os artigos coletados estejam alinhados com os objetivos da pesquisa. Este assunto é abordado na seção 3.4.3.1.

Uma vez definidas as palavras-chave, utiliza-se ferramentas de busca avançadas disponíveis nas bases de dados para localizar e extrair os artigos científicos pertinentes. Durante essa fase, são aplicados filtros para restringir os resultados a publicações recentes e de alta relevância, garantindo a atualidade e a qualidade dos dados coletados. Além disso, é importante assegurar a inclusão de estudos de diversas regiões e contextos, para obter uma visão ampla e diversificada do problema em análise.

Após a coleta, os dados textuais são armazenados de forma organizada em um banco de dados estruturado, facilitando o acesso e a manipulação nas etapas subsequentes do método. Este banco de dados deve ser capaz de suportar grandes volumes de dados e permitir operações eficientes de consulta e recuperação de informações. A organização cuidadosa e a catalogação dos textos coletados são essenciais para a eficácia das etapas seguintes, especialmente durante o EDA e a representação de conhecimento. Esta fase estabelece a base sobre a qual todo o

processo de representação de conhecimento será construído, tornando-a fundamental para o sucesso do método proposto.

4.2.2 Etapa 2: *Exploratory Data Analysis*

A de EDA é fundamental para compreender a estrutura e as características dos dados coletados na Etapa 1. Esta fase envolve subetapas importantes, como o pré-processamento, a transformação de dados, a análise estatística e a modelagem de tópicos. O pré-processamento inclui a limpeza, normalização e transformação dos dados para eliminar inconsistências e ruídos. A análise estatística ajuda a compreender a distribuição dos dados, enquanto a modelagem de tópicos (Topic Modelling) identifica os principais temas abordados nos textos científicos. O EDA prepara os dados para as etapas subsequentes e proporciona insights preliminares valiosos para a pesquisa.

4.2.3 Etapa 3: Representação de Conhecimento

Esta etapa é dedicada à extração e organização das informações presentes nos textos coletados. Utilizando técnicas de NER e RE, são identificadas as entidades e suas relações dentro dos textos científicos. As entidades podem incluir termos como doenças, medicamentos, genes, entre outros, enquanto as relações descrevem como essas entidades interagem entre si.

As informações extraídas são organizadas em triplas (sujeito, predicado, objeto), que representam as relações entre as entidades de forma estruturada. Essas triplas são então integradas em um KG, que é uma estruturada do conhecimento extraído, e que possui fácil representação visual. O KG permite visualizar as interconexões entre as entidades, facilitando a identificação de padrões e insights valiosos.

Esta etapa é crucial para transformar os dados textuais não estruturados em uma forma que pode ser facilmente analisada e utilizada para descobertas e inferências futuras. O KG gerado serve como a base para a análise e interpretação das informações extraídas, proporcionando uma visão clara e organizada do conhecimento presente nos textos científicos.

4.2.4 Etapa 4: Avaliação

A etapa de avaliação envolve a análise contínua dos dados e do KG gerado. Esta fase é essencial para mensurar a qualidade e a coerência das informações extraídas e representadas. A avaliação é realizada em várias etapas, começando pela observação inicial das triplas resultantes do processo de extração. Esta análise qualitativa permite identificar possíveis inconsistências e refinar o método de extração.

Além da observação qualitativa, são aplicadas métricas estruturais e de centralidade utilizando a biblioteca NetworkX. As métricas estruturais incluem a densidade, o diâmetro e o coeficiente de agrupamento do grafo. A densidade mede o quão interconectadas estão as entidades no grafo; valores altos indicam uma rede mais densa de conexões. O diâmetro representa a maior distância entre dois nós no grafo, fornecendo uma ideia da “largura” do conhecimento representado; um diâmetro menor sugere que as informações estão mais próximas umas das outras. O coeficiente de agrupamento mede o grau de interconexão entre os vizinhos de um nó; valores altos indicam uma rede bem estruturada, com grupos fortemente interligados.

As métricas de centralidade incluem a centralidade de grau, a centralidade de proximidade e a centralidade de intermediação. A centralidade de grau mede a importância de uma entidade com base no número de conexões diretas que ela possui; entidades com alta centralidade de grau são bem conectadas e potencialmente influentes. A centralidade de proximidade avalia a proximidade de uma entidade em relação a todas as outras entidades no grafo; entidades com alta centralidade de proximidade podem acessar rapidamente outras partes do grafo, sugerindo eficiência na propagação de informações. A centralidade de intermediação mede a influência de uma entidade no controle do fluxo de informações entre outras entidades; entidades com alta centralidade de intermediação são cruciais para a mediação de informações e podem conectar diferentes partes do grafo.

A avaliação contínua também pode incluir o uso de ontologias específicas da área de biomedicina para validar automaticamente as triplas extraídas. Esta validação automática assegura a precisão e a relevância das informações

representadas, aumentando a confiabilidade do KG. A etapa de avaliação é um ciclo iterativo que pode levar a ajustes no pré-processamento e na extração de informações, garantindo a melhoria contínua da qualidade dos dados e do KG.

4.2.5 Etapa 5: Apresentação dos Resultados

A etapa final se destina a comunicar os achados da pesquisa à comunidade acadêmica e ao público em geral. Esta fase é crucial para disseminar o conhecimento gerado e validar o método proposto. Os resultados são compartilhados através de publicações científicas em periódicos e apresentações em conferências, detalhando a metodologia, as descobertas e a relevância do KG gerado. Além disso, os resultados e considerações finais são detalhados na dissertação, que fornece um relato abrangente de todo o processo de pesquisa, desde a coleta de textos até a avaliação do KG.

Também é prevista a solicitação de registro de patente do protótipo desenvolvido, que inclui módulos de coleta de dados, extração de texto, representação de conhecimento, monitoramento e análise da qualidade do KG, e consulta do KG. A patente protege a inovação e facilita a transferência de tecnologia, permitindo que a solução desenvolvida seja utilizada por outras instituições e pesquisadores. A comunicação eficaz dos resultados garante que o conhecimento gerado pela pesquisa seja amplamente disseminado e aproveitado, contribuindo para o avanço científico e tecnológico na área de estudo.

4.3 INSTANCIAÇÃO DO MODELO

O propósito desta seção é demonstrar, através de um exemplo, a instanciação do método proposto, visando clarificar o seu funcionamento. Pretende-se, desta forma, apresentar os componentes do método e as interligações entre eles, assim como as técnicas e tecnologias envolvidas, que serão descritas em cada etapa.

4.3.1 Etapa 1: Coleta de Textos

A coleta de textos iniciou-se com a consulta aos metadados dos artigos científicos utilizando o pacote BioPython, especificamente o módulo Entrez. Onde, para operar corretamente o pacote, é necessário fornecer um endereço de e-mail, com a opção de realizar consultas de forma anônima ou identificada por uma chave de acesso. A principal diferença entre essas modalidades está no limite de consultas por minuto: três para consultas anônimas e dez para consultas identificadas.

O processo de consulta foi dividido em duas etapas principais. Na primeira etapa, foram obtidos os IDs dos artigos a partir do termo de pesquisa "*diabetes mellitus"*, resultando em 518.432 IDs. Na segunda etapa, procedeu-se à obtenção dos metadados correspondentes a cada ID. Devido ao volume substancial de dados e ao tempo de resposta, as consultas foram realizadas em lotes de 4.000 artigos cada. A API permite a definição do formato da resposta através de parâmetros, e optou-se pelo formato JSON para facilitar a seleção de características relevantes.

Um exemplo de registro obtido pode ser observado na Figura 7 apresentada anteriormente.

4.3.2 Etapa 2: Exploratory Data Analysis

A etapa de EDA envolveu várias atividades essenciais para preparar e analisar os dados coletados. Uma das primeiras atividades foi a Seleção de Atributos, que apesar de ser realizada uma única vez para cada fonte de dados que possa ser usada, é crucial para reduzir o uso de memória na base de dados e preservar apenas os atributos relevantes para a compreensão dos dados e o alcance dos objetivos da pesquisa.

A seleção de atributos ocorreu em duas etapas principais. Primeiramente, selecionam-se atributos relacionados à geolocalização, idioma, palavras-chave, ID, título e resumo. O conjunto original de metadados contém uma ampla variedade de atributos, como autores e instituições de filiação, que não são relevantes para este estudo específico sobre *diabetes mellitus*. Ao filtrar esses atributos irrelevantes, reduz-se o ruído e o risco de incluir dados que possam prejudicar a análise.

Após esta primeira filtragem, é acrescentado o atributo "DB", uma constante ("PubMed"), introduzida por este pesquisador devido à possibilidade de incluir outras bases de dados no futuro. A Figura 11 apresenta os atributos selecionados nesta

primeira fase. Esta etapa garante que os dados utilizados nas análises subsequentes sejam concisos e diretamente relacionados ao tema de interesse.

Figura 11 - Estrutura resultante da primeira seleção de features.

```
RangeIndex: 518432 entries, 0 to 518431
Data columns (total 16 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   DB                 518432 non-null  object
 1   CLASS              518432 non-null  object
 2   ID                 518432 non-null  object
 3   DateCompleted      495100 non-null  object
 4   DateCompletedOld   495100 non-null  object
 5   DateRevised        518432 non-null  object
 6   DateRevisedOld     518432 non-null  object
 7   Language           518432 non-null  object
 8   ELocationID        518432 non-null  object
 9   ArticleTitle       518432 non-null  object
 10  MedlinePgn         509869 non-null  object
 11  Country            518424 non-null  object
 12  KeywordList        518432 non-null  object
 13  AbstractText       518432 non-null  object
 14  PublicationStatus  518432 non-null  object
 15  ArticleIdList      518432 non-null  object
dtypes: object(16)
```

Fonte: Elaborado pelo autor (2024)

A segunda etapa de seleção de atributos ocorre durante o EDA e envolve a análise da relevância de cada atributo, avaliando seu potencial valor para a etapa de análise dos abstracts. O conjunto de atributos resultante desta análise compõe-se de:

- ID
- Language
- Country
- AbstractText

A análise estatística da distribuição dos papers por idioma é uma etapa importante, uma vez que a produção científica em Biomedicina tende a ser predominantemente em inglês. Esta observação é crucial para justificar a escolha do inglês como idioma principal para a pesquisa. Ao analisar a produção por idioma, é possível identificar padrões de publicação globais e regionais, fornecendo insights sobre a acessibilidade e a disseminação do conhecimento científico. As quantidades de papers para os 15 países com mais papers obtidos pode ser observado na Figura 12

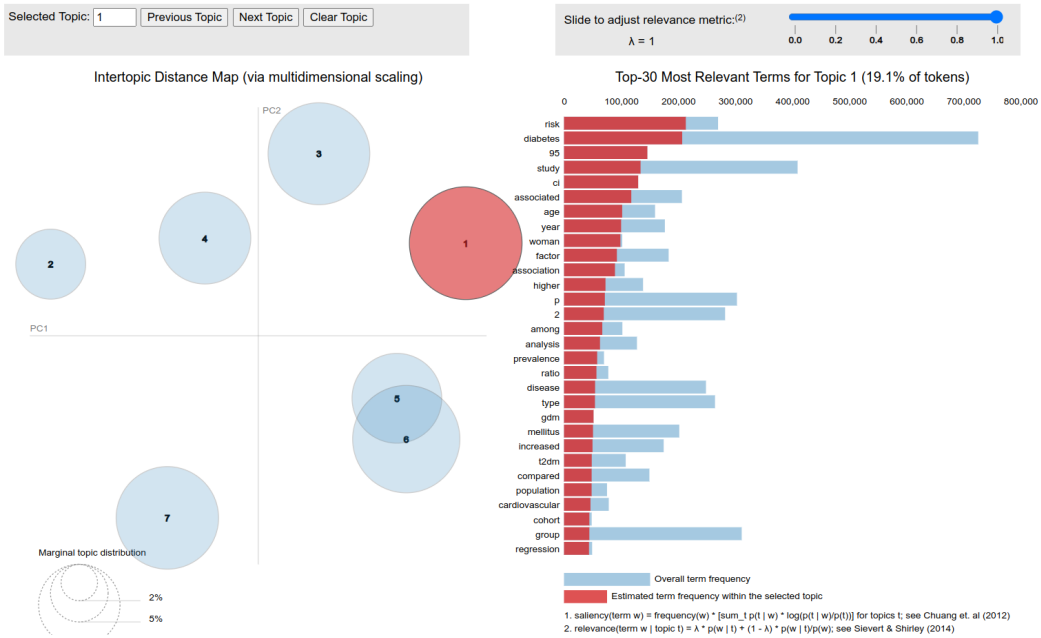
Figura 12 - Quantidade de papers por idioma

	Language	count
0	eng	443314
1	ger	13632
2	fre	11359
3	jpn	8502
4	rus	7788
5	spa	6667
6	chi	4917
7	ita	4159
8	pol	3361
9	cze	1832
10	und	1744
11	por	1325
12	hun	1068
13	eng, spa	905
14	dut	901

Fonte: Elaborado pelo autor (2024)

A aplicação de Topic Modelling foi de grande ajuda para o entendimento dos temas abordados nos textos coletados. Utilizando a técnica de Latent Dirichlet Allocation (LDA), foi possível identificar os principais tópicos presentes nos artigos científicos sobre diabetes mellitus. O LDA permitiu a categorização automática dos documentos em diferentes tópicos, proporcionando uma visão clara e organizada dos temas mais recorrentes. Para a visualização dos resultados, foi utilizado o LDAvis, uma ferramenta interativa que facilita a exploração dos tópicos gerados pelo LDA. O LDAvis permite visualizar, conforme Figura 13 a distribuição dos tópicos nos documentos, bem como a relevância de cada palavra em cada tópico, tornando o processo de análise mais intuitivo e eficiente.

Figura 13 - Visualização da aplicação do LDA em Topic Modelling



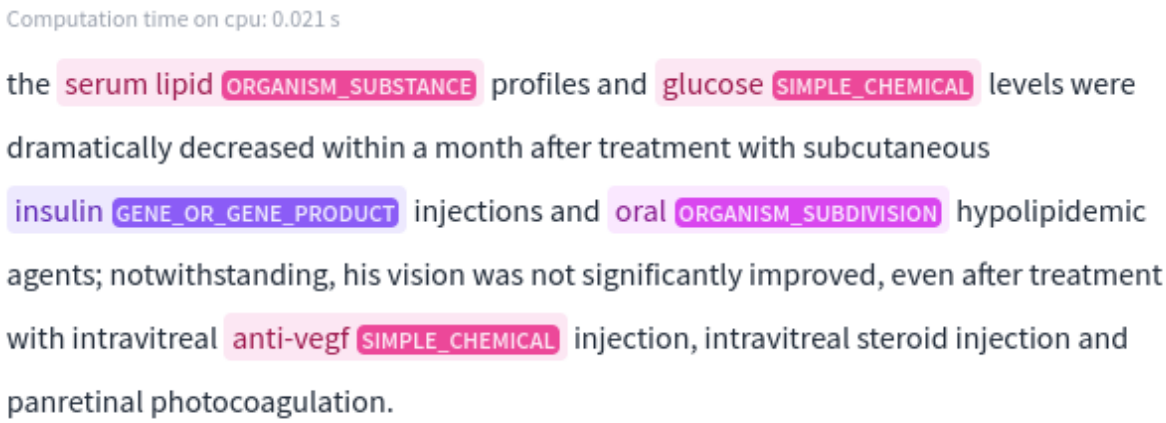
Fonte: Elaborado pelo autor (2024)

Tendo sido realizada a escolha do idioma e identificados tópicos com termos relevantes, o texto passa pelo pré-processamento a fim de prepará-lo para a próxima etapa.

4.3.3 Etapa 3: Representação de Conhecimento

Ao entrar na etapa de Representação de Conhecimento, o texto preprocessado foi submetido a um conjunto de modelos para a identificação de entidades relevantes ao método em execução. Utilizando as bibliotecas spaCy e Transformers (Hugging Face), foram executados modelos pré-treinados em NER especificamente voltados para a área da Biomedicina. Esses modelos possuem conjuntos distintos de entidades que podem ser identificadas no texto, conforme ilustrado na Figura 14.

Figura 14 - Resultado da aplicação do modelo ner_bionlp13cg_md



Fonte: Elaborado pelo autor (2024)

Este modelo em particular foi treinado para identificar 16 classes de entidades, abrangendo uma ampla gama de categorias relevantes para a pesquisa. Ao aplicar este modelo ao conjunto completo de textos, foi gerado um vasto conjunto de entidades identificadas, totalizando alguns milhões de ocorrências. Essas entidades incluem classes como doenças, medicamentos, sintomas, genes, entre outros, que são cruciais para a representação do conhecimento na área da Biomedicina. A Figura 15 ilustra a distribuição e a variedade das entidades identificadas, demonstrando a quantidade de informações detalhadas e relevantes dos textos científicos que pode ser capturada.

Figura 15 - Quantidade e entidades identificadas pelo modelo ner_bionlp13cg_md

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	ORGANISM	3492	1304959
02	GENE_OR_GENE_PRODUCT	2726	1008113
03	SIMPLE_CHEMICAL	2540	1009669
04	CELL	1010	340947
05	MULTI_TISSUE_STRUCTURE	954	335288
06	ORGAN	894	309235
07	ORGANISM_SUBSTANCE	891	349812
08	TISSUE	832	274371
09	CANCER	396	153042
10	CELLULAR_COMPONENT	341	137821
11	PATHOLOGICAL_FORMATION	241	85026
12	ANATOMICAL_SYSTEM	231	88242
13	ORGANISM_SUBDIVISION	226	98157
14	IMMATERIAL_ANATOMICAL_ENTITY	103	43417
15	AMINO_ACID	50	20272
16	DEVELOPING_ANATOMICAL_STRUCTURE	6	1595

Fonte: Elaborado pelo autor (2024)

Após a identificação das entidades em cada sentença, foi realizada a identificação dos relacionamentos possíveis entre as diversas entidades presentes na sentença em processamento. Para esta atividade foi utilizado o modelo BERT-base-SRL, o qual, em um conjunto de 1000 sentenças, identificou 2795 triplas compostas de sujeito-predicado-objeto, um relatório da execução pode ser confirmado na Figura 16.

Figura 16 - Relatório da execução do modelo BERT-base-SRL

```
RangeIndex: 2795 entries, 0 to 2794
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sujeito    2795 non-null   object
 1   Predicado  2795 non-null   object
 2   Objeto     2795 non-null   object
dtypes: object(3)
memory usage: 65.6+ KB
```

Fonte: Elaborado pelo autor (2024)

Uma amostra das triplas pode ser observada na Figura 17.

Figura 17 - Amostra de triplas geradas pelo modelo BERT-base-SRL

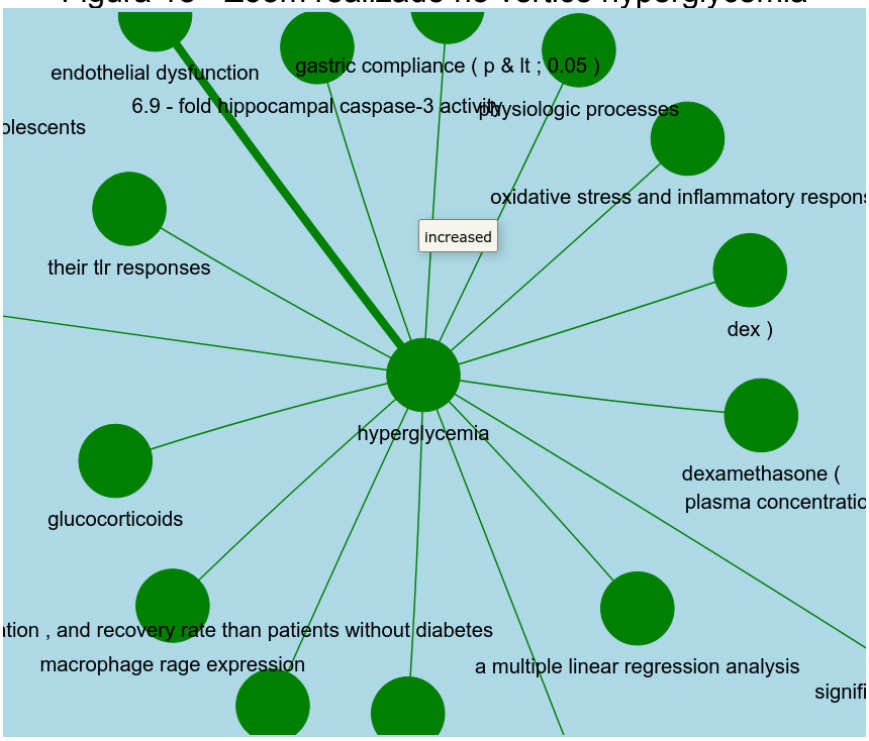
	Sujeito	Predicado	Objeto
0	reovirus type 3 , passaged in pancreatic beta ...	produced	an insulitis
1	beta cells	containing	insulin
2	by protein wasting secondary to hypergluconeog...	characterized	cushing 's syndrome
3	by insulin treatment of rats with chronic step...	induced	the increased liver dna
4	steptozotocin	induced	diabetes
5	by an active process having a half - time of 3...	restored	the dna content of the organ
6	an active process	having	a half - time of 32 days
7	a mechnanism	acts	to restore normal liver cellularity when an ov...
8	severe diabetic ketoacidosis and hyperkalemia	presented	with an ecg resembling an acute anterior wall ...
9	subsequent work - up including exercise testin...	ruled	any significant coronary artery disease

Fonte: Elaborado pelo autor (2024)

Após a geração das triplas, é possível calcular diversas métricas e visualizar as triplas geradas. Além disso, é viável aplicar algoritmos que comparam os resultados com ontologias da Biomedicina. A plotagem (visualização gráfica) dos resultados também é possível, utilizando a biblioteca NetworkX da linguagem

Python, conforme ilustrado na Figura 18. As triplas obtidas foram salvas em uma base de dados própria no formato Parquet, além de serem exportadas para os formatos RDF e HTML. Este último formato oferece uma visualização interativa das triplas geradas.

Figura 18 - Zoom realizado no vértice hyperglycemia



Fonte: Elaborado pelo autor (2024)

4.3.4 Etapa 4: Avaliação

Esta etapa tem por finalidade avaliar a qualidade e a coerência do KG gerado a partir dos textos científicos sobre diabetes mellitus. A avaliação é realizada em várias fases, utilizando diferentes abordagens para mensurar e analisar a precisão e a relevância das informações extraídas.

Inicialmente, as triplas geradas são observadas qualitativamente para identificar possíveis inconsistências e refinar o método de extração. Em seguida, são aplicadas métricas estruturais e de centralidade utilizando a biblioteca NetworkX. Esta avaliação pode ser realizada tanto utilizando a representação em triplas, conforme Figura 17, quanto através de sua forma gráfica, apresentada na Figura 18.

As métricas estruturais, que são calculadas objetivamente, incluem a

densidade, o diâmetro e o coeficiente de agrupamento do grafo. A densidade mede o quão interconectadas estão as entidades no grafo; valores altos indicam uma rede mais densa de conexões. O diâmetro representa a maior distância entre dois nós no grafo, fornecendo uma ideia da “largura” do conhecimento representado. O coeficiente de agrupamento mede o grau de interconexão entre os vizinhos de um nó; valores altos indicam uma rede bem estruturada e com grupos fortemente interligados.

As métricas de centralidade, como a centralidade de grau, a centralidade de proximidade e a centralidade de intermediação, são calculadas para avaliar a importância de cada entidade no KG. A centralidade de grau mede a importância de uma entidade com base no número de conexões diretas que ela possui. A centralidade de proximidade avalia a proximidade de uma entidade em relação a todas as outras entidades no grafo, indicando sua eficiência na propagação de informações. A centralidade de intermediação mede a influência de uma entidade no controle do fluxo de informações entre outras entidades. As métricas geradas podem ser observadas na Figura 19, mas serão discutidas no próximo capítulo.

Além das métricas quantitativas, é possível aplicar algoritmos que comparam os resultados com ontologias específicas da área de Biomedicina. Essa comparação permite validar automaticamente as triplas extraídas, assegurando a precisão e a relevância das informações representadas.

Figura 19 - Métricas geradas a partir do KG criado.

	Métrica	Valor
0	Número de Nodos	4.948000e+03
1	Número de Arestas	2.763000e+03
2	Densidade do Grafo	1.128780e-04
3	Número de Componentes Conexas	2.187000e+03
4	Tamanho do Maior Componente Conexo	6.500000e+01
5	Grau Médio	1.116815e+00
6	In-degree Médio	5.584074e-01
7	Out-degree Médio	5.584074e-01
8	Coeficiente de Agrupamento	0.000000e+00
9	Centralidade de Grau Média	2.257560e-04
10	Centralidade de Proximidade Média	1.207327e-04
11	Centralidade de Intermediação Média	3.683911e-09
12	Diâmetro do Grafo	0.000000e+00

Fonte: Elaborado pelo autor (2024)

4.3.5 Etapa 5: Apresentação dos Resultados

Esta etapa tem por finalidade comunicar os achados da pesquisa à comunidade acadêmica e ao público em geral. Os resultados obtidos, incluindo a geração do HTML e do RDF, assim como a base de triplas gerada e as métricas calculadas, serão disponibilizados através de publicações científicas e em repositórios públicos.

As publicações detalharão a metodologia, as descobertas e a relevância do KG gerado, permitindo que outros pesquisadores avaliem e construam sobre o trabalho realizado. Além disso, os resultados e considerações finais serão apresentados na dissertação, fornecendo um relato compreensivo de todo o processo de pesquisa. A disponibilização dos dados em repositórios públicos assegura que o conhecimento gerado seja amplamente acessível, contribuindo para o avanço científico e tecnológico na área de Biomedicina.

4.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o método proposto, detalhando cada uma de suas etapas e expondo seu funcionamento. Demonstrou-se a coleta e transformação dos dados necessários para viabilizá-lo, passando pela extração de entidades e relações e a construção do KG. Foram aplicadas técnicas avançadas de NLP, como NER e RE, utilizando modelos pré-treinados das bibliotecas spaCy e Transformers (Hugging Face), que permitiram identificar entidades específicas e suas interações dentro dos textos científicos.

Em seguida, as etapas do método foram apresentadas com um enfoque técnico, visando clarificar seu funcionamento por meio de uma instanciação prática. Os componentes do método, as conexões entre eles e as técnicas e tecnologias envolvidas foram detalhados, incluindo o uso de NetworkX para a visualização e análise das triplas e do KG.

A avaliação do método incluiu a aplicação de métricas estruturais e de centralidade para assegurar a qualidade e coerência do KG, assim como a análise visual das triplas formadas. A comparação com ontologias específicas da Biomedicina não foi possível devido a exigir maior processamento para haver conformidade entre os nomes de entidades existentes no KG e as apresentadas nas

ontologias, já que uma pequena diferença leva a uma comparação infrutífera entre dois strings. Os resultados, incluindo a geração de arquivos HTML e RDF, a base de triplas e as métricas calculadas, serão disponibilizados através de publicações científicas e em repositórios públicos, garantindo ampla acessibilidade e contribuindo para o avanço científico na área de Biomedicina.

A apresentação e discussão dos resultados obtidos na instanciação do método serão abordadas no próximo capítulo, onde serão analisados os impactos e a eficácia do método proposto.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Esta seção apresenta os resultados obtidos a partir da instanciação do método proposto, utilizando o conjunto de dados descrito na seção 3.4.3.1 e a avaliação conforme a seção 3.4.5. A análise dos resultados visa discutir a execução de cada uma das atividades que compõem o método, os experimentos que levaram à sua construção e os resultados obtidos com o Knowledge Graph (KG) gerado. A seguir, são detalhadas as etapas do processo, as metodologias aplicadas e os principais achados desta pesquisa.

5.1 AVALIAÇÃO DO CONJUNTO DE DADOS

O conjunto total de registros de artigos obtidos foi de 518.432, conforme descrito na seção 3.4.3.1. Este conjunto foi reduzido à medida que o EDA foi aplicado, devido à necessidade de que os registros contivessem abstracts, o que não ocorreu em todos os casos. Para simplificar o processo, optou-se também por adotar um único idioma, realizando uma análise preliminar para identificar qual idioma possuía o maior número de registros. Estas atividades ocorreram na etapa de EDA, a qual será avaliada a seguir.

5.2 AVALIAÇÃO DA ETAPA 2: EDA

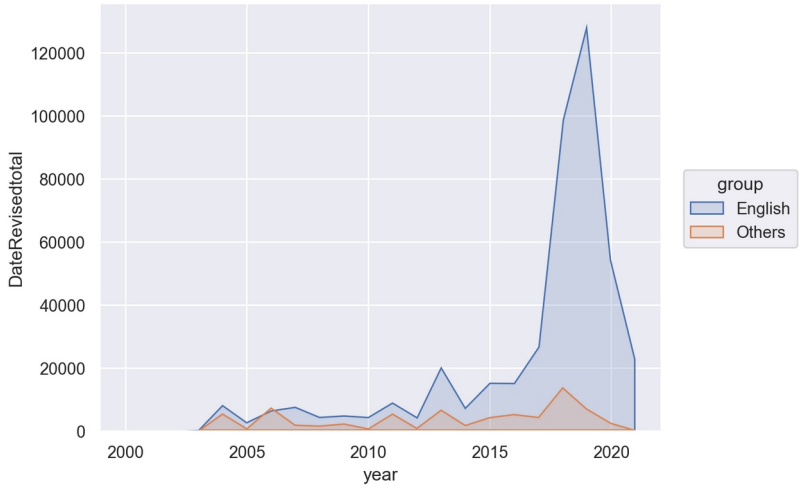
5.2.1 Análises de Distribuição dos Papers

A análise da distribuição dos artigos por idioma é uma etapa crucial, uma vez que a produção científica em Biomedicina tende a ser predominantemente realizada em inglês. Esta observação é fundamental para justificar a escolha do inglês como idioma principal para a pesquisa. Ao analisar a produção por idioma, é possível identificar padrões de publicação globais e regionais, fornecendo dados sobre a acessibilidade e a disseminação do conhecimento científico. Este processo não apenas confirma a importância do inglês como ferramenta de comunicação internacional na comunidade científica, mas também apoia a decisão de concentrar esforços na coleta de dados em fontes de língua inglesa. Isso aumenta a

representatividade e a relevância das informações obtidas para a pesquisa sobre diabetes mellitus.

Dessa forma, esta etapa não apenas reflete preferências linguísticas, mas também facilita o acesso aos dados mais relevantes e atualizados disponíveis na literatura biomédica. A análise global já foi apresentada na Figura 12, e mostrou que a produção de artigos em inglês é muito superior a outros idiomas, representando 86% do total, confirmando a escolha do idioma como apropriada para esta pesquisa. Na Figura 20 podemos ver a evolução da produção científica cadastrada no PubMed relativa a diabetes mellitus, também visualizamos a produção realizada na língua inglesa e outras.

Figura 20 - Evolução da produção anual na língua inglesa e outras

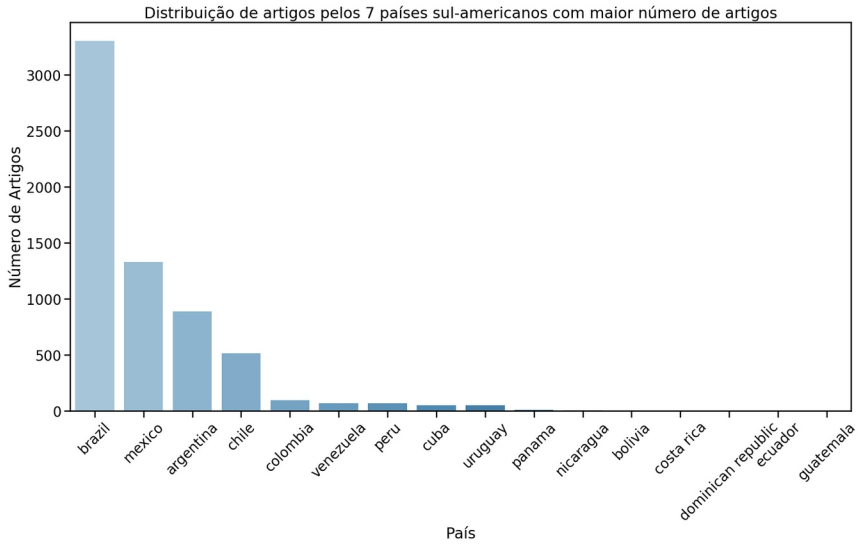


Fonte: Elaborado pelo autor (2024)

A predominância da língua inglesa na produção científica pode ser atribuída a vários fatores. Em muitos países de língua inglesa, há um grande incentivo à pesquisa, associado a um forte poder econômico e altos níveis de qualidade na educação. Além disso, a língua inglesa se consolidou como um meio de comunicação universal, de modo que, para aumentar o impacto de uma pesquisa, é vantajoso publicá-la em inglês. Esta combinação de fatores contribui para a ampla utilização do inglês na disseminação do conhecimento científico.

A análise específica da produção científica na América Latina pode ser vista na Figura 21. Os resultados não apresentam grandes surpresas quando considerados fatores como o tamanho das economias, a população e o nível de educação dos países da região.

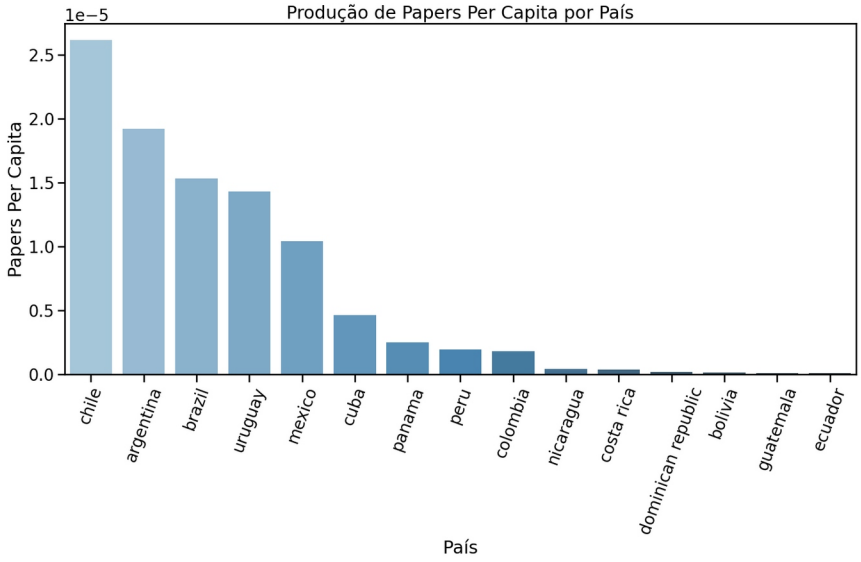
Figura 21 - Produção de papers relativos ao tema diabetes mellitus na América Latina



Fonte: Elaborado pelo autor (2024)

Adicionalmente, uma análise da produção científica per capita na América Latina foi realizada, revelando mudanças interessantes que podem ser observadas na Figura 22. Embora os mesmos fatores – tamanho das economias, população e nível de educação dos países – continuem relevantes, o peso da população é menor nesta análise. Como resultado, países com menor população se posicionam mais próximos ao topo da lista, destacando-se na produção científica per capita.

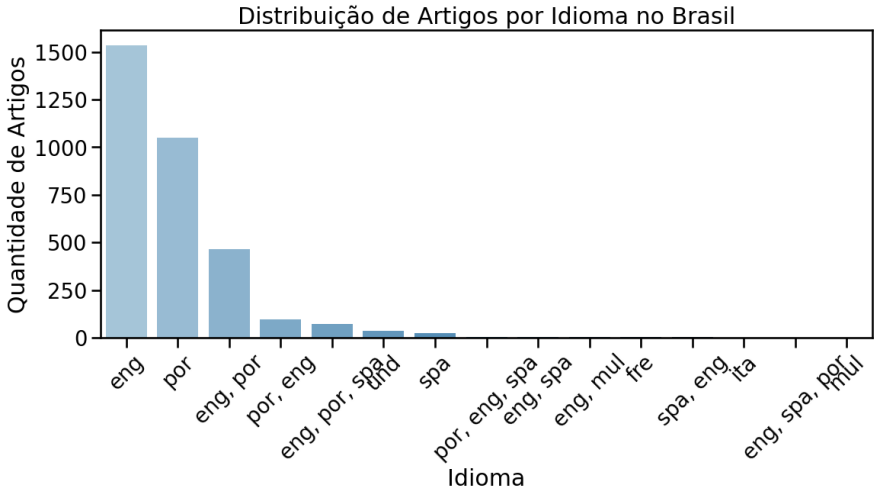
Figura 22 - Produção de artigos per capita relacionados ao tema diabetes mellitus na América Latina



Fonte: Elaborado pelo autor (2024)

De forma semelhante, foi realizada uma análise do idioma utilizado nas publicações científicas específicas do Brasil, onde o inglês também se destaca como o idioma predominante conforme exposto na Figura 23.

Figura 23 - Produção por idioma no Brasil.



Fonte: Elaborado pelo autor (2024)

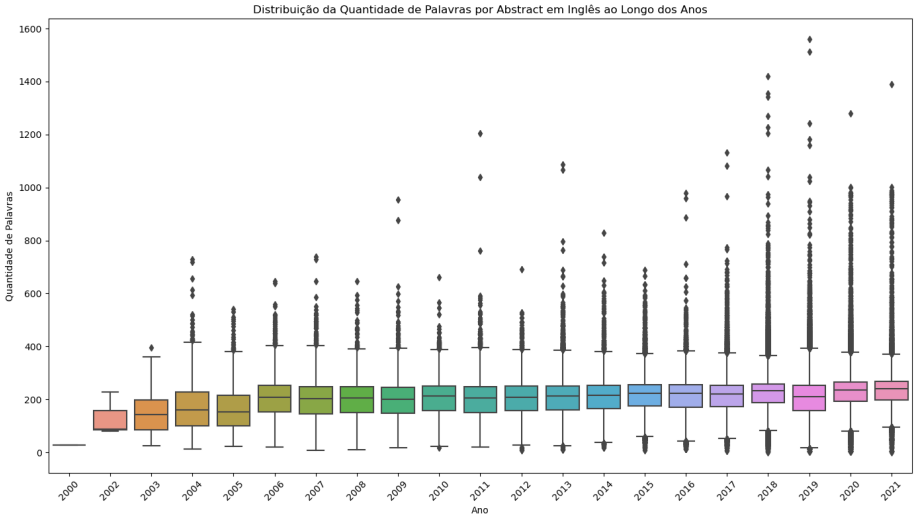
5.2.2 Análises de Distribuição das Palavras

A análise da distribuição das palavras é essencial para compreender a evolução e o foco da pesquisa científica sobre diabetes mellitus. Essa etapa revela tendências no uso de terminologia ao longo do tempo e identifica áreas de concentração e interesse na biomedicina. O aumento no número de palavras nos resumos (*abstracts*) indica pesquisas mais complexas e temas mais amplos. A análise de BoW destaca os termos mais frequentes, fornecendo insights sobre conceitos-chave e metodologias predominantes, o que é crucial para a construção de grafos de conhecimento.

O boxplot na Figura 24 mostra uma tendência de aumento na média de palavras por abstract ao longo dos anos, desde o início do milênio, com um crescimento contínuo. A maior variação no tamanho dos abstracts ao longo do tempo, refletida nos intervalos interquartílicos, e a presença de outliers, especialmente a partir de 2014, indicam um aumento significativo de abstracts mais extensos. Esse crescimento pode ser atribuído à complexidade crescente das pesquisas e às normas de publicação que exigem abstracts mais detalhados, como as diretrizes da American Psychological Association (APA).

A distribuição assimétrica na quantidade de palavras sugere que a maioria dos abstracts é mais curta, enquanto alguns são significativamente mais longos. Isso pode ser explicado pela variedade de tipos de abstracts e pela ausência de limites de palavras em algumas revistas. A evolução no perfil dos abstracts de diabetes mellitus implica a necessidade de processamento avançado de linguagem natural e no desenvolvimento de habilidades de leitura e análise pelos profissionais de saúde para lidar com textos mais extensos e complexos.

Figura 24 - Evolução por ano da quantidade de palavras usadas nos abstracts



Fonte: Elaborado pelo autor (2024)

Com os textos preprocessados, foi realizada uma análise utilizando a técnica de Bag of Words (BOW), que fornece uma visão geral dos termos mais frequentes em abstracts de artigos sobre diabetes mellitus. As palavras maiores e mais proeminentes, como “diabetes”, “mellitus”, “glucose”, “insulin” e “blood”, evidenciam o foco central da pesquisa nessa área. O resultado pode ser apreciado na Figura 25.

Figura 25 - BoW aplicado aos abstracts

Fonte: Elaborado pelo autor (2024)

A presença de outras palavras de grande importância, como “effect”, “level”, “associated”, “result” e “rat”, indica possíveis estudos laboratoriais, algo que será analisado mais detalhadamente na seção sobre Topic Modelling.

A análise léxica também incluiu a avaliação das palavras-chave atribuídas a cada artigo. Ao contrário dos resumos, onde as palavras estão inseridas em um contexto, as palavras-chave são termos que expressam os conceitos centrais do estudo. Uma análise das associações mais frequentes entre esses termos permitiu identificar pares de palavras-chave que frequentemente aparecem juntos, revelando suas relações potenciais. Essa análise resultou na criação de um grafo de palavras-chave, baseado no princípio de que os termos utilizados em um único paper estão, de alguma forma, interligados. A Figura 26 apresenta os pares de palavras-chave mais recorrentes, fornecendo um panorama das inter-relações temáticas nos estudos analisados. Os dados são apresentados em um dicionário contendo o par de palavras-chave e um valor numérico representando o número de artigos em que elas são citadas.

Figura 26 - Termos, em duplas, mais utilizados nos papers

```
{('diabetes mellitus', 'type 2'): 1285,
 ('diabetes mellitus', 'insulin'): 1143,
 ('diabetes', 'obesity'): 1107,
 ('diabetes mellitus', 'hypertension'): 1041,
 ('diabetes mellitus', 'obesity'): 929,
 ('diabetes', 'insulin'): 698,
 ('diabetes', 'hypertension'): 686,
 ('obesity', 'type 2 diabetes'): 624,
 ('insulin resistance', 'obesity'): 575,
 ('diabetes', 'oxidative stress'): 503,
 ('diabetes mellitus', 'insulin resistance'): 503,
 ('diabetes', 'inflammation'): 488,
 ('diabetes', 'insulin resistance'): 484,
 ('metabolic syndrome', 'obesity'): 481,
 ('insulin resistance', 'type 2 diabetes'): 480,
 ('diabetes mellitus', 'hyperglycemia'): 469,
 ('inflammation', 'oxidative stress'): 437,
 ('diabetes mellitus', 'risk factors'): 434,
 ('obesity', 'type 2 diabetes mellitus'): 434,
 ('diabetes mellitus', 'oxidative stress'): 433}
```

Fonte: Elaborado pelo autor (2024)

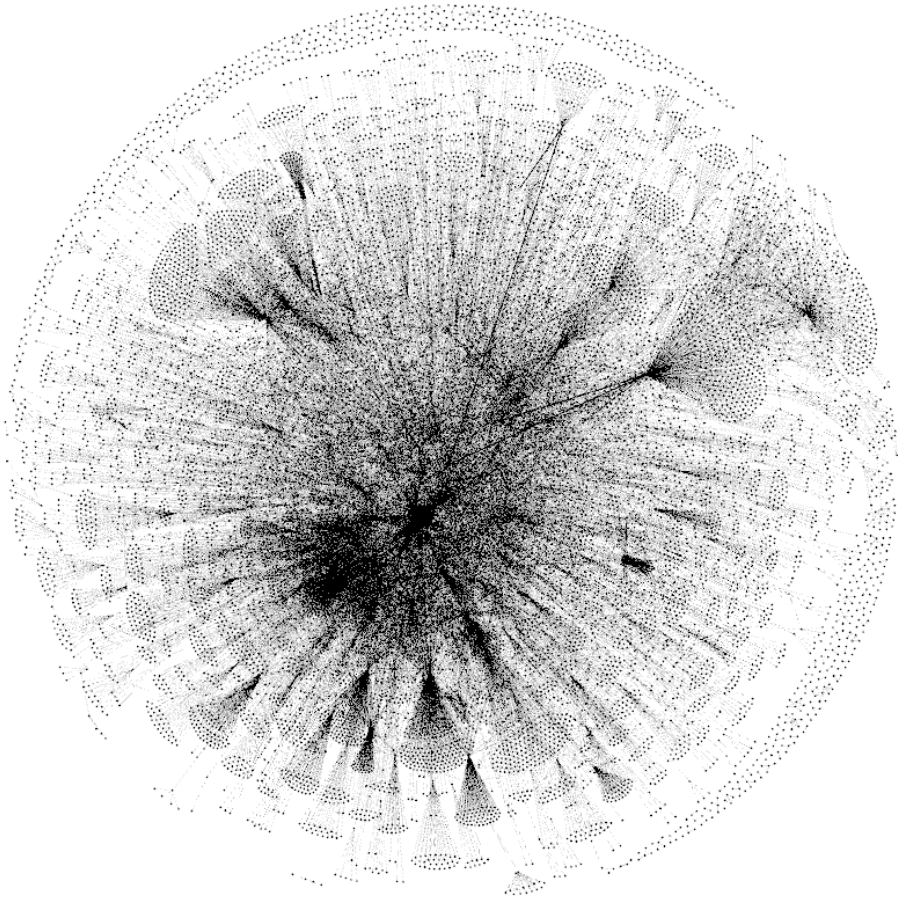
Diante das informações apresentadas na Figura 26, é possível notar que a aplicação de análises baseadas em Semantic Relation é uma abordagem promissora para o tratamento de dados. A *Semantic Relation* é uma técnica de processamento de linguagem natural que identifica e utiliza as interconexões de significado entre termos e conceitos em um corpus. Isso pode resolver os desafios apresentados por termos diversos que se referem ao mesmo conceito, como nas pesquisas sobre Diabetes Mellitus.

Inicialmente, a aplicação da Semantic Relation requer a criação de um modelo semântico capaz de representar a equivalência entre termos como “Diabetes Mellitus Type 2” e “Type 2 Diabetes”. Este modelo pode ser baseado em ontologias biomédicas ou criado por algoritmos de aprendizado de máquina treinados em grandes conjuntos de dados textuais.

A análise semântica permite identificar sinônimos e agrupá-los de forma coerente, assegurando que a referência unificada esteja contextualizada. Também distingue termos com similaridades superficiais, mas significados distintos – como “Type 1 Diabetes” e “Type 2 Diabetes”. Essa metodologia aumenta a consistência e clareza das informações em estudos de meta-análise e aprimora a recuperação de dados e a obtenção de insights significativos.

Os dados contidos na estrutura apresentada foram utilizados para a criação do grafo mostrado na Figura 27.

Figura 27 - Grafo criado utilizando relacionamento entre palavras-chave



Fonte: Elaborado pelo autor (2024)

Este grafo baseia-se na premissa de que, dentro de um único estudo científico, as palavras-chave utilizadas estão interligadas de alguma forma, embora a natureza exata dessa conexão não seja explicitada. A partir dessa premissa, diversos estudos podem ser conduzidos utilizando o grafo, incluindo:

- Agrupamentos de Palavras-Chave: O grafo revela clusters de palavras-chave que sugerem temas recorrentes em estudos sobre diabetes mellitus. Ao analisar esses agrupamentos, é possível identificar focos de pesquisa específicos e destacar áreas temáticas predominantes.
- Peso das Conexões: As arestas do grafo variam de espessura, refletindo o peso das conexões entre as palavras-chave. Linhas mais grossas indicam uma associação frequente e robusta, sugerindo que essas palavras-chave

são frequentemente mencionadas juntas, enquanto linhas mais finas podem indicar relações menos frequentes ou emergentes.

- Palavras-Chave Centrais: Algumas palavras-chave aparecem como nodais centrais no grafo, com diversas ligações, demonstrando sua relevância na literatura sobre diabetes mellitus. A análise desses termos centrais pode revelar os principais tópicos em discussão e as tendências predominantes na pesquisa. Como nodos centrais, foi possível identificar termos relativos a Diabetes Mellitus, assim como outros tipicamente associados a esta doença.
- Comunidades: O grafo também revela algumas comunidades densamente interligadas de palavras-chave, possivelmente representando subáreas especializadas no estudo do diabetes mellitus. Essas comunidades podem indicar campos de estudo bem definidos ou nichos emergentes.

Além desses elementos, a análise também identifica características, tais como:

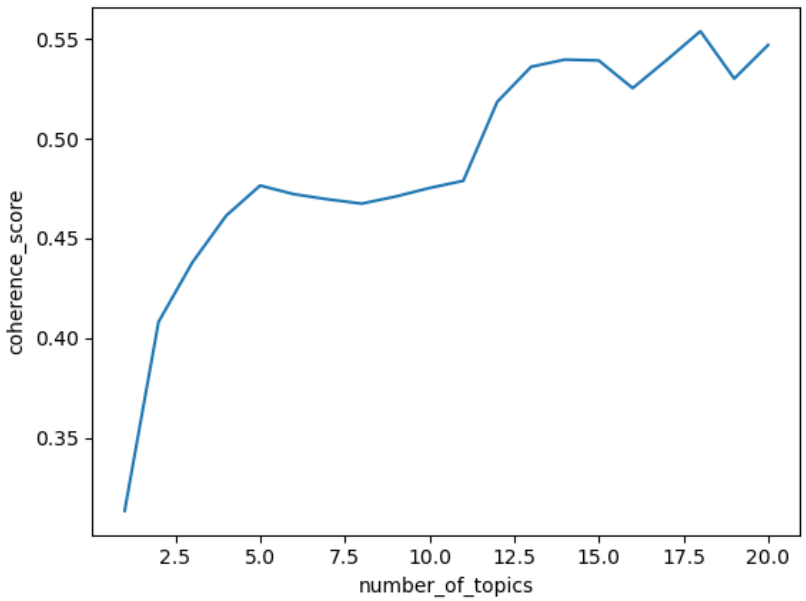
- Termos Pouco Relevantes: Duplas isoladas no grafo podem indicar termos menos relevantes para a pesquisa principal em diabetes mellitus, aparecendo em menos estudos e, consequentemente, tendo menos conexões.
- Termos Específicos e Novos: Essas duplas podem representar termos extremamente específicos ou novos na área, que estão sendo estudados em um número limitado de trabalhos. Isso pode incluir jargões emergentes ou conceitos que estão começando a ganhar notoriedade.
- Limitações da Análise: As duplas isoladas podem também indicar limitações na metodologia de coleta ou análise dos dados, resultando em uma representação incompleta das relações entre as palavras-chave ou na inclusão de termos irrelevantes.

O reconhecimento de padrões repetitivos nas duplas de termos contribui para a compreensão do campo, fornecendo insights críticos que podem orientar futuras pesquisas e aprofundar o entendimento dos temas tratados nos estudos sobre *diabetes mellitus*.

5.2.3 Topic Modelling – Latent Dirichlet Allocation

As descobertas utilizando Topic Modelling, especificamente com o uso do LDA, foram interessantes e promissoras para o meu entendimento sobre a doença diabetes mellitus. O LDA é uma técnica poderosa para identificar tópicos subjacentes em um grande conjunto de textos, mas requer que o número de tópicos seja definido previamente. Uma maneira eficaz de determinar esse número é através do cálculo da coherence, que mede a consistência e a semântica dos tópicos gerados. A análise de coherence é crucial para garantir que os tópicos identificados sejam significativos e úteis para a pesquisa. Esta abordagem é ilustrada na Figura 28, que demonstra como a coherence pode ser utilizada para otimizar a seleção do número de tópicos, resultando em uma modelagem mais precisa e informativa.

Figura 28 - Gráfico para a escolha da melhor métrica de coherence



Fonte: Elaborado pelo autor (2024)

A escolha do melhor número de tópicos em um modelo de Latent Dirichlet Allocation (LDA) é realizada através da análise de coherence. O gráfico resultante da coherence apresenta valores que indicam a consistência e a semântica dos tópicos gerados, permitindo identificar o ponto ideal onde o número de tópicos oferece a melhor representação dos dados.

No gráfico, o eixo X representa o número de tópicos, enquanto o eixo Y mostra o valor da coherence. O melhor número de tópicos é geralmente identificado

no ponto onde a coherence atinge seu valor máximo ou começa a se estabilizar, indicando que os tópicos são suficientemente coerentes e semânticos.

Para o experimento, foram utilizados apenas 7 tópicos. Esta decisão foi baseada na complexidade que surge com números muito elevados de tópicos, o que pode tornar a análise mais difícil e menos clara. Como o objetivo final desta pesquisa não é a modelagem exaustiva de todos os possíveis tópicos, optou-se por simplificar o processo. A escolha de 7 tópicos permite uma análise prática e manejável, facilitando a extração de insights significativos sobre diabetes mellitus sem sobrecarregar o modelo com uma quantidade excessiva de informações.

Para a análise visual dos resultados do modelo LDA, foi utilizada a biblioteca LDAvis do Python, que fornece uma interface interativa para a exploração e interpretação dos tópicos gerados. A Figura 29 apresenta a visualização LDAvis, onde é possível observar um mapa de distância entre tópicos e os termos mais relevantes para um tópico específico.

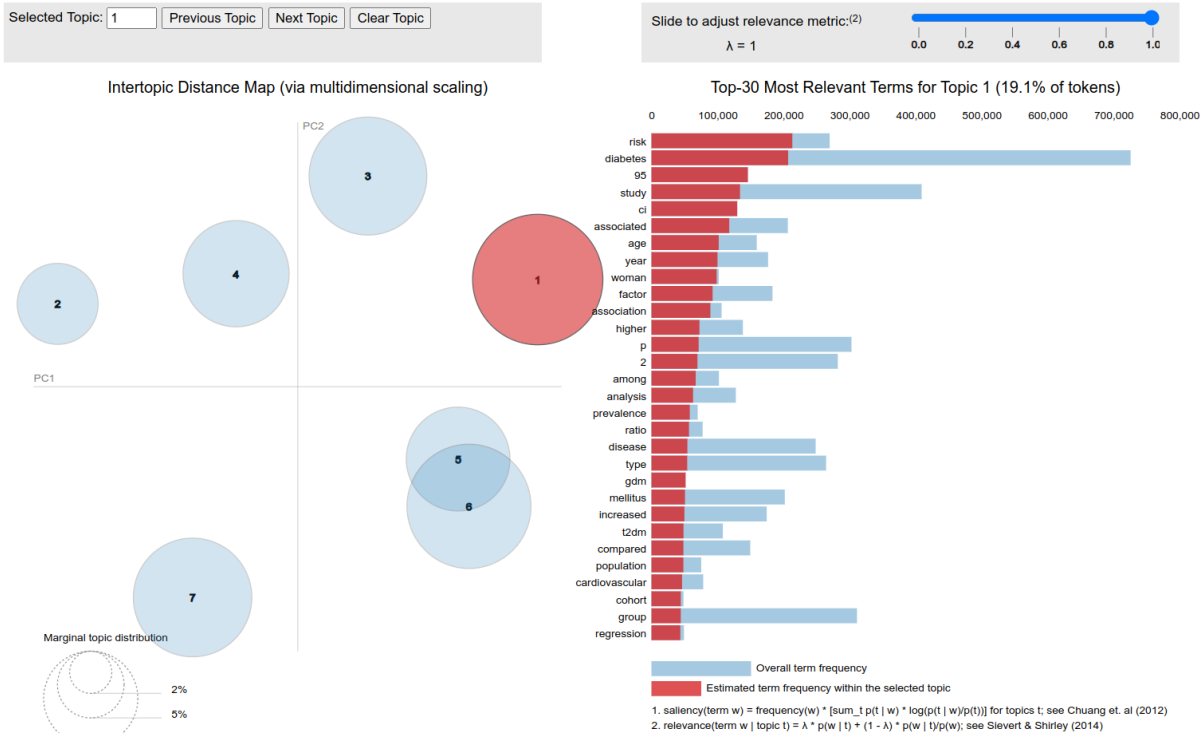
No lado esquerdo da figura, temos o “Intertopic Distance Map” que utiliza escalonamento multidimensional (MDS) para mostrar as distâncias entre os tópicos. Cada círculo representa um tópico, e a proximidade entre os círculos indica a semelhança entre os tópicos. Tópicos mais próximos compartilham mais termos em comum, enquanto tópicos mais distantes são mais distintos entre si. O tamanho dos círculos representa a prevalência dos tópicos no corpus; tópicos maiores são mais comuns.

No lado direito da figura, é apresentada uma lista dos 30 termos mais relevantes para o tópico selecionado, que neste exemplo é o Tópico 1, responsável por 19,1% dos tokens. A barra azul clara representa a frequência geral dos termos no corpus, enquanto a barra vermelha indica a frequência estimada dos termos dentro do tópico selecionado. Termos como “risk”, “diabetes”, “study”, “associated”, e “age” destacam-se como especialmente relevantes para o Tópico 1, sugerindo que este tópico está relacionado a estudos sobre os riscos e fatores associados ao diabetes.

A relevância dos termos pode ser ajustada utilizando o controle deslizante λ, permitindo uma análise mais detalhada e focada dos tópicos. Termos com barras vermelhas maiores são mais específicos para o tópico selecionado, enquanto termos com barras azuis maiores são mais gerais no corpus.

Essa visualização facilita a compreensão e interpretação dos tópicos gerados pelo LDA, permitindo identificar rapidamente os principais temas e suas relações dentro do conjunto de dados sobre diabetes mellitus.

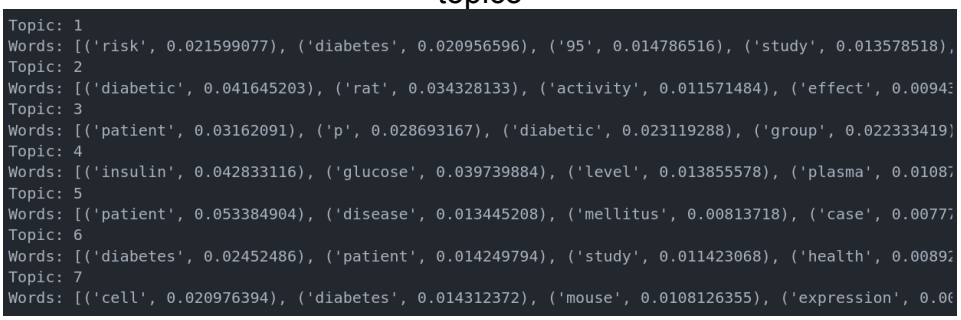
Figura 29 - Resultado do LDA visualizado utilizando o LDAvis



Fonte: Elaborado pelo autor (2024)

Os dados internos do modelo criado pelo LDA podem ser acessados como conjuntos de pares, compostos por termo e a probabilidade do termo ocorrer no tópico ao qual pertence o conjunto. Esses pares permitem uma análise detalhada dos termos mais representativos de cada tópico, fornecendo insights adicionais sobre a distribuição e a relevância dos termos no contexto dos tópicos gerados pelo modelo. A visualização da formatação, desses conjuntos de pares, utilizada nas análises posteriores pode ser observada na Figura 30.

Figura 30 - Pares de termo e a probabilidade que compõe cada tópico



Fonte: Elaborado pelo autor (2024)

A análise das palavras contidas em cada tópico resultou em possíveis assuntos ou rótulos dos tópicos extraídos. A interpretação fornecida, que pode ser subjetiva, é apresentada a seguir:

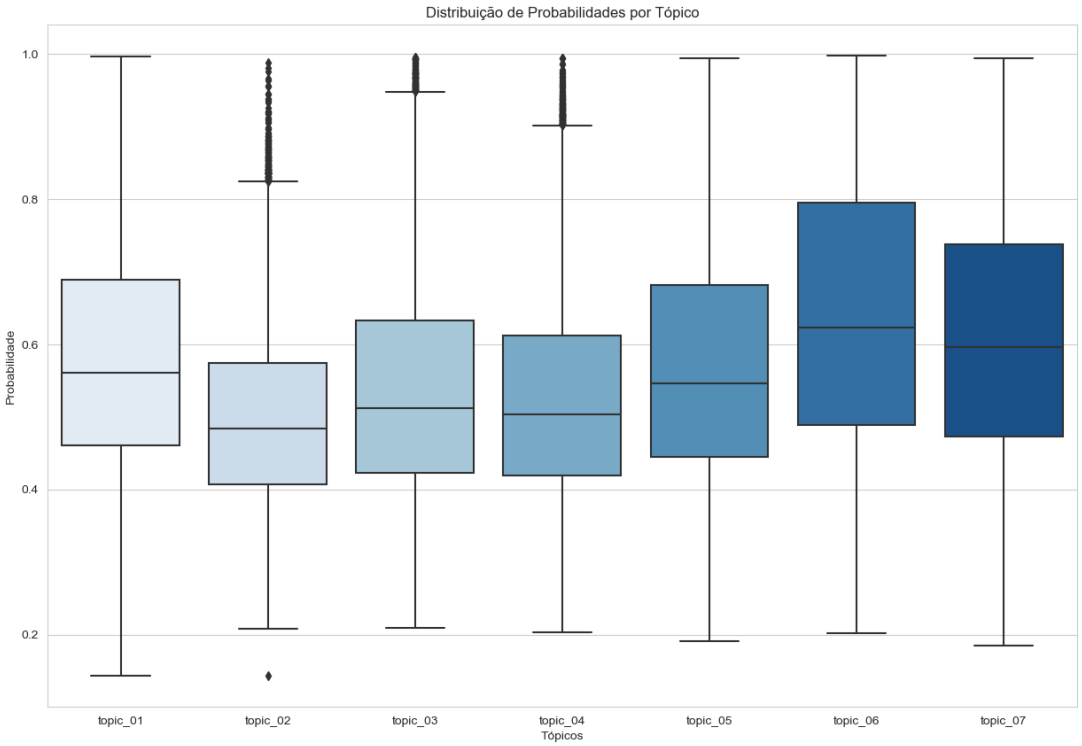
- Tópico 1: Epidemiologia e Fatores de Risco. Este tópico parece se concentrar na epidemiologia da diabetes, destacando riscos associados, prevalência e demografia (incluindo a idade e o gênero). As palavras-chave “risk”, “study”, “associated”, “age”, “factor” indicam uma análise estatística de fatores de risco e associações para diabetes. A presença de termos como “95”, “ci” (intervalo de confiança) e “p” (valor-p) sugere que este tópico apresenta resultados de estudos epidemiológicos quantitativos.
- Tópico 2: Pesquisas Experimentais em Modelos Animais. Este tópico possivelmente trata de experimentos em animais, especialmente ratos, e analisa os efeitos de tratamentos ou condições patológicas. Alguns termos como “rat”, “activity”, “effect”, “animal” e “control” indicam um foco em estudos experimentais. “Diabetic”, “wound” e “nerve” sugerem um foco nas complicações da diabetes nos modelos.
- Tópico 3: Análise Clínica de Diabetes. Aparentemente, o foco é nos resultados clínicos de pacientes com diabetes, incluindo comparações com grupos de controle e análises de biomarcadores, como “serum”, “level” e “blood”. As expressões “significantly”, “p” e “lt” (menos que) sugerem análises estatísticas dos dados coletados em estudos clínicos.
- Tópico 4: Estudo sobre Insulina e Glucose. Este tópico tem como foco principal o estudo bioquímico e farmacológico da insulina e glucose, utilizando termos como “insulin”, “glucose”, “plasma” e “concentration”. Além disso,

menciona-se “metformin”, um medicamento usado para diabetes, o que sugere um foco no tratamento e gestão metabólica.

- Tópico 5: Complicações e Gestão de Casos Clínicos. Os termos apresentados apontam para discussões sobre complicações da diabetes e gestão de casos em ambientes clínicos. Palavras como “patient”, “disease”, “clinical”, “treatment” e “complication” são predominantes, indicando um foco nos desafios de tratar pacientes com condições complexas.
- Tópico 6: A Saúde Pública e a Gestão da Diabetes. Possivelmente aqui se aborda a diabetes sob a perspectiva da saúde pública e da gestão da doença, com ênfase nos conceitos de “health”, “care”, “data” e “management”. A utilização de termos como “review”, “trial” e “system” sugere uma abordagem fundamentada em revisões sistemáticas e ensaios clínicos.
- Tópico 7: Pesquisa Básica e Molecular. Com foco na pesquisa molecular e celular, este tópico discute os mecanismos celulares e genéticos ligados à diabetes. As expressões “cell”, “expression”, “protein”, “gene” e “insulin” enfatizam uma abordagem mais fundamental da doença.

No gráfico apresentado na Figura 31, é possível notar uma média de probabilidades entre 0,4 e 0,6, o que indica que, em geral, não há um tópico que domine completamente a distribuição, apesar de haver variações significativas na dispersão e nos outliers entre os tópicos.

Figura 31 - Distribuição de probabilidade de cada tópico



Fonte: Elaborado pelo autor (2024)

Analisando cada caso, podemos chegar às seguintes conclusões:

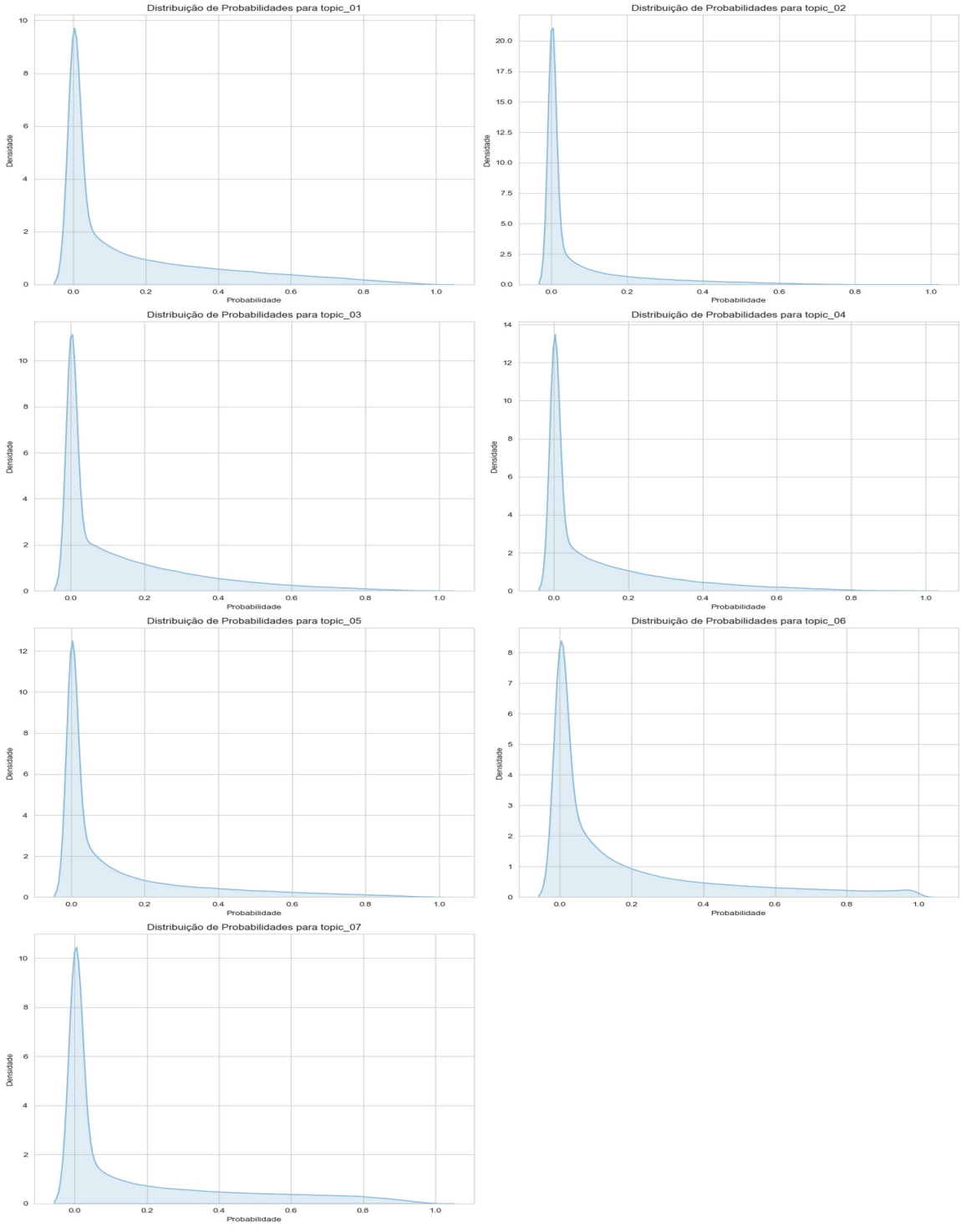
- Tópico 1 apresenta uma distribuição mais ampla e uma mediana mais baixa em comparação com os outros tópicos, sugerindo que as probabilidades associadas a este tópico são mais variáveis entre os documentos. Também há alguns outliers inferiores, indicando que alguns documentos têm uma probabilidade muito baixa de se alinhar com este tópico.
- Tópicos 2 e 3 possuem medianas semelhantes e distribuições mais estreitas do que o Tópico 1, mas com muitos outliers superiores. Isso sugere que, embora a maioria dos documentos tenha uma probabilidade moderada de se alinhar com esses tópicos, alguns documentos têm uma alta probabilidade de corresponder a eles.
- Tópicos 4 e 5 têm medianas semelhantes e quartis superiores mais elevados, o que implica uma tendência de ter uma probabilidade consistentemente mais alta em um conjunto substancial de documentos.
- Tópico 6 exibe a mediana mais alta entre todos, indicando que é o mais representativo na coleção de documentos analisados. Sua distribuição de quartis também é compacta, demonstrando menos variabilidade entre os

documentos para este tópico. Cabe ressaltar que este tópico, identificado como “A Saúde Pública e a Gestão da Diabetes”, é possivelmente o mais genérico.

- Tópico 7, embora tenha uma mediana comparável aos Tópicos 4 e 5, possui um intervalo interquartil mais largo e apresenta alguns outliers inferiores, mostrando que há documentos que se alinham muito fracamente com este tópico.

Adicionalmente, foi realizada a análise das distribuições de densidade de probabilidade para cada tópico, conforme mostrado na Figura 32. Esses gráficos mostram picos pronunciados perto de zero, indicando que a maioria dos documentos tem uma baixa probabilidade de estar associada fortemente a um único tópico. Esta tendência é consistente em todos os tópicos e pode ser atribuída ao fato de que os documentos analisados pertencem a uma área temática altamente especializada, que é a diabetes mellitus. Dado que o assunto é específico, é provável que haja uma sobreposição substancial de vocabulário e conceitos entre os documentos, levando a uma distribuição de probabilidade que favorece valores mais baixos, uma vez que nenhum tópico é exclusivamente dominante. Além disso, a especialização do tema pode significar que os documentos abordem a diabetes de múltiplas perspectivas inter-relacionadas, resultando em atribuições de tópico mais difusas e menos polarizadas.

Figura 32 - Análise das distribuições de densidade de probabilidade



Fonte: Elaborado pelo autor (2024)

A análise com base na metodologia LDA revelou que, apesar de sete tópicos terem sido identificados dentro do corpus especializado em diabetes mellitus, a distribuição das probabilidades sugere que a complexidade do assunto poderia ser melhor captada por um número maior de tópicos. A ampla variação observada no Tópico 1 indica uma heterogeneidade nas discussões sobre epidemiologia e fatores

de risco, sugerindo nuances que um número maior de tópicos poderia potencialmente desvendar. A predominância de probabilidades baixas entre os documentos indica a sobreposição de conceitos e vocabulário, o que aponta para a interconexão das áreas temáticas. Isso reforça a ideia de que divisões mais amplas poderiam facilitar a aquisição de conhecimento.

Os resultados preliminares sugerem a necessidade de mais pesquisas com diferentes técnicas de modelagem de tópicos para uma construção mais robusta e detalhada de grafos de conhecimento. A abordagem inicial com sete tópicos proporcionou insights valiosos, mas há um potencial significativo para uma análise mais granular que possa capturar a diversidade e complexidade dos estudos sobre diabetes mellitus. A continuidade desta pesquisa, explorando outras técnicas e ajustando o número de tópicos, será apresentada nas próximas seções, visando aprimorar a compreensão e a visualização das inter-relações temáticas dentro da literatura científica analisada.

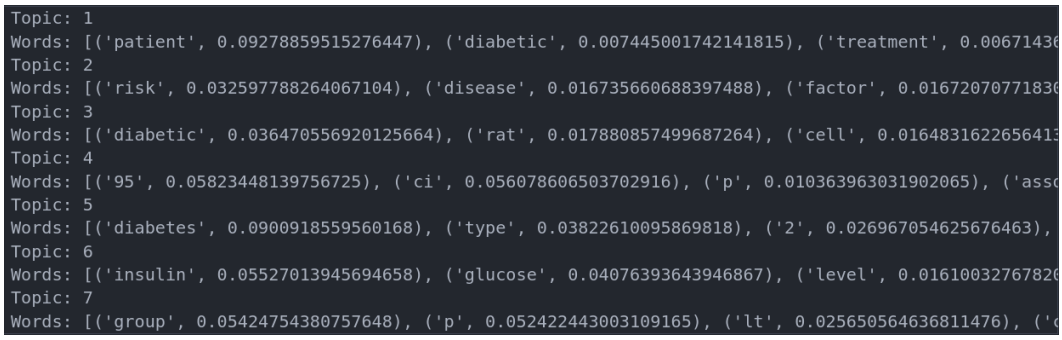
5.2.4 Topic Modelling – Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) é uma técnica de redução de dimensionalidade e aprendizado de representações que se mostra bastante útil na análise de tópicos, especialmente no processamento de coleções de documentos. Diferente do LDA, que modela documentos como misturas de tópicos e tópicos como misturas de palavras, o NMF decompõe uma matriz de alta dimensão em duas matrizes de baixa dimensão, com a restrição de que todas as matrizes tenham apenas valores não negativos. Isso permite que o NMF atribua pesos específicos a palavras em cada tópico, facilitando a interpretação.

Enquanto o LDA é um modelo probabilístico que usa uma abordagem baseada em inferência para estimar a distribuição de tópicos, o NMF é um algoritmo algébrico que utiliza técnicas de otimização para encontrar a melhor representação de fatores, minimizando o erro de reconstrução. O NMF tende a resultar em tópicos mais esparsos, o que pode ser mais interpretável, pois cada documento é representado por um número menor de tópicos ativos. Além disso, o NMF pode ser mais rápido em termos de tempo de computação comparado ao LDA, mas é mais sensível à escolha do número de tópicos, tornando desafiador determinar o número ótimo a priori.

Embora o LDA ofereça uma interpretação probabilística mais rica, o NMF proporciona uma utilidade prática em termos de simplicidade e velocidade, sendo valioso em grandes conjuntos de dados ou quando há restrições de tempo. Para esta abordagem, não foram geradas visualizações específicas, mas na Figura 33 pode-se observar um segmento da lista de termos e suas probabilidades para cada tópico, demonstrando a efetividade do NMF na identificação e interpretação dos tópicos.

Figura 33 - Pares de termo e a probabilidade que compõe cada tópico NMF



Fonte: Elaborado pelo autor (2024)

Para manter a consistência com a abordagem anterior, optou-se por aplicar a técnica de Non-negative Matrix Factorization (NMF), configurando-a para identificar sete tópicos distintos. A análise subsequente dos resultados obtidos é apresentada a seguir:

- Tópico 1: Concentra-se predominantemente no manejo do paciente diabético, com destaque para tratamentos, estudos clínicos e o curso da doença. A forte presença do termo “patient” sugere um enfoque em estudos centrados no paciente, enquanto a variedade de termos relacionados a tratamentos e desfechos clínicos indica uma discussão rica em torno da gestão clínica do diabetes.
- Tópico 2: Aborda os fatores de risco e doenças associadas ao diabetes, como doenças cardiovasculares e hipertensão. O termo “risk” aparece com uma forte probabilidade, destacando-se como um tópico voltado para estudos epidemiológicos e a associação entre diabetes e outras comorbidades.
- Tópico 3: Engloba pesquisa experimental, com foco em estudos com modelos animais, evidenciado pelos termos “rat” e “mouse”. Além disso, termos como

"cell" e "expression" apontam para um tópico que trata da pesquisa celular e molecular na diabetes.

- Tópico 4: Caracteriza-se pela presença de termos estatísticos como "95", "ci" e "p", indicando que esse tópico possivelmente compreende estudos que relatam análises estatísticas detalhadas, como intervalos de confiança e valores-p, típicos em pesquisas que quantificam associações ou efeitos.
- Tópico 5: Está fortemente associado ao diabetes mellitus tipos 1 e 2, sugerindo uma abordagem mais ampla da doença, abarcando aspectos de prevalência, gestão e intervenção de saúde pública.
- Tópico 6: Relaciona-se a aspectos bioquímicos do diabetes, como o papel da insulina e da glicose no metabolismo e tratamento. A alta probabilidade dos termos "insulin" e "glucose" destaca o enfoque em processos metabólicos e respostas ao tratamento.
- Tópico 7: Apresenta uma concentração de termos estatísticos e comparativos, como "p", "lt" e "compared", sugerindo um tópico que envolve comparação de grupos em estudos, avaliação de efeitos significativos e análise de níveis de biomarcadores.

Ao identificar certas semelhanças nos conjuntos de palavras, foi realizada uma análise cruzando os dois conjuntos (LDA e NMF) para tentar identificar equivalências. O resultado demonstrou que existem equivalências relativamente claras, porém as diferenças nas probabilidades das palavras mostraram claramente o enfoque distinto entre as duas técnicas. A análise comparativa sugere que, enquanto o LDA fornece uma distribuição mais uniforme das palavras-chave dentro de tópicos, o NMF tende a se concentrar mais intensamente em menos palavras, potencialmente oferecendo um foco mais nítido, porém menos diversificado. Ambas as técnicas revelam insights complementares sobre a estrutura temática dos documentos analisados.

5.2.5 Topic Modelling – BERTopic

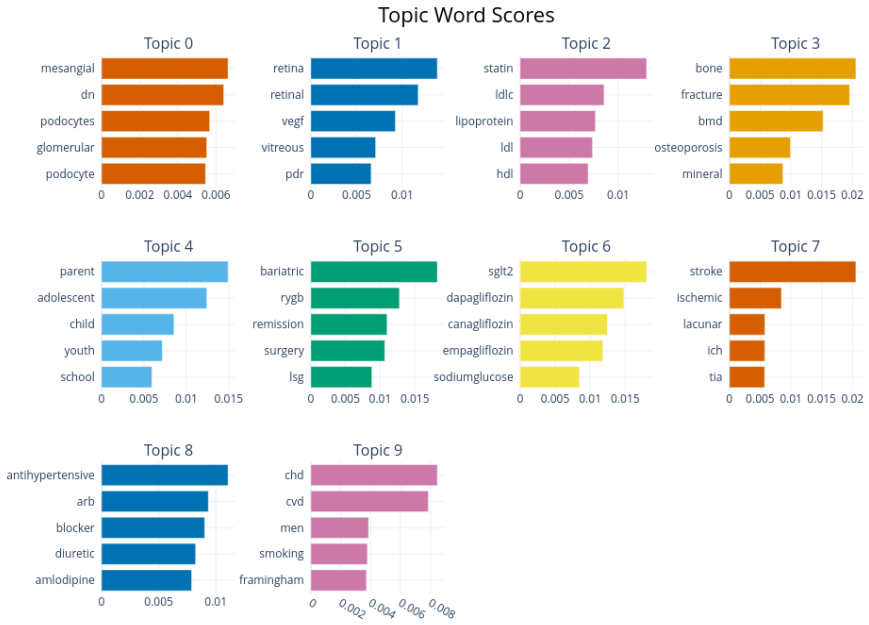
Dois experimentos foram conduzidos utilizando a técnica BERTopic, cujos resultados serão expostos a seguir, organizados em dois segmentos distintos. O propósito aqui não é mergulhar profundamente na análise da modelagem de tópicos,

já que o foco central deste trabalho não reside na exploração exaustiva de Topic Modelling, mas sim em demonstrar como essa técnica pode contribuir para a construção de KGs e enriquecer o entendimento sobre o tema diabetes mellitus, facilitando a execução do projeto proposto.

No primeiro experimento com o BERTopic, optou-se por não definir um número específico de tópicos, resultando na identificação automática de 2147 tópicos. Embora essa quantidade seja excessiva para as necessidades da análise, a utilização dos relatórios gerados forneceu insights valiosos. Iniciarei examinando os conjuntos de palavras que representam os tópicos identificados, no relatório Topic Word Scores, como demonstrado na Figura 34.

Um ponto de atenção é o fato do BERTopic criar um tópico de índice -1. Este identificador possui uma função especial: ele é um tópico “padrão” para os documentos cujas análises não conseguem determinar, dentro de uma probabilidade aceitável, a qual tópico pertencem. Notavelmente, para este experimento, o número foi alto, reforçando a hipótese de alta semelhança entre os documentos ao serem de uma área tão específica.

Figura 34 - Escore de palavras de tópico do BERTopic



Fonte: Elaborado pelo autor (2024)

Este relatório exibe, por uma questão de praticidade, apenas os 10 tópicos mais proeminentes, pois abordar integralmente os 2147 tópicos identificados pelo modelo seria inviável em um formato de apresentação conciso. Contudo, cada tópico

pode ser investigado individualmente para permitir uma análise mais detalhada e direcionada. Detalhes e insights adicionais sobre cada tópico específico apresentados no relatório serão discutidos a seguir:

- Tópico 0: Palavras como “mesangial”, “dn” (possivelmente nefropatia diabética), “podocytes”, e “glomerular” indicam um tópico relacionado à nefrologia, focado em aspectos da doença renal.
- Tópico 1: Com termos como “retina”, “retinal”, “vegf” e “vitreous”, este tópico aparentemente se concentra em condições oftalmológicas, particularmente aquelas afetando a retina.
- Tópico 2: Palavras como “statin”, “ldlc”, “lipoprotein” e “hdl” estão ligadas ao tópico de saúde cardiovascular, especialmente no contexto do gerenciamento do colesterol.
- Tópico 3: As palavras “bone”, “fracture”, “bmd” (densidade mineral óssea) e “osteoporosis” apontam para um foco em doenças ósseas e ortopedia.
- Tópico 4: Termos como “parent”, “adolescent”, “child”, e “school” indicam discussões sobre a saúde infantil e juvenil e possivelmente a psicologia do desenvolvimento.
- Tópico 5: “bariatric”, “rygb” (bypass gástrico em Y de Roux), “remission”, e “surgery” sugerem que este tópico trata de cirurgia bariátrica e suas implicações na saúde.
- Tópico 6: Termos como “sglt2”, “dapagliflozin”, “canagliflozin”, e “empagliflozin” estão todos relacionados a uma classe de medicamentos usados no tratamento de diabetes tipo 2, os inibidores de SGLT2.
- Tópico 7: Este tópico inclui palavras associadas a condições cerebrovasculares: “stroke”, “ischemic”, “lacunar”, “ich” (hemorragia intracerebral) e “tia” (ataque isquêmico transitório).
- Tópico 8: Termos como “antihypertensive”, “arb” (bloqueador do receptor de angiotensina), “blocker”, e “diuretic” indicam um tópico dedicado a medicamentos anti-hipertensivos.

A análise apresentada revela uma discrepância considerável em comparação aos resultados anteriores, emergindo tópicos com um grau de especificidade notavelmente maior. Contudo, é importante ressaltar que, nos casos do LDA e do NMF, o número de tópicos foi intencionalmente fixado em sete. Assim,

os 10 tópicos aqui demonstrados pelo relatório do BERTopic representam apenas uma fração dos tópicos detectados, proporcionando uma visão parcial do espectro completo identificado pelo modelo.

Como ilustração de uma análise focada, a Figura 35 destaca as informações pertinentes ao tópico número 423. Este tópico parece abordar questões relacionadas à saúde e nutrição na infância e adolescência e pode conter estudos que diferenciam resultados ou comportamentos por gênero. Isso é evidenciado pela distinção clara entre as palavras “child”, “girl” e “boy”, sugerindo que o tópico engloba uma discussão diferenciada em termos de sexo na população infanto-juvenil.

Figura 35 - Probabilidade dos termos para o tópico 423

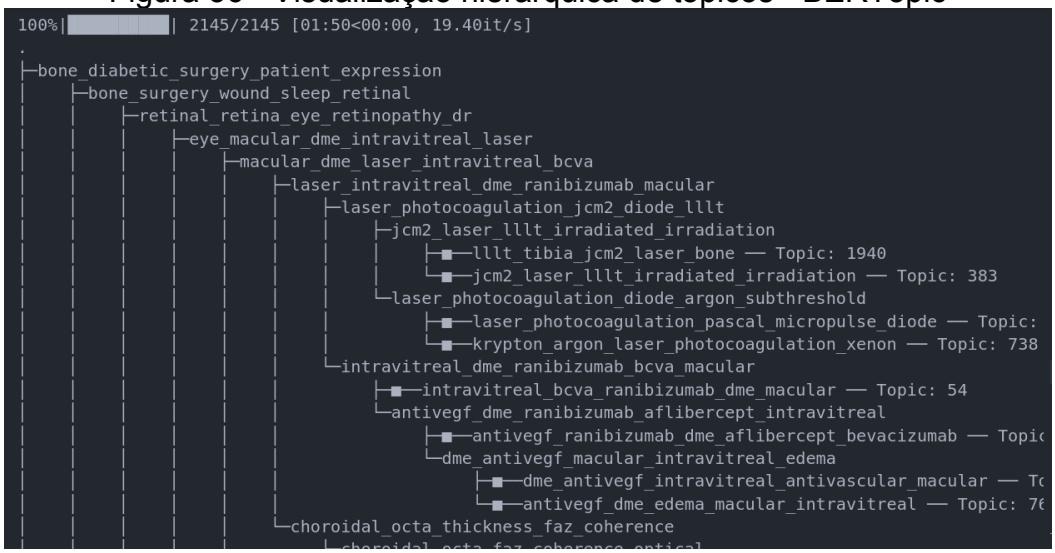
```
child: 0.012628419676952572
girl: 0.009008236012598243
adolescent: 0.008542039487676185
boy: 0.008378671307550975
ifg: 0.008240829040524655
childhood: 0.007773254596169204
homair: 0.007579433671078604
igt: 0.0069930477548569596
obese: 0.006744521538853459
ir: 0.006002675038562914
```

Fonte: Elaborado pelo autor (2024)

Além do tópico 423, uma série de outros tópicos foi analisada, revelando uma ampla variedade de áreas de pesquisa. Entre os tópicos destacados, encontram-se a saúde animal, com ênfase em estudos veterinários, a importância dos exercícios físicos e as mudanças hormonais. Esses exemplos ilustram o vasto espectro temático que o BERTopic é capaz de capturar e classificar, fornecendo insights valiosos em diversos domínios do conhecimento.

A natureza interconectada deste campo de estudo sugere que os tópicos podem ser naturalmente divididos em uma série de subtópicos mais detalhados. Esse fenômeno está em consonância com a visão do pesquisador de que temas abrangentes se desdobram em nichos mais focados de discussão. Na Figura 36, um mapa hierárquico reforça essa perspectiva, oferecendo uma representação visual similar à de uma clusterização hierárquica, e ilustra claramente a relação entre tópicos e seus subtópicos correspondentes.

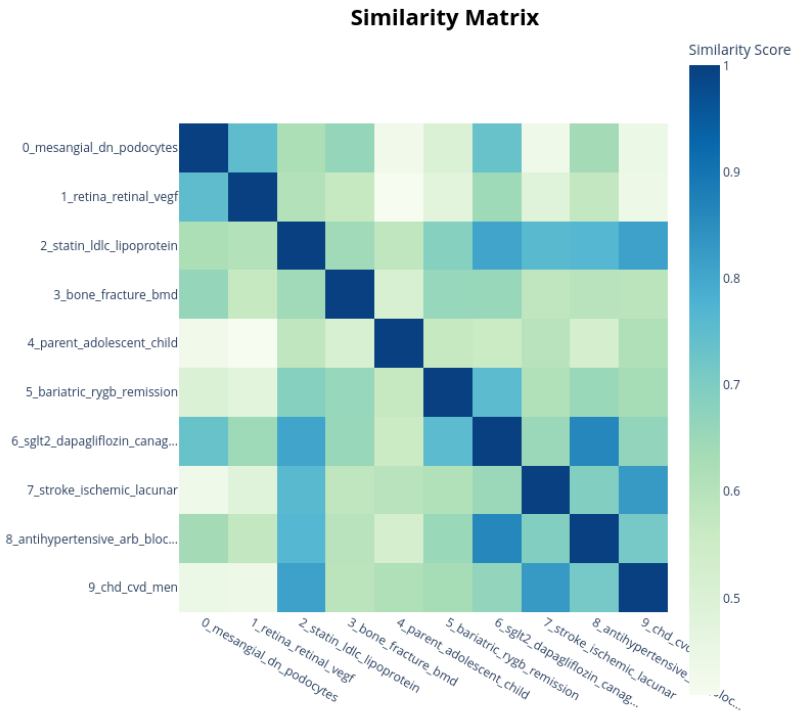
Figura 36 - Visualização hierárquica de tópicos - BERTopic



Fonte: Elaborado pelo autor (2024)

O último relatório de análise exposto neste trabalho sobre BERTopic, para um número de tópicos automático, é chamado de Similarity Matrix, apresentado como um mapa de calor. Este relatório aponta similaridades entre os tópicos, funcionando como uma forma alternativa de visualizar o mapa de distâncias. No entanto, o relatório não se mostrou prático para um grande número de tópicos, pois não conseguiu plotar os 423 tópicos descobertos, gerando uma matriz 33 x 33 preenchida com tópicos selecionados de maneira aparentemente aleatória ou baseados em importância, embora os critérios de seleção não tenham sido claramente identificados. Devido à sua baixa utilidade visual para um grande número de tópicos, essa plotagem não será apresentada. Em vez disso, uma matriz de similaridade para apenas 10 tópicos será exibida na Figura 37, proporcionando uma visualização mais clara e compreensível.

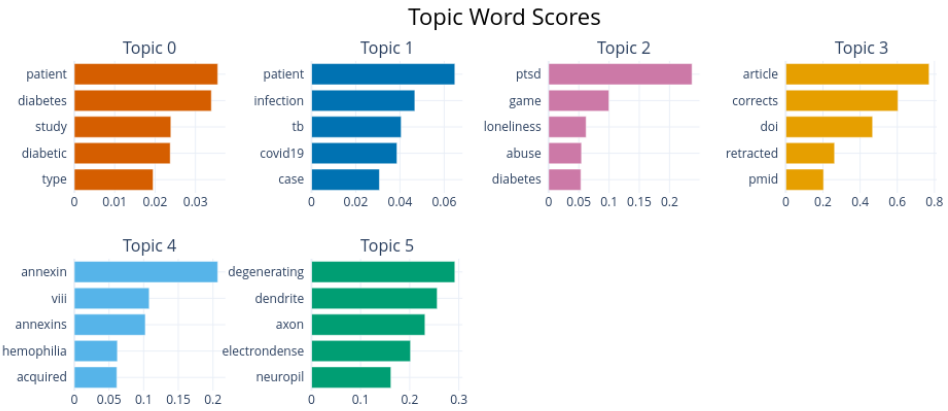
Figura 37 - Matriz de similaridade intertópicos - BERTopic



Fonte: Elaborado pelo autor (2024)

Na segunda abordagem realizada com o BERTopic, o número de tópicos foi estabelecido preliminarmente em sete (k = 7). Contudo, é importante notar que o número definido inicialmente não corresponde obrigatoriamente à quantidade de tópicos que o modelo identificará. O BERTopic emprega uma abordagem que ajusta o número de tópicos com base nos dados, tratando o valor estipulado mais como uma orientação para o número máximo de tópicos a serem gerados. Neste experimento específico, o BERTopic determinou que seis tópicos era o mais adequado para representar o conjunto de dados. A Figura 38 ilustra os termos mais significativos que definem cada um dos seis tópicos descobertos.

Figura 38 - Escore de palavras por tópico - BERTopic (k = 7)



Fonte: Elaborado pelo autor (2024)

Ao estabelecer um limite para a quantidade de tópicos, o BERTopic identificou conjuntos de termos que diferem significativamente daqueles revelados por métodos anteriores, como o LDA. Essa distinção sugere que o BERTopic possui uma capacidade superior de discernir tópicos com maior especificidade. Em seguida, apresentarei uma análise detalhada dos tópicos identificados pelo BERTopic, destacando suas principais características e relevância no contexto da pesquisa.

- Tópico 0: Este tópico está focado em termos como “patient”, “diabetes”, “study”, “diabetic”, e “type”, sugerindo que provavelmente aborda estudos clínicos ou pesquisas relacionadas ao diabetes, possivelmente diferenciando entre os tipos de diabetes.
- Tópico 1: Com termos como “patient”, “infection”, “tb” (tuberculose), “covid19”, e “case”, este tópico parece se concentrar em doenças infecciosas, particularmente TB e COVID-19, e suas incidências em pacientes.
- Tópico 2: Inclui termos como “ptsd” (transtorno de estresse pós-traumático), “game”, “loneliness”, “abuse”, e “diabetes”, sugerindo um tópico que liga o impacto psicológico e social de condições crônicas ou de abuso com o diabetes.
- Tópico 3: Este tópico contém termos como “article”, “corrects”, “doi”, “retracted”, e “pmid”, implicando um foco relacionado à publicação científica e à integridade da pesquisa, onde artigos são corrigidos ou retratados.

- Tópico 4: Inclui termos como “annexin”, “viii”, “annexins”, “hemophilia”, e “acquired”, que estão associados a coagulopatias, especificamente à hemofilia e às proteínas relacionadas à coagulação.
- Tópico 5: Com palavras como “degenerating”, “dendrite”, “axon”, “electrondense”, e “neuropil”, este tópico aponta para a neurologia, com um foco potencial em doenças degenerativas e anatomia neuronal.

A diferença entre os termos pode ser atribuída às diferenças fundamentais nos métodos de análise de tópicos. O BERTopic, ao contrário do LDA, usa embeddings de palavras com a arquitetura Transformer, capturando de forma eficaz o contexto e as nuances semânticas dos dados. Esta técnica adapta de forma flexível o número de tópicos identificados, considerando o valor estipulado pelo usuário como uma diretriz ou um limite. Dessa forma, o BERTopic é capaz de desvendar tópicos com uma maior profundidade, fornecendo uma visão mais aprofundada e contextualizada dos temas presentes nos dados.

5.2.6 EDA – Avaliação de Resultados

A análise inicial de dados (EDA) revelou aspectos cruciais para estudos bibliométricos e cienciométricos, destacando a importância do inglês como idioma predominante na disseminação da produção científica em biomedicina. Essas descobertas corroboram a escolha estratégica do inglês para a seleção de documentos, alinhando-se ao idioma predominante da comunidade científica internacional e assegurando a inclusão de um amplo espectro de pesquisas globais. A análise também evidenciou um aumento consistente na produção científica na área, refletindo o crescimento dinâmico e a diversificação do conhecimento biomédico. Este panorama indica a relevância dos achados para compreender tendências editoriais e fortalecer colaborações internacionais, destacando a contribuição significativa de diversas regiões na pesquisa global, sem barreiras linguísticas.

A execução da técnica de Topic Modelling revelou-se uma ferramenta valiosa para o aprofundamento do conhecimento sobre a diversidade de contextos e temas relacionados à diabetes mellitus, contribuindo significativamente para a compreensão da abrangência do Knowledge Graph (KG) a ser construído. Esta

técnica oferece uma visão ampla sobre os termos e conceitos predominantes no domínio em questão, além de auxiliar na desambiguação de termos, um fator crucial na geração de um KG preciso e rico em informações.

Além disso, o BERTopic, o qual faz uso da arquitetura Transformer, mostrou-se uma abordagem promissora no campo de Topic Modelling. A capacidade desta técnica de compreender e processar o contexto e a semântica dos textos permite identificar tópicos com mais clareza e detalhes, o que é crucial para a criação de um KG bem estruturado e informativo. Com o BERTopic, é possível extrair tópicos extremamente relevantes que refletem as tendências atuais e emergentes na pesquisa sobre diabetes mellitus, aumentando a qualidade e a utilidade do KG desenvolvido.

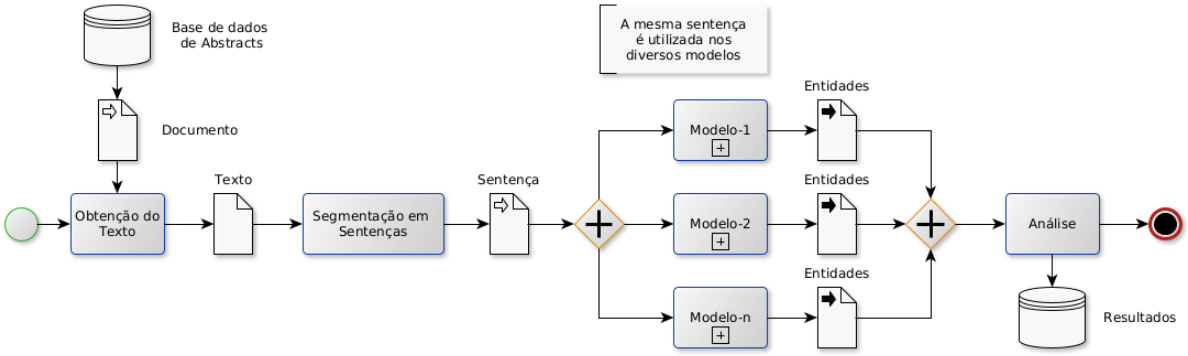
5.3 AVALIAÇÃO DA ETAPA 3: REPRESENTAÇÃO DE CONHECIMENTO

5.3.1 Named Entity Extraction

Para esta pesquisa, foram selecionados dois pacotes Python amplamente reconhecidos no campo da pesquisa. O primeiro é o scispaCy, uma extensão do spaCy, destacando-se pelo seu enfoque em textos científicos e utilizando modelos que aparentemente empregam arquitetura de CNN. O segundo é o pacote Transformers da Hugging Face, que oferece uma vasta gama de modelos baseados na inovadora arquitetura de Transformers.

A pesquisa utilizou um conjunto de dados uniforme, composto por resumos científicos que foram subdivididos em sentenças. Nesse contexto, aplicaram-se os modelos de NER para extrair entidades, permitindo uma análise comparativa das entidades identificadas por cada modelo. É importante destacar que esta comparação tem limitações, considerando que o conjunto de documentos utilizado não é rotulado. O processo é detalhado na Figura 39, facilitando a visualização do fluxo de trabalho empregado.

Figura 39 - Fluxo da extração das entidades nomeadas



Fonte: Elaborado pelo autor (2024)

No Quadro 5, detalham-se os modelos empregados na fase de NER desta investigação. Este quadro inclui informações cruciais que permitem distinguir os modelos conforme suas características distintivas, abrangendo aspectos como a arquitetura subjacente, as capacidades específicas de cada um e as ferramentas ou bibliotecas Python em que foram implementados. A apresentação destes detalhes visa fornecer uma base sólida para a compreensão das variações de desempenho e da aplicabilidade no contexto da presente análise.

Quadro 5 - Modelos de NER da área de Biomedicina utilizados

Indice	Nome	Arquitetura base	Meio de uso	Número de entidades	Tipo de entidades
01	en_ner_bionlp13cg_md	CNN	ScispaCy	16	Especifica
02	en_core_sci_sm	CNN	scispaCy	1	Genérica
03	en_core_sci_scibert	Transformer	scispaCy	1	Genérica
04	en_core_sci_lg	CNN	scispaCy	1	Genérica
05	en_ner_bc5cdr_md	Bi-LSTM-CRF	scispaCy	2	Especifica
06	biomedical-ner-all	BERT	HugginFace	83	Especifica

Fonte: Elaborado pelo autor (2024)

Conforme apresentado, os modelos de NER demonstram variação na quantidade de categorias de entidades nomeadas que conseguem reconhecer. Isso não implica necessariamente uma menor capacidade de detecção de entidades, mas sim que alguns modelos optam por um rótulo unificador para todas as entidades, classificando-as como “Genérica”. Em contraste, os modelos com o tipo

“Específica” são capazes de atribuir diversos rótulos distintos, refletindo uma gama mais ampla de categorias reconhecíveis para cada entidade identificada. É fundamental frisar que a natureza e o escopo das entidades identificadas estão diretamente relacionados ao treinamento específico pelo qual cada modelo passou.

Embora todos os modelos utilizados se concentrem em corpora da Biomedicina, eles diferem em suas especializações: alguns são direcionados para a identificação de doenças, enquanto outros se especializam em substâncias químicas relacionadas a medicamentos, alimentação ou procedimentos laboratoriais, entre outras categorias específicas. Nesse contexto, o modelo “biomedical-ner-all” se destaca por suas vantagens notáveis. Este modelo é treinado para abranger uma vasta gama de categorias dentro do domínio biomédico, o que lhe confere uma capacidade excepcional de reconhecer uma diversidade maior de entidades, tornando-o particularmente útil em análises que demandam um alto grau de detalhamento e abrangência.

Nas seções subsequentes, detalharei individualmente cada modelo utilizado nesta pesquisa, descrevendo o experimento específico aplicado e os resultados alcançados. Para ilustrar as diferenças de desempenho e facilitar a comparação direta entre os modelos, apresentarei os resultados obtidos a partir da análise da mesma sentença em determinados casos. No Quadro X, os resultados são apresentados com o “Id” codificado da seguinte forma: [A]bstract + índice + [S]entence + índice, e o “Texto” referente à sentença analisada. Esta metodologia permite uma apreciação mais clara das capacidades únicas de cada modelo dentro do escopo do estudo.

É importante destacar que não foram identificados modelos de NER específicos para diabetes mellitus. Por essa razão, optou-se pela utilização de modelos NER amplamente reconhecidos na área da Biomedicina em geral, a fim de assegurar a qualidade e a aplicabilidade dos resultados obtidos. Através dessa abordagem, é possível garantir que os modelos empregados são robustos e possuem um alto grau de precisão, fornecendo uma base sólida para as análises e interpretações subsequentes.

Quadro 6 - Sentença a serem utilizadas nas análises

Id	Texto
A0S6	extreme dyslipidemia (serum cholesterol 1311 mgdl and triglycerides 6356 mgdl) and diabetes mellitus (fasting plasma glucose 325 mgdl and hba1 c 12.1%) were first diagnosed.
A0S7	the serum lipid profiles and glucose levels were dramatically decreased within a month after treatment with subcutaneous insulin injections and oral hypolipidemic agents; notwithstanding, his vision was not significantly improved, even after treatment with intravitreal anti-vegf injection, intravitreal steroid injection and panretinal photocoagulation.

Fonte: Elaborado pelo autor (2024)

5.3.2 Modelo - en_ner_bionlp13cg_md

No contexto de NER para textos biomédicos, o modelo “en_ner_bionlp13cg_md” da scispaCy representa um recurso robusto e eficiente. Treinado especificamente com o conjunto de dados BioNLP13CG, este modelo médio possui a especialização necessária para discernir e categorizar entidades pertinentes à genética clínica e ao câncer, uma exigência primordial quando se lida com literatura científica biomédica complexa. A utilidade desse modelo transcende a simples identificação de termos, proporcionando uma identificação precisa e diversificada de conceitos biomédicos, fundamentais para a análise detalhada de textos.

A escolha desse modelo equilibra a necessidade de detalhamento na extração de entidades e a eficiência computacional, sendo ideal para o processamento ágil de grandes volumes de dados sem prejuízo à qualidade do reconhecimento. Dessa forma, ele apoia diretamente a exploração de informações científicas em publicações biomédicas, facilitando a extração de dados valiosos e aprofundados. A presumida arquitetura CNN (ou potencialmente BiLSTM) contribui para a performance satisfatória em tarefas de NER, embora as informações exatas sobre a arquitetura sejam escassas e oriundas de fontes não oficiais.

Adicionalmente, as 16 categorias de entidades biomédicas que o modelo pode identificar e classificar ampliam as capacidades de extração de informação específica, essas categorias podem ser verificadas no Quadro X. Isso permite um mapeamento detalhado de conceitos como doenças, procedimentos e agentes químicos, essenciais para a estruturação de bancos de dados biomédicos e para a

concepção de análises textuais criteriosas. Por fim, ressalta-se a importância de reconhecer as limitações de um modelo treinado em domínios específicos, como a potencial inadequação em áreas fora da biomedicina ou na identificação de neologismos pós-treinamento.

O modelo é composto por 4.087.446 tokens e 50.000 vetores distintos, o que contribui para sua robustez e eficácia na tarefa de reconhecimento de entidades nomeadas.

Quadro 7 - Categorias de entidades reconhecidas pelo modelo en_ner_bionlp13cg_md

Indice	Rótulo
01	AMINO_ACID
02	ANATOMICAL_SYSTEM
03	CANCER
04	CELL
05	CELLULAR_COMPONENT
06	DEVELOPING_ANATOMICAL_STRUCTURE
07	GENE_OR_GENE_PRODUCT
08	IMMATERIAL_ANATOMICAL_ENTITY
09	MULTI_TISSUE_STRUCTURE
10	ORGAN
11	ORGANISM
12	ORGANISM_SUBDIVISION
13	ORGANISM_SUBSTANCE
14	PATHOLOGICAL_FORMATION
15	SIMPLE_CHEMICAL
16	TISSUE

Fonte: Elaborado pelo autor (2024)

Foram conduzidos dois experimentos com variações no volume de dados: inicialmente, utilizei uma amostra aleatória de 1.000 registros para validar a pipeline, ajustar parâmetros e formatar os dados e visualizações de forma apropriada.

Posteriormente, a análise foi estendida para o corpus completo, contendo 361.688 registros. Essa abordagem foi aplicada a todos os modelos utilizados na pesquisa, permitindo uma comparação consistente entre eles. Os resultados podem ser observados na Tabela X.

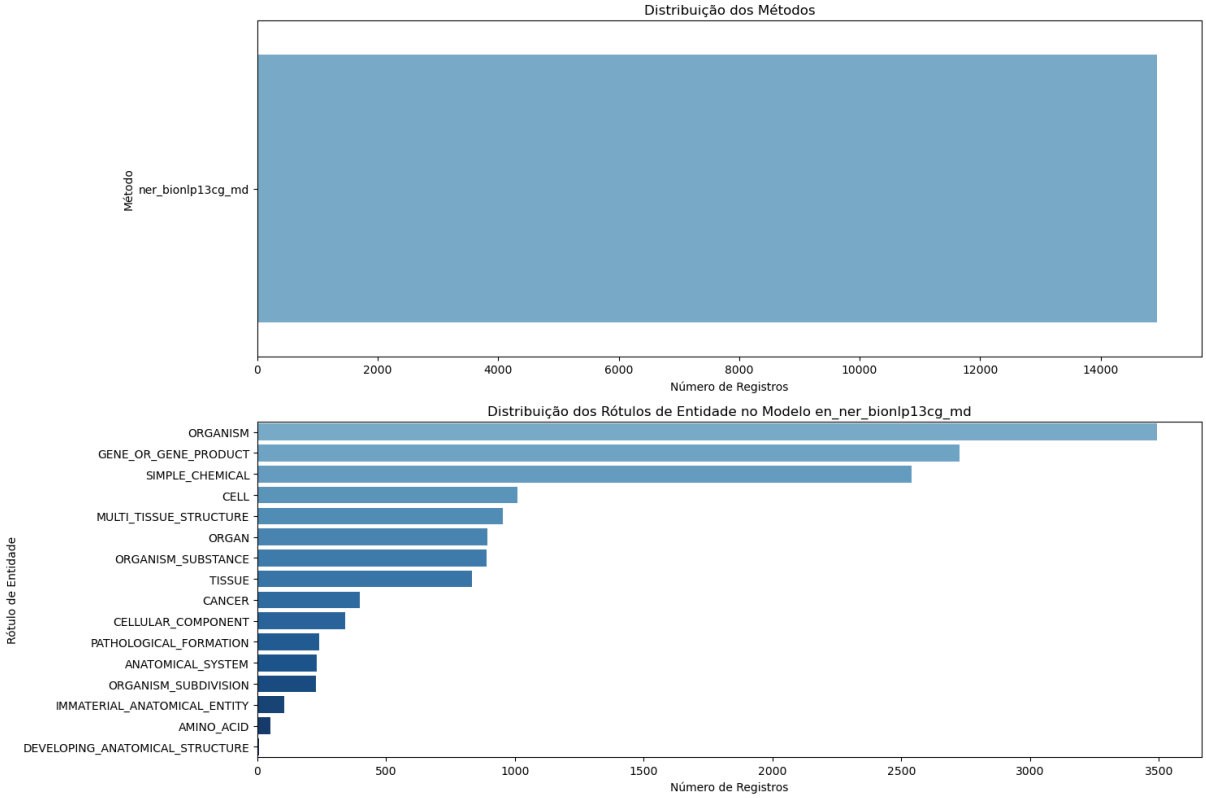
Tabela 1 - Quantidade de entidades reconhecidas pelo modelo en_ner_bi-onlp13cg_md

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	ORGANISM	3492	1304959
02	GENE_OR_GENE_PRODUCT	2726	1008113
03	SIMPLE_CHEMICAL	2540	1009669
04	CELL	1010	340947
05	MULTI_TISSUE_STRUCTURE	954	335288
06	ORGAN	894	309235
07	ORGANISM_SUBSTANCE	891	349812
08	TISSUE	832	274371
09	CANCER	396	153042
10	CELLULAR_COMPONENT	341	137821
11	PATHOLOGICAL_FORMATION	241	85026
12	ANATOMICAL_SYSTEM	231	88242
13	ORGANISM_SUBDIVISION	226	98157
14	IMMATERIAL_ANATOMICAL_ENTITY	103	43417
15	AMINO_ACID	50	20272
16	DEVELOPING_ANATOMICAL_STRUCTURE	6	1595

Fonte: Elaborado pelo autor (2024)

A representação gráfica dos dados para a amostra de 1.000 registros pode ser observada na Figura 40.

Figura 40 - Distribuição de entidades reconhecidas pelo modelo en_ner_bionlp13cg_md

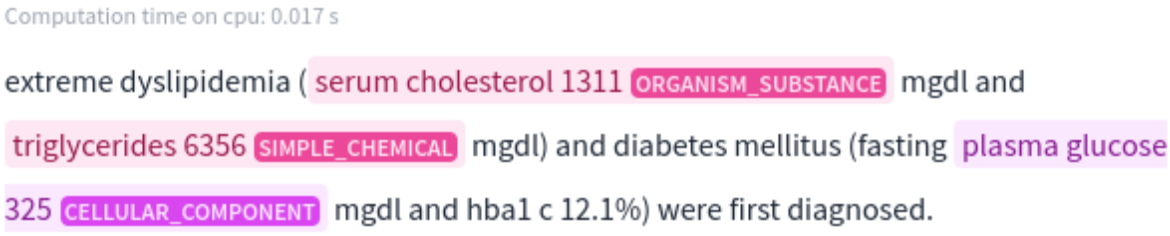


Fonte: Elaborado pelo autor (2024)

A análise dos dados destaca uma clara diferenciação entre as categorias de entidades, com “ORGANISM”, “GENE_OR_GENE_PRODUCT” e “SIMPLE_CHEMICAL” predominando significativamente. Esses termos são centrais em várias subáreas da Biomedicina, abrangendo desde participantes de pesquisa até componentes bioquímicos cruciais para estudos de tratamentos e respostas metabólicas. Por outro lado, categorias menos frequentes, que focam em estruturas anatômicas, são tipicamente mais específicas para áreas como a oncologia e aparecem com menos frequência nos estudos sobre diabetes, refletindo sua aplicabilidade limitada, porém válida, nesse contexto específico.

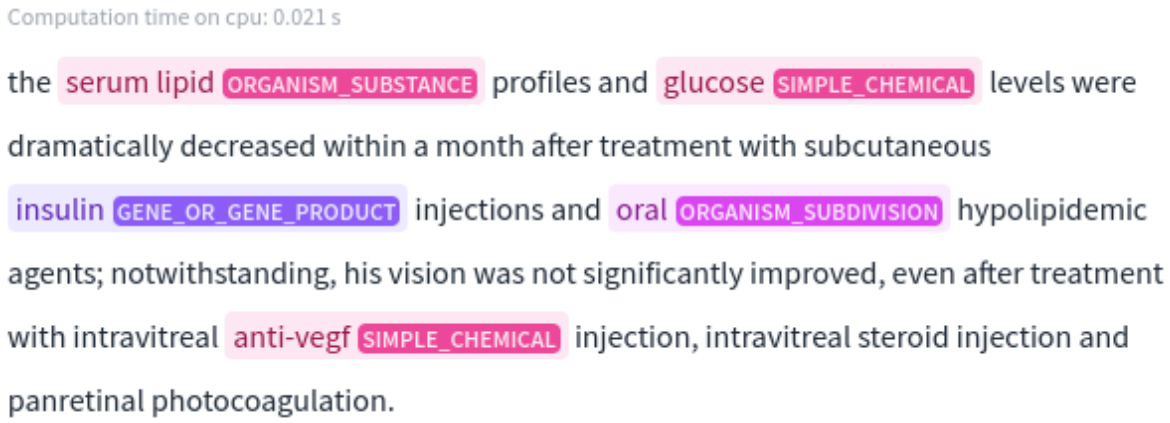
A aplicação do modelo nas sentenças pode ser visualizada detalhadamente: para a sentença A0S6, as informações estão na Figura 41, e para a sentença A0S7, na Figura 42.

Figura 41 - Entidades identificadas em A0S6 - en_ner_bionlp13cg_md



Fonte: Elaborado pelo autor (2024)

Figura 42 - Entidades identificadas em A0S7 - en_ner_bionlp13cg_md



Fonte: Elaborado pelo autor (2024)

O modelo demonstra ser válido devido à sua ampla gama de rótulos identificáveis. No entanto, enfrenta limitações significativas por estar focado em uma subárea distinta da diabetes mellitus, não conseguindo identificar entidades críticas como 'hypolipidemic agents', 'vision', 'intravitreal', 'steroid', e 'panretinal photocoagulation'. Essa especialização pode restringir sua utilidade no contexto específico desta pesquisa sobre diabetes, sugerindo a necessidade de adaptações ou do uso complementar de outros modelos mais alinhados com os requisitos do estudo.

5.3.3 Modelo - en_core_sci_sm

O modelo “en_core_sci_sm”, acessado através da biblioteca scispaCy, é especializado em textos científicos e biomédicos. Ele é eficiente para processar

textos rapidamente, ideal para tarefas que demandam velocidade e eficiência, embora ofereça menos detalhes em comparação com modelos maiores. Utiliza uma arquitetura de rede neural convolucional (CNN), adequada para reconhecer padrões textuais de forma eficaz em análises em grande escala.

Este modelo identifica apenas um rótulo genérico “Entity”, sem distinguir tipos específicos de entidades biomédicas, o que pode simplificar a integração em sistemas que não necessitam de uma classificação detalhada. Apesar dessa limitação, ele é útil para a identificação básica de entidades. O Quadro 8 lista os rótulos de entidades, mantendo uma abordagem padronizada para facilitar a comparação com outros modelos utilizados na pesquisa.

Quadro 8 - Categorias existentes no modelo en_core_sci_sm

Índice	Rótulo
01	ENTITY

Fonte: Elaborado pelo autor (2024)

O resultado dos dois experimentos, sendo o primeiro com uma amostra aleatória de 1000 registros e o segundo contendo todos os 361.688 registros, pode ser observado na Tabela 2.

Tabela 2 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_sm

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	ENTITY	66420	24539167

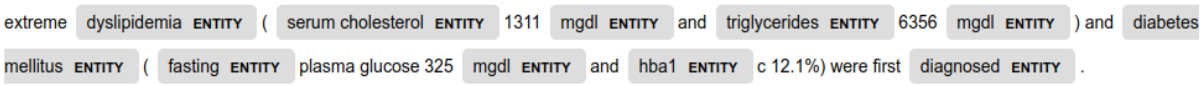
Fonte: Elaborado pelo autor (2024)

Ao contrário do modelo 'en_ner_bionlp13cg_md', que identificou apenas 14.933 entidades na amostra e 5.559.966 no corpus completo, o modelo 'en_core_sci_sm' detectou um número significativamente maior de entidades. Apesar de utilizar um rótulo genérico ('ENTITY'), este modelo reconheceu 66.420 entidades na amostra, representando um aumento de aproximadamente 344,79% em relação ao 'en_ner_bionlp13cg_md'. Para o corpus completo, identificou 24.539.167 entidades, o que representa um aumento de cerca de 341%. Esses

dados destacam a capacidade expansiva do 'en_core_sci_sm' em capturar entidades, mesmo com uma categorização menos específica.

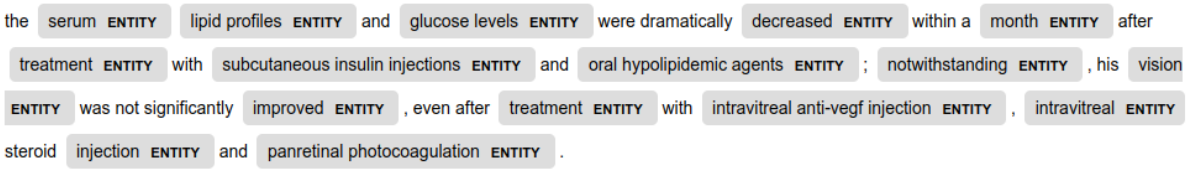
A aplicação do modelo nas sentenças pode ser visualizada detalhadamente: as informações para a sentença A0S6 estão na Figura X, e para a sentença A0S7, na Figura Y.

Figura 43 - Entidades identificadas em A0S6 - en_core_sci_sm



Fonte: Elaborado pelo autor (2024)

Figura 44 - Entidades identificadas em A0S7 - en_core_sci_sm



Fonte: Elaborado pelo autor (2024)

Embora o modelo tenha identificado um número consideravelmente maior de entidades, ele não conseguiu reconhecer todas, como “steroid” (A0S7) e “plasma glucose” (A0S6). Contudo, houve uma melhoria notável na capacidade de identificação de entidades, mesmo sob um rótulo genérico. Diante disso, surgiu a hipótese de uma possível melhora ao combinar modelos com rótulos genéricos a ontologias e bases de dados sobre doenças, entre outras definições relevantes, formando uma estratégia integrada para uma rotulação mais precisa das entidades identificadas. Esta abordagem será explorada mais detalhadamente nas seções subsequentes.

Modelo - en_core_sci_scibert

O modelo “en_core_sci_scibert” é baseado na tecnologia SciBERT da AllenAI, especializado em textos biomédicos e científicos, e integrado ao framework spaCy. Ele oferece robusta eficiência em tarefas de processamento de linguagem natural, utilizando a base “allenai/scibert”, um modelo Transformer treinado especificamente em literatura científica. Isso o torna particularmente eficaz para

compreender e processar terminologias e conceitos em textos de diabetes mellitus, facilitando a identificação precisa de termos técnicos e entidades biomédicas, essenciais para a extração e análise de informações complexas em pesquisas médicas.

A integração desse modelo em projetos de mineração de dados na área da Biomedicina pode potencializar a análise de dados, permitindo uma identificação mais precisa de entidades e relações específicas do campo biomédico, contribuindo significativamente para a construção de um Grafo de Conhecimento mais rico e informativo sobre diabetes mellitus.

O resultado dos dois experimentos — o primeiro com uma amostra aleatória de 1000 registros e o segundo contendo todos os 361.688 registros — pode ser observado na Tabela 3.

Tabela 3 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_scibert

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	ENTITY	65319	24075959

Fonte: Elaborado pelo autor (2024)

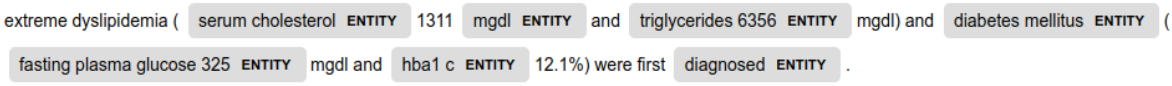
Os modelos “en_core_sci_sm” e “en_core_sci_scibert”, focados em Biomedicina e utilizando um rótulo genérico “ENTITY”, demonstraram desempenho superior na identificação de entidades comparados ao “en_ner_bionlp13cg_md”, que possui rótulos específicos. No corpus completo de 361.688 documentos, o “en_core_sci_sm” identificou 24.539.167 entidades, um aumento de aproximadamente 344% em relação ao “en_ner_bionlp13cg_md”, que identificou 5.559.966 entidades. Similarmente, o “en_core_sci_scibert” identificou 24.075.959 entidades, um aumento de cerca de 333%.

Essas diferenças mostram a eficácia dos modelos com rótulo genérico em capturar uma maior quantidade de entidades, apesar da menor especificidade. Isso sugere um trade-off entre especificidade e cobertura no reconhecimento de entidades. A criação de corpus com rótulos genéricos pode ser facilitada por técnicas não supervisionadas ou abordagens de zero-shot learning, que não exigem anotações detalhadas para cada tipo de entidade. Confirmar essas hipóteses requer um estudo detalhado sobre os corpus usados para treinar cada modelo,

proporcionando insights sobre as metodologias de treinamento e capacidades dos modelos de reconhecimento de entidades nomeadas.

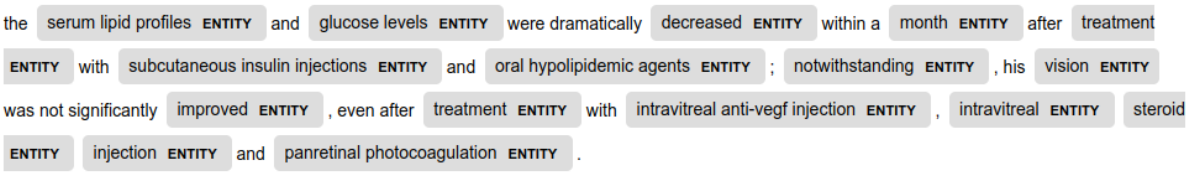
A aplicação do modelo nas sentenças pode ser visualizada em detalhes: as informações referentes à sentença A0S6 estão apresentadas na Figura 45, enquanto os dados da sentença A0S7 podem ser observados na Figura 46.

Figura 45 - Entidades identificadas em A0S6 - en_core_sci_scibert



Fonte: Elaborado pelo autor (2024)

Figura 46 - Entidades identificadas em A0S7 - en_core_sci_scibert



Fonte: Elaborado pelo autor (2024)

Os dois modelos, “en_core_sci_sm” e “en_core_sci_scibert”, mostram variações significativas na precisão da identificação de entidades em textos biomédicos. Por exemplo, na sentença A0S6, o modelo “en_core_sci_sm” reconhece “triglycerides” mas não associa o número relevante 6356, enquanto o “en_core_sci_scibert” fez essa distinção corretamente. No entanto, ambos falham em reconhecer “mgdl” como “mg/dl” devido à ausência do símbolo “/”. Na expressão “fasting plasma glucose 325”, “en_core_sci_sm” identifica apenas “fasting”, omitindo partes críticas, e ambos os modelos parcialmente reconhecem “hba1 c 12.1%”. Essas imprecisões indicam a necessidade de padronização na rotulação durante o treinamento dos modelos e uma curadoria cuidadosa do corpus de treinamento para melhorar a identificação de termos especializados. A discussão sobre como tratar medidas e percentuais como partes integrantes das entidades ou relações

separadas permanece relevante para aprimorar a modelagem de dados em biomedicina.

5.3.4 Modelo – en_core_sci_lg

Os modelos “en_core_sci_sm” e “en_core_sci_lg” da scispaCy são especializados em NLP para textos científicos e biomédicos em inglês. A principal diferença entre eles é o volume de treinamento, com “en_core_sci_sm” treinado com 6.7 bilhões de tokens e “en_core_sci_lg” com 13.7 bilhões de tokens. O “en_core_sci_lg” possui 600.000 vetores de palavras únicos, resultando em maior precisão e eficácia, mas exigindo mais recursos computacionais.

O “en_core_sci_sm” é otimizado para eficiência e velocidade, adequado para ambientes com recursos limitados. Já o “en_core_sci_lg” oferece maior precisão, sendo ideal para projetos onde a precisão é crucial e há disponibilidade de recursos. A escolha entre os modelos deve considerar as necessidades específicas do projeto e os recursos computacionais disponíveis.

O resultado dos dois experimentos, o primeiro com uma amostra aleatória de 1000 registros e o outro contendo todos os 361688 registros pode ser observada na Tabela 4

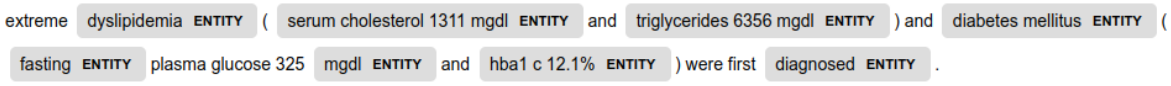
Tabela 4 - Quantidade de entidades reconhecidas pelo modelo en_core_sci_lg

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	ENTITY	64475	23747637

Fonte: Elaborado pelo autor (2024)

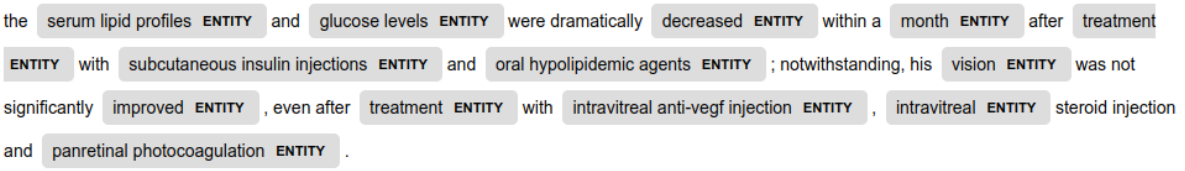
Curiosamente, o modelo “en_core_sci_lg” identificou menos entidades em comparação ao “en_core_sci_sm”, com contagens de 64.475 e 23.747.637 contra 66.420 e 24.539.167 para a amostra e para o corpus completo, respectivamente. Para uma avaliação mais detalhada, é essencial analisar as entidades identificadas nas sentenças A0S6 e A0S7. As análises correspondentes podem ser visualizadas nas Figuras 47 e 48, respectivamente.

Figura 47 - Entidades identificadas em A0S6 - en_core_sci_lg



Fonte: Elaborado pelo autor (2024)

Figura 48 - Entidades identificadas em A0S7 - en_core_sci_lg



Fonte: Elaborado pelo autor (2024)

Na análise da sentença A0S6, o modelo “en_core_sci_lg” demonstrou maior precisão, identificando corretamente as entidades “serum cholesterol 1311 mg/dl”, “triglycerides 6356 mg/dl” e “hba1c 12.1%”. No entanto, ambos os modelos falharam em identificar “plasma glucose 325”, sugerindo uma possível lacuna nos rótulos de treinamento para esse termo específico. Na sentença A0S7, o “en_core_sci_lg” mostrou superioridade ao reconhecer “serum lipid profiles” como uma única entidade, enquanto o modelo “en_core_sci_sm” fragmentou o termo em “serum” e “lipid profiles”. Além disso, o modelo “en_core_sci_lg” corretamente não reconheceu “notwithstanding” como uma entidade, refletindo seu uso gramatical como uma preposição ou advérbio que indica concessão, semelhante a “apesar de”. Isso reforça a assertividade do modelo “en_core_sci_lg” na identificação de entidades, consolidando sua aplicabilidade em análises que exigem discernimento contextual.

5.3.5 Modelo - en_ner_bc5cdr_md

O modelo “en_ner_bc5cdr_md”, oferecido pela biblioteca scispaCy, é especializado em textos biomédicos. Foi treinado utilizando o corpus BC5CDR, que inclui anotações detalhadas sobre doenças e compostos químicos, extraídas de aproximadamente 1.5 milhão de artigos da PubMed. Essa base de treinamento robusta permite ao modelo identificar e classificar com precisão entidades

biomédicas específicas, tornando-o especialmente valioso para pesquisadores em áreas como farmacologia e genética.

Utilizando uma arquitetura neural sofisticada baseada em Bi-LSTM-CRF, o “en_ner_bc5cdr_md” processa textos com uma compreensão profunda do contexto, capturando informações de ambas as direções do texto e garantindo previsões consistentes através do uso de Conditional Random Fields (CRF). Esta configuração melhora significativamente a precisão da identificação de entidades e sua eficácia em prever relações e contextos dentro dos textos.

Na prática, o “en_ner_bc5cdr_md” é utilizado para a extração de informações específicas em artigos científicos e na curadoria de grandes bases de dados biomédicos. Sua capacidade de identificar termos críticos, como doenças e agentes químicos, o torna uma ferramenta essencial para a análise avançada de textos biomédicos, contribuindo para o avanço da pesquisa médica e científica. O modelo reconhece dois rótulos principais, detalhados no Quadro 9.

Quadro 9 - Categorias existentes no modelo en_ner_bc5cdr_md

Indice	Rótulo
01	CHEMICAL
02	DISEASE

Fonte: Elaborado pelo autor (2024)

O resultado dos dois experimentos, o primeiro com uma amostra aleatória de 1000 registros e o segundo contendo todos os 361.688 registros, pode ser observado na Tabela X.

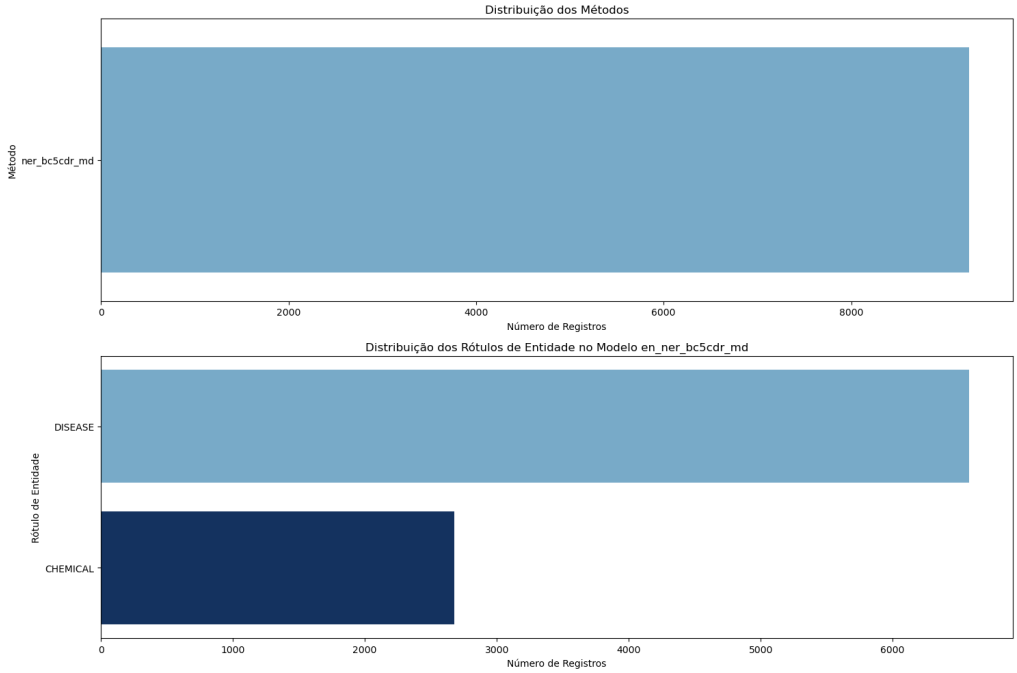
Tabela 5 - Quantidade de entidades reconhecidas pelo modelo en_ner_bc5cdr_md

Índice	Rótulo	Descobertos na Amostra	Descobertos no Corpus Completo
01	CHEMICAL	64475	1084190
02	DISEASE	6582	2388576

Fonte: Elaborado pelo autor (2024)

A representação gráfica dos dados para a amostra de 1000 registros pode ser observada na Figura 49.

Figura 49 - Distribuição de entidades reconhecidas pelo modelo en_ner_bc5cdr_md

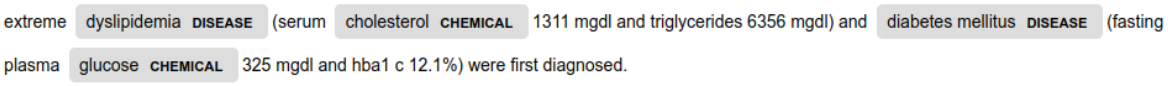


Fonte: Elaborado pelo autor (2024)

O modelo “en_ner_bc5cdr_md” possui o menor número de entidades reconhecidas em comparação a outros modelos. Essa menor diversidade nas classes de entidades reconhecíveis é compreensível, dada a especialização do modelo. Comparar “en_ner_bc5cdr_md” com o modelo “en_ner_bionlp13cg_md”, que também possui rótulos específicos, é complexo devido à falta de uma correspondência direta entre os rótulos de ambos. Notavelmente, o modelo “en_ner_bionlp13cg_md” demonstrou uma capacidade superior, identificando 62,01% mais entidades na amostra de 1000 registros e 62,46% mais no corpus completo de 361.688 documentos. Essa diferença quantitativa possivelmente reflete o maior número de classes reconhecíveis (16), proporcionando assim uma maior abrangência.

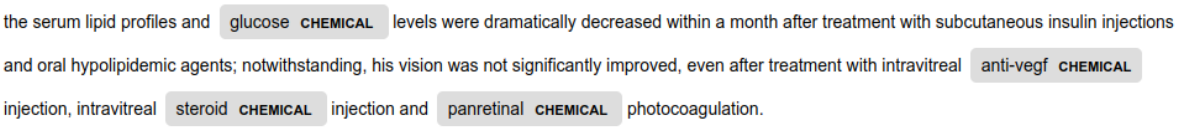
A aplicação do modelo nas sentenças pode ser visualizada detalhadamente: para a sentença A0S6, as informações estão na Figura X, e para a sentença A0S7, na Figura Y.

Figura 50 - Entidades identificadas em A0S6 - en_ner_bc5cdr_md



Fonte: Elaborado pelo autor (2024)

Figura 51 - Entidades identificadas em A0S7 - en_ner_bc5cdr_md



Fonte: Elaborado pelo autor (2024)

Para a sentença A0S6, o modelo identificou apenas parte do termo “serum cholesterol 1311 mg/dl”, reconhecendo somente “cholesterol” e omitindo tanto a medida quanto seu valor. Esta identificação é imprecisa, pois o termo correto é “serum cholesterol”, que se refere especificamente ao colesterol no soro sanguíneo. Uma situação semelhante ocorreu com “fasting plasma glucose”, onde foi reconhecido apenas “glucose”. É crucial enfatizar que a identificação parcial de termos pode levar à atribuição incorreta de classes. Erros similares foram observados também na sentença A0S7, onde a identificação parcial resultou em uma compreensão incompleta do contexto biomédico.

5.3.6 Modelo - biomedical-ner-all

O modelo “biomedical-ner-all” representa um avanço significativo no reconhecimento de entidades nomeadas (NER) para textos biomédicos. Focado nos domínios de biomedicina e epidemiologia, este modelo foi treinado com um vasto corpus que inclui uma variedade de entidades nomeadas, abrangendo desde fatores de risco médicos e sinais vitais até funções biológicas e drogas. Utilizando a arquitetura baseada em Transformers, especificamente BERT, o modelo possui uma compreensão profunda do contexto nos textos, essencial para a precisão na identificação de entidades complexas em documentos biomédicos.

Além de reconhecer diversos tipos de entidades clínicas, o modelo considera fatores não clínicos como idade e gênero, que são determinantes para os resultados de saúde. Isso torna o “biomedical-ner-all” uma ferramenta valiosa para pesquisadores e profissionais da saúde, permitindo uma análise completa e detalhada das informações biomédicas. A flexibilidade do modelo é demonstrada por sua adaptabilidade a outras tarefas de NER dentro dos domínios biomédicos, sendo uma escolha robusta para a curadoria de dados e análise textual.

A disponibilidade do modelo na plataforma Hugging Face Hub enfatiza sua acessibilidade e praticidade para uso em pesquisa e aplicações práticas. Além de extrair informações, o modelo permite a análise de relações entre as entidades e a geração de insights sobre a disseminação de doenças e estatísticas associadas. Assim, o “biomedical-ner-all” não só cumpre os padrões de precisão em benchmarks de NER biomédicos, mas também oferece uma plataforma adaptável e fácil de usar para a pesquisa biomédica e epidemiológica.

A arquitetura do “biomedical-ner-all”, baseada no BERT, é reconhecida por sua capacidade de processar o contexto de palavras em textos, considerando a posição delas. Isso é crucial para tarefas de NER em textos biomédicos, onde o contexto pode alterar significativamente o significado e a classificação de termos técnicos. O modelo foi treinado em um corpus substancial, com cerca de 1.5 milhão de artigos científicos, contendo anotações detalhadas de uma vasta gama de entidades biomédicas. Este extenso conjunto de dados garante que o modelo aprenda a reconhecer uma ampla variedade de entidades biomédicas — especificamente 84 tipos diferentes — com alta precisão, adaptando-se às nuances específicas de termos e relações essenciais para análise e interpretação eficazes em textos na área da saúde.

Entre as 84 classes de entidades biomédicas disponíveis no modelo, 38 foram identificadas no corpus analisado. Para facilitar a visualização, apenas as classes identificadas serão apresentadas, na Tabela 6 pode se observar as 16 classes com maior número de indentificações. No entanto, para aqueles interessados em explorar o conjunto completo de classes, é possível acessar a documentação detalhada do modelo. Este modelo foi aplicado apenas no conjunto de documentos reduzido, a amostra aleatória de 1000 elementos.

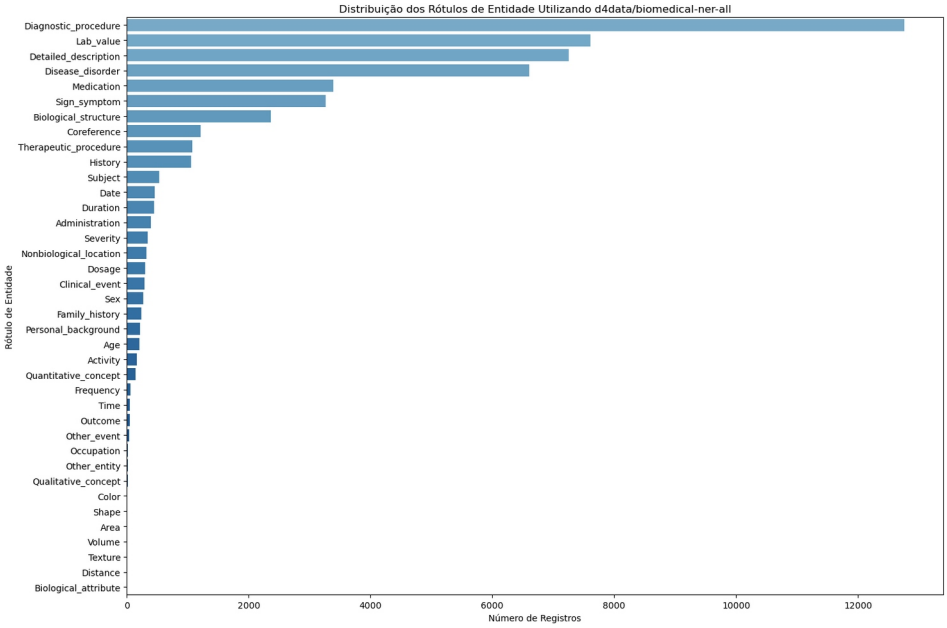
Tabela 6 - Quantidade de entidades reconhecidas pelo modelo biomedical-ner-all. Primeiros 16 registros

Índice	Rótulo	Descobertos na Amostra
01	Diagnostic_procedure	12763
02	Lab_value	7611
03	Detailed_description	7259
04	Disease_disorder	6606
05	Medication	3386
06	Sign_symptom	3266
07	Biological_structure	2364
08	Coreference	1216
09	Therapeutic_procedure	1076
10	History	1054
11	Subject	527
12	Date	462
13	Duration	444
14	Administration	399
15	Severity	346
16	Nonbiological_location	325

Fonte: Elaborado pelo autor (2024)

A representação gráfica dos dados para a amostra de 1000 registros pode ser observada na Figura 52.

Figura 52 - Distribuição de entidades reconhecidas pelo modelo biomedical-ner-all



Fonte: Elaborado pelo autor (2024)

O modelo “biomedical-ner-all” demonstrou uma capacidade notável ao identificar um total de 51.216 entidades distribuídas em 38 classes distintas na amostra de 1.000 documentos, tornando-se o modelo com rótulos específicos que mais identificou entidades. Esta extensa identificação destaca sua eficácia em captar uma ampla gama de informações biomédicas relevantes, refletindo a riqueza e a diversidade dos dados contidos nos textos científicos analisados.

A identificação de “Disease_disorder” e “Sign_symptom” é particularmente relevante para o estudo da diabetes mellitus, mostrando que o modelo consegue captar as nuances das complicações e sintomas associados à doença. Isso é crucial para a extração de conhecimento e formação de uma base de dados estruturada que pode contribuir para um entendimento mais profundo da doença e suas múltiplas manifestações clínicas.

No entanto, categorias como “Medication” e “Therapeutic_procedure” apresentam números mais baixos em comparação com as categorias diagnósticas e descritivas, possivelmente refletindo a natureza do corpus utilizado, que pode estar mais focado em descrições do estado da doença do que em tratamentos ou intervenções médicas. Este aspecto sugere a necessidade de cruzamento com os resultados obtidos na etapa de Topic Modelling para entender melhor essas discrepâncias, revelando conexões entre os tópicos predominantes e a frequência das entidades, proporcionando insights adicionais sobre as áreas de foco dentro do corpus.

A presença de categorias como “Coreference”, “History”, e “Personal_background” demonstra a habilidade do modelo de captar informações contextuais e históricas, essenciais para a compreensão completa dos relatos de casos e estudos longitudinais. A identificação de entidades como “Age” e “Sex” também é relevante, pois são fatores que podem influenciar a prevalência e o manejo da diabetes.

Em conclusão, a aplicação do modelo “biomedical-ner-all” em textos específicos sobre diabetes mellitus mostra que, apesar de algumas limitações na identificação de tratamentos, o modelo é eficaz na extração de uma ampla gama de informações biomédicas cruciais para a pesquisa e prática clínica. Isso reforça o valor da aplicação de técnicas avançadas de NER para a curadoria de dados e análise de textos em biomedicina, possibilitando insights mais aprofundados e formando a base para futuras investigações e aplicações clínicas.

O modelo “biomedical-ner-all”, assim como outros modelos implementados na plataforma Transformers da Hugging Face, oferece uma estrutura de saída altamente funcional. Esta estrutura consiste em uma lista de dicionários que detalham diversos atributos críticos relacionados à identificação das classes de entidades. Informações importantes como “entity_group”, “score” e “word”, somado a atributos adicionais que indicam a localização no texto, são fornecidas. Esta estrutura pode ser observada detalhadamente na Figura 53.

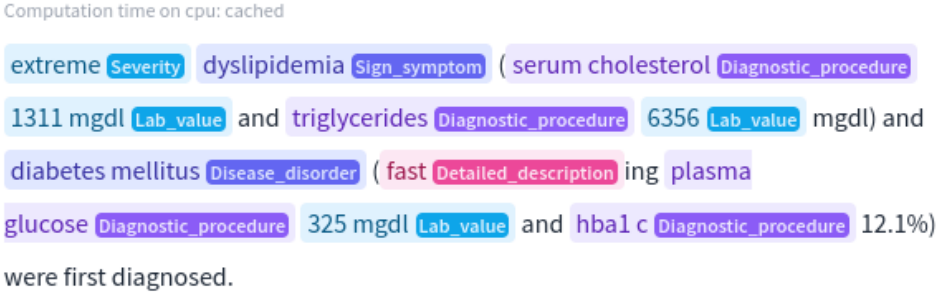
Figura 53 - Amostra do registro da classificação de entidades - biomedicall-ner-all

```
[{'entity_group': 'Diagnostic_procedure',
  'score': 0.999905,
  'word': 'serum lipid profiles',
  'start': 4,
  'end': 24},
 {'entity_group': 'Diagnostic_procedure',
  'score': 0.99990404,
  'word': 'glucose levels',
  'start': 29,
  'end': 43},
 {'entity_group': 'Lab_value',
  'score': 0.99989057,
  'word': 'decreased',
  'start': 62,
  'end': 71},
 {'entity_group': 'Date',
  'score': 0.99894303,
  'word': 'within a month',
  'start': 72,
  'end': 86},
```

Fonte: Elaborado pelo autor (2024)

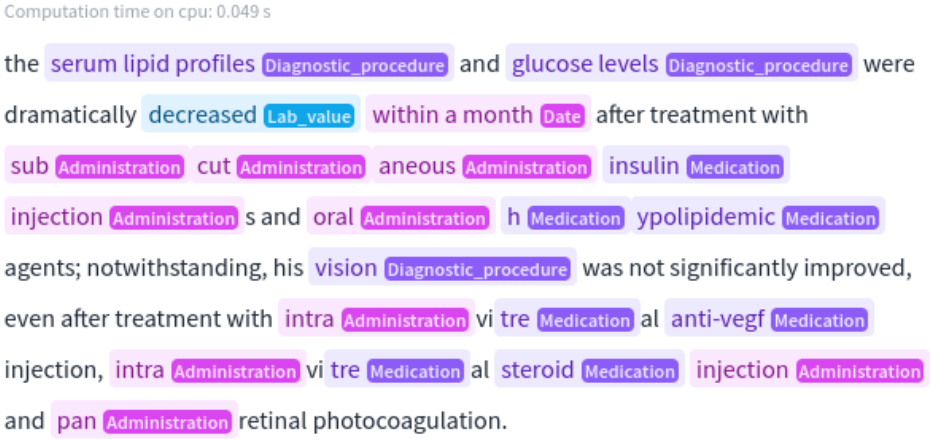
A aplicação do modelo nas sentenças pode ser visualizada detalhadamente nas figuras apresentadas: para a sentença A0S6, as informações estão na Figura 54, e para a sentença A0S7, na Figura 55.

Figura 54 - Entidades identificadas em A0S6 - biomedical-ner-all



Fonte: Elaborado pelo autor (2024)

Figura 55 - Entidades identificadas em A0S7 - biomedical-ner-all



Fonte: Elaborado pelo autor (2024)

Os resultados obtidos com o modelo foram interessantes. Na sentença A0S6, ele foi o único modelo a reconhecer o termo “extreme” como uma entidade, categorizando-o na classe “Severity”. Esse reconhecimento destacou a subsequente entidade “dyslipidemia” na classe “Sign_symptom”. Além disso, o modelo adequadamente separou os termos “plasma glucose 325 mg/dl” e “serum cholesterol 1311 mg/dl” em duas partes distintas, classificando a primeira parte de cada termo como “Diagnostic_procedure” e a segunda parte como “Lab_value”, respectivamente. No entanto, apresentou uma falha ao não considerar a unidade “mg/dl” para “triglycerides”. Por outro lado, na sentença A0S7, o modelo fragmentou erroneamente o termo “intravitreal” em quatro partes: “intra”, “vi”, “tre” e “al”, identificando erroneamente duas dessas partes como entidades enquanto descartava as outras.

5.3.7 Entidades NER e Ontologias: Um Estudo Experimental

O último experimento na área de NER envolveu a compilação de uma lista abrangente de doenças a partir de arquivos de ontologias e bancos de dados especializados. Utilizando essas fontes, foram extraídos nomes de doenças, incluindo sinônimos, abreviações e siglas, para formar uma base unificada e pesquisável. Esta consolidação foi necessária devido ao grande volume de dados gerados, totalizando inicialmente 11.525.905 registros. Processos subsequentes de

eliminação de redundâncias reduziram este número para 1.710.204 entradas únicas, facilitando a gestão e pesquisa dos dados com o auxílio de um indexador, que melhorou significativamente o desempenho das consultas. As fontes utilizadas para compilar esta lista estão detalhadas no Quadro 10.

Quadro 10 - Ontologias e bases de dados de doenças utilizadas na comparação

Arquivo	Tipo	URL
DMTO.owl	Ontologia	https://bioportal.bioontology.org
doid.owl	Ontologia	https://disease-ontology.org/
mondo.owl	Ontologia	https://bioportal.bioontology.org
human_disease_textmining_full.tsv	Banco de Dados	https://download.jensenlab.org/
medicinenet-diseases.json	Banco de Dados	https://github.com/Shivanshu-Gupta/web-scrapers/tree/master/medical_ner
diseases.json	Banco de Dados	https://github.com/dhimmel/het.io-rep-data/tree/master/browser-tables

Fonte: Elaborado pelo autor (2024)

Com a base de dados estruturada, prosseguiu-se com a investigação das entidades genéricas identificadas pelos modelos NER, visando verificar a presença de doenças potencialmente não detectadas por modelos com categorizações mais específicas. Esta análise foi motivada pela hipótese de que o extenso volume de entidades identificadas poderia incluir referências a condições médicas não capturadas explicitamente pelos modelos focados em classes específicas de entidades. Detalhes adicionais sobre as metodologias de pesquisa aplicadas neste contexto serão omitidos nesta descrição.

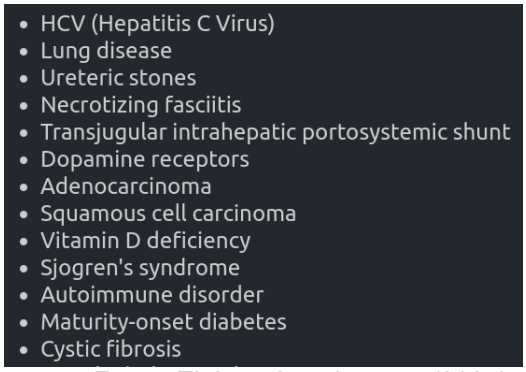
Os resultados deste experimento confirmaram a existência de termos associados a doenças que não haviam sido reconhecidos pelos modelos com foco em categorias específicas. A contagem desses termos descobertos é apresentada no Quadro 11, e uma seleção exemplificativa dessas doenças está disponível para consulta na Figura 56.

Quadro 11 - Quantidade de entidades não reconhecidas pelo modelo ner_bc5cdr_md

Modelo	Quantidade não Identificada
ner_bc5cdr_md	1213

Fonte: Elaborado pelo autor (2024)

Figura 56 - Amostra de entidades não reconhecidas

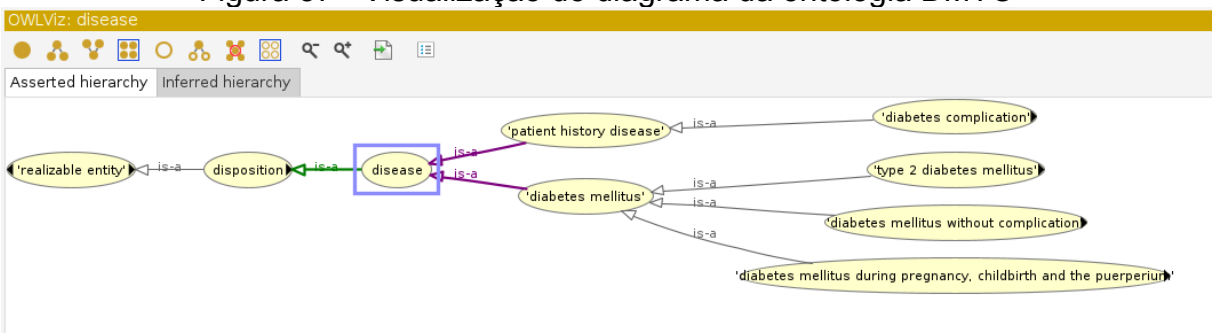


Fonte: Elaborado pelo autor (2024)

Algumas imprecisões na identificação de entidades podem ser atribuídas ao uso de nomenclaturas atípicas, como, por exemplo, 'type II diabetes', onde foi empregado numeral romano em vez de algarismo arábico. Além disso, foi observado que certos termos erroneamente identificados como doenças podem refletir contaminações na base de dados provenientes de uma das ontologias utilizadas.

Um desafio particular reside no fato de que nem todas as ontologias apresentam um atributo de classificação explícito, como 'doença', ou possuem uma estrutura lógica clara devido aos seus propósitos de design. Durante a manipulação da ontologia 'DMTO.owl', ao explorar subclasses de termos que incluem 'diabete' em seus rótulos para identificar doenças, descobriu-se que o termo 'patient history disease' foi, de forma estranha, categorizado como subclasse de 'disease', apesar de não ser uma doença propriamente dita. Este exemplo específico pode ser visualizado na Figura 57.

Figura 57 - Visualização do diagrama da ontologia DMTO



Fonte: Elaborado pelo autor (2024)

É importante destacar que as ontologias representam um recurso inestimável não somente para o aprimoramento dos modelos de NER e ER, mas também contribuem significativamente para o enriquecimento dos KGs. As ontologias frequentemente incluem informações detalhadas sobre sinônimos e estabelecem relações que podem adicionar profundidade e precisão ao conhecimento estruturado. Essa riqueza de informações é fundamental para a construção de KGs mais completos e interconectados, permitindo uma melhor análise e compreensão dos dados científicos.

5.3.8 NER – Avaliação de Resultados

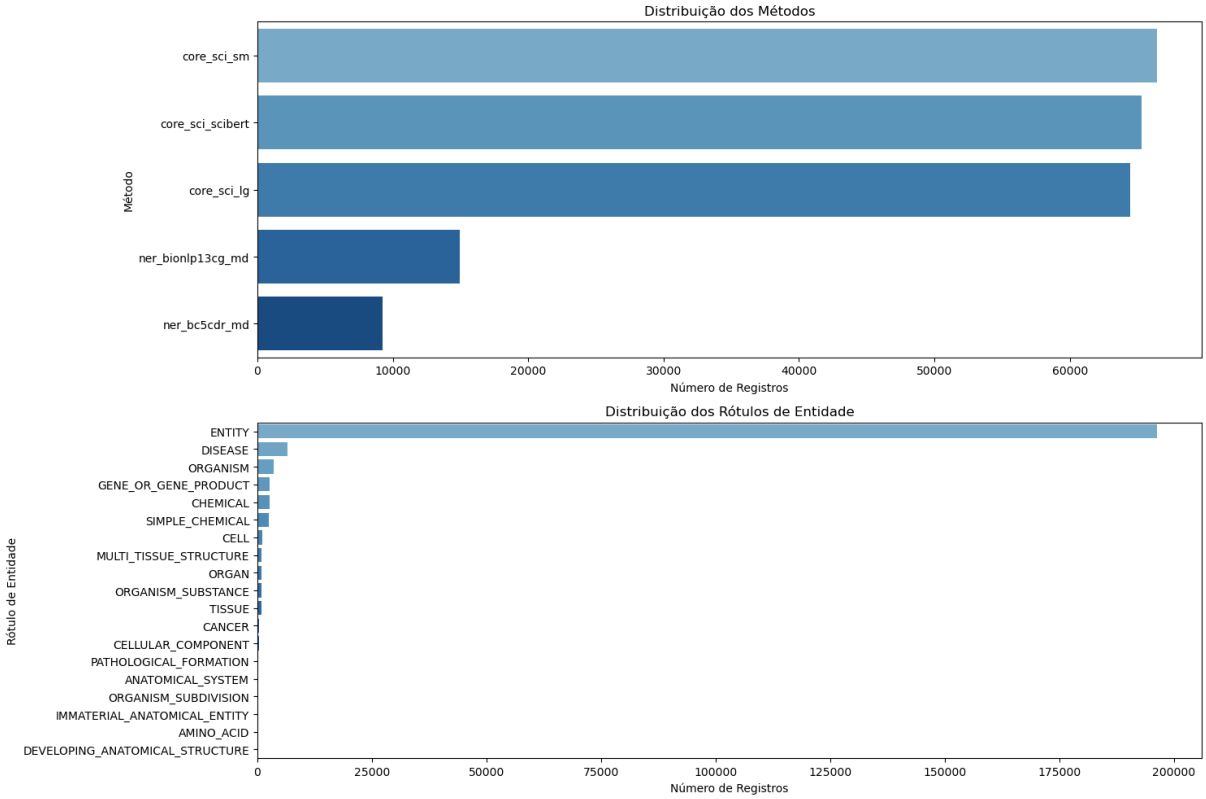
A conclusão dos experimentos que envolvem o NER no campo da biomedicina, particularmente em textos sobre diabetes mellitus, revela insights sobre a funcionalidade e aplicabilidade dos modelos testados. Ao longo dos experimentos, foram utilizados diversos modelos de NER, cada um com suas particularidades arquiteturais e de classificação, para a análise de um amplo corpus de 361,688 abstracts. O modelo en_ner_bionlp13cg_md, com 16 classes de entidades e arquitetura Bi-LSTM-CRF, e o biomedical-ner-all, com 83 classes e base em BERT, mostraram-se capazes de identificar uma grande variedade de entidades biomédicas, o que demonstra sua capacidade de reconhecimento aprofundado.

Os modelos com rótulos genéricos, como o en_core_sci_sm, o en_core_sci_scibert e o en_core_sci_lg, embora menos especializados nas categorias de entidades, identificaram um maior número de entidades. Isto sugere uma vantagem em termos de cobertura de entidades, embora possa comprometer a especificidade necessária para aplicações clínicas precisas. Conforme mostrado na Figura X, esses modelos destacaram-se ao identificar mais de 60.000 registros cada um, enquanto os modelos com rótulos específicos, como o en_ner_bc5cdr_md e o en_ner_bionlp13cg_md, identificaram menos registros, refletindo sua especialização em categorias mais restritas.

A distribuição dos rótulos de entidade, conforme Figura 58, evidencia a predominância do rótulo genérico “ENTITY”, com mais de 175.000 registros. Outros rótulos importantes, como “DISEASE”, “ORGANISM”, e “GENE_OR_GENE_PRODUCT”, também foram identificados, embora em menor quantidade. Esta análise detalha a capacidade dos modelos em reconhecer uma

ampla gama de termos biomédicos, variando desde entidades genéricas até categorias específicas e detalhadas.

Figura 58 - Distribuição de entidades reconhecidas por todos os modelos



Fonte: Elaborado pelo autor (2024)

5.3.9 Entity Relation Extraction

A RE é uma tarefa essencial em NLP, focada em identificar e classificar relações semânticas entre entidades mencionadas em textos. Esse processo é vital para transformar textos não estruturados em conhecimento estruturado, facilitando a criação de KGs e sistemas de resposta a perguntas.

Com o avanço das técnicas de deep learning, a RE tem se beneficiado de modelos como CNNs, RNNs e Transformers. Os Transformers, em particular, são eficazes em entender contextos complexos e relações sutis entre entidades, sendo a escolha predominante para tarefas que exigem uma compreensão aprofundada do contexto.

A fim de implementar essas técnicas foi utilizada a linguagem Python, e de bibliotecas como spaCy, Hugging Face's Transformers, TensorFlow e PyTorch.

Estas ferramentas facilitam o desenvolvimento e a experimentação com diferentes abordagens de RE, especialmente para textos biomédicos complexos.

Durante a pesquisa, foram identificados modelos de RE com foco em biomedicina. No entanto, poucos atendem aos requisitos específicos deste estudo, como a capacidade de formar triplas de sujeito, predicado e objeto. Detalhes sobre os modelos considerados estão listados no Quadro 12.

Quadro 12 - Modelos de RE identificados e analizados

Modelo	Requisito	URL
BERT-based nominal SRL	GPU	https://github.com/CogComp/SRL-English
transformer-srl	Não Identificado	https://github.com/Riccorl/transformer-srl
srl-en_mbert-base	Não Identificado	https://huggingface.co/liaad/srl-en_mbert-base
AllenNLP BERT-base-SRL	Transformer	https://storage.googleapis.com/allennlp-public-models/bert-base-srl-2020.11.19.tar.gz
OpenNRE	Não Identificado	https://github.com/thunlp/OpenNRE
BERT-GT	NVIDIA Tesla V100 SXM2	https://github.com/ncbi/bert_gt
BioREx	NVIDIA Tesla V100 SXM2	https://github.com/ncbi/BioREx?tab=readme-ov-file

Fonte: Elaborado pelo autor (2024)

Neste trabalho, optou-se por empregar a biblioteca AllenNLP para implementar um algoritmo de SRL (Semantic Role Labeling), utilizando o modelo “BERT-base-SRL”. O objetivo foi extrair triplas a partir de textos, utilizando as entidades previamente identificadas nos experimentos de NER.

Para o experimento, foi criada uma função responsável por processar uma sentença e identificar os papéis semânticos associados às entidades nela presentes. Essa função utiliza previsões de SRL para localizar verbos e seus argumentos correspondentes, classificados como “agente” e “paciente”. Quando um agente e um paciente coincidem com as entidades reconhecidas pelos experimentos de NER na

mesma sentença, uma tripla é formada, com o verbo atuando como predicado, estabelecendo uma relação semântica clara entre as duas entidades.

Uma visão macro do algoritmo pode ser apreciada na Figura 59.

Figura 59 - Algoritmo para identificação de triplas

```
Para cada sentença
    texto = sentança.texto
    entidades = sentança.entidades
    predicoes_srl = modelo.prediçoes(texto)
    tripla = vazio
    Para cada Verbo em predicoes_srl.verbos
        verbo_texto = verbo.texto
        verbo_tags = verbo.tags
        Para cada Tag em verbo_tags
            Se a Tag é igual a agente entao faça
                palavras_agente adiciona palavra
            Se a Tag é igual a paciente entao faça
                palavras_paciente adiciona palavra
        Se existem agentes e pacientes em entidades para o verbo então faça
            tripla = monta_tripla(agente, verbo, paciente)
    retorna tripla
```

Fonte: Elaborado pelo autor (2024)

Esse método permite uma análise mais profunda dos textos, facilitando a construção de grafos de conhecimento e a compreensão de relações contextuais complexas dentro dos dados biomédicos explorados. O experimento foi realizado sobre a mesma amostra de sentenças, contendo 1000 registros, utilizada nos experimentos de NER. A quantidade de triplas geradas pode ser observada na Tabela X.

Quadro 13 - Quantidade de triplas identificadas para a amostra de 1000 abstracts

Modelo	Qtd. Sentenças	Qtd. Triplas Geradas
BERT-base-SRL	1000	2795

Fonte: Elaborado pelo autor (2024)

Para o armazenamento das triplas extraídas, adotou-se inicialmente uma estrutura composta por três colunas, conforme visualizado na Figura 60. Contudo, considera-se a incorporação de elementos adicionais nesta estrutura, especificamente o ID do abstract e o ID da sentença. Essa expansão visa enriquecer a capacidade de realizar investigações analíticas mais profundas em etapas futuras deste estudo. Tal metodologia permitirá uma interrogação mais precisa e contextualizada dos dados, contribuindo significativamente para a robustez e a profundidade analítica da pesquisa em curso.

Figura 60 - Estrutura do dataframe que armazena as triplas

```
RangeIndex: 2795 entries, 0 to 2794
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sujeito    2795 non-null   object
 1   Predicado  2795 non-null   object
 2   Objeto     2795 non-null   object
dtypes: object(3)
memory usage: 65.6+ KB
```

Fonte: Elaborado pelo autor (2024)

Uma amostra das triplas extraídas é apresentada na Figura 61, onde é possível observar que diversas entidades, inicialmente dispersas nas sentenças e delimitadas por vírgulas, foram consolidadas em uma única entidade. Este fenômeno, previamente observado, requer atenção meticulosa para garantir a construção adequada do Grafo de Conhecimento. A correta individualização das entidades é fundamental para a integridade e precisão da estrutura de dados, sendo um desafio crucial a ser superado no desenvolvimento subsequente do projeto.

Figura 61 - Amostra de triplas descobertas

	Sujeito	Predicado	Objeto
0	reovirus type 3 , passaged in pancreatic beta ...	produced	an insulitis
1	beta cells	containing	insulin
2	by protein wasting secondary to hypergluconeog...	characterized	cushing 's syndrome
3	by insulin treatment of rats with chronic step...	induced	the increased liver dna
4	steptozotocin	induced	diabetes
5	by an active process having a half - time of 3...	restored	the dna content of the organ
6	an active process	having	a half - time of 32 days
7	a mechnanism	acts	to restore normal liver cellularity when an ov...
8	severe diabetic ketoacidosis and hyperkalemia	presented	with an ecg resembling an acute anterior wall ...
9	subsequent work - up including exercise testin...	ruled	any significant coronary artery disease

Fonte: Elaborado pelo autor (2024)

A análise revelou que o termo “diabetes” aparece tanto na posição de sujeito quanto de objeto nas triplas extraídas, indicando que a diabetes desempenha papéis variados dentro das relações semânticas abordadas. Os detalhes específicos,

incluindo a frequência e o contexto em que “diabetes” assume cada papel, podem ser observados na Tabela 7.

Tabela 7 - Identificação do termo diabetes como sujeito e como objeto na tripla

Papel	Frequência
Sujeito	163
Objeto	279

Fonte: Elaborado pelo autor (2024)

As triplas geradas foram armazenadas em uma base de dados, seguindo o procedimento padrão adotado para todas as etapas deste estudo. Adicionalmente, foi gerado um arquivo em formato RDF para facilitar a interoperabilidade e a análise dos dados. Uma amostra deste arquivo RDF pode ser visualizada na Figura 62, ilustrando a estrutura e o conteúdo das triplas em um formato padronizado e acessível.

Figura 62 - Amostra do RDF gerado ao salvar as triplas

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ns1: <http://pablovigneaux.org/diabetes/kg#> .

ns1:10072 a ns1:Projeto_Dissertacao ;
    ns1:attended ns1:the_second_examination_of_the_atherosclerosis_risk_in_comm

ns1:1058_patients_at_53_centers_in_the_united_states a ns1:Projeto_Dissertacao
    ns1:had ns1:a_newly_diagnosed_lesion_in_a_native_coronary_artery .

ns1:108_patients_with_type_2_diabetes a ns1:Projeto_Dissertacao ;
    ns1:have ns1:structural_heart_disease .

ns1:10_patients__32_%25_ a ns1:Projeto_Dissertacao ;
    ns1:had ns1:recanalized_residual_laa_cavities__which_were_morphologically_s
    ns1:recanalized ns1:residual_laa_cavities__which_were_morphologically_simil

ns1:110_poorly_controlled_insulin_-_naive_type_2_diabetic_patients a ns1:Projet
    ns1:receive ns1:metformin .

ns1:111_subjects_with_diabetes_mellitus a ns1:Projeto_Dissertacao ;
    ns1:had ns1:hba1c_and_arterial_elasticity_determined_in_an_academic_outpati
```

Fonte: Elaborado pelo autor (2024)

O modelo BERT aplicado ao SRL apresentou um desempenho satisfatório, evidenciando uma excelente capacidade na identificação e extração de verbos. No entanto, recomenda-se a avaliação de outros modelos disponíveis para SRL,

incluindo a experimentação com variantes do BERT. Essa abordagem permitirá explorar e comparar diferentes capacidades de identificação, visando aprimorar ainda mais a precisão da extração semântica.

5.3.10 Avaliação e visulização do Knowledge Graph

Esta atividade consistiu na análise visual e métrica dos resultados obtidos na etapa anterior. As análises foram realizadas tanto sobre o conteúdo das triplas, apresentadas em formato de dataframe, quanto na visualização gráfica dessas triplas. Além disso, foram avaliadas as métricas obtidas do grafo, conforme proposto no projeto.

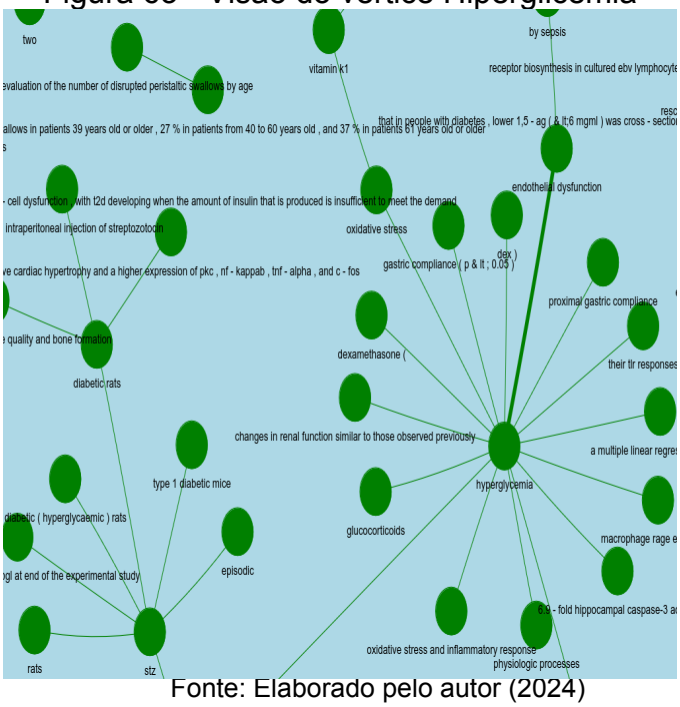
A análise do conteúdo das triplas envolveu a verificação da precisão e coerência das entidades e relações extraídas durante as fases de NER e ER, garantindo que as triplas representassem corretamente as relações semânticas identificadas nos textos.

Dado o escopo definido para este trabalho, após a montagem e visualização do grafo, não foram necessárias ações adicionais. Uma representação com,pleta do Grafo de Conhecimento gerado não possui muito valor analítico, por isto são apresentadas visões focadas em vértices espacíficos como na Figura 63. Essa abordagem visual foi realizada com a biblioteca NetworkX, que pode oferecer uma visualização clara e intuitiva das inter-relações entre as entidades e seus respectivos predicados. No entanto, quando o tamanho do grafo excede as dezenas de vértices, a visualização pode perder valor, não tendo maior utilidade do que fornecer uma noção do tamanho. Para contornar isso, foram feitas análises utilizando zoom do grafo gerado, o que é possibilitado pela mesma ferramenta.

As métricas do grafo foram calculadas e analisadas para avaliar a densidade, centralidade e outras características estruturais importantes, proporcionando uma compreensão mais profunda da organização e conectividade das entidades no KG.

Essa atividade de visualização e análise métrica permitiu validar a eficácia do método utilizado na construção do grafo e ofereceu insights valiosos sobre a estrutura e interconexões das entidades biomédicas, cumprindo assim os objetivos propostos inicialmente para o projeto.

Figura 63 - Visão do vértice Hiperglicemia



Fonte: Elaborado pelo autor (2024)

Uma visão detalhada do nó “diabetes”, com rótulos claramente visíveis para facilitar o entendimento, está disponível na Figura 64. Esta perspectiva ampliada permite uma melhor apreciação das conexões e dos atributos associados a este nó específico dentro do Grafo de Conhecimento, destacando sua centralidade e relevância no contexto estudado.

Figura 64 - Zoom dado no Grafo gerado, no centro o vértice da diabetes

Fonte: Elaborado pelo autor (2024)

A imagem mostra um grafo com o termo central “diabetes”, que está conectado a várias outras entidades. Essas conexões representam as relações semânticas extraídas dos documentos científicos analisados. Aqui estão algumas observações específicas sobre o grafo:

- Conectividade do Nó Central (Diabetes): O nó “diabetes” serve como um hub central, conectando-se a diversos outros termos como “streptozotocin”, “type 1 diabetic mice”, “diabetic rats”, “fasting blood glucose”, “renal replacement therapy”, entre outros. Isso indica a centralidade e a relevância do termo “diabetes” no contexto dos documentos analisados.
- Diversidade de Termos Relacionados: As entidades relacionadas cobrem uma ampla gama de tópicos dentro da pesquisa sobre diabetes, incluindo tratamentos (ex. “streptozotocin”), populações estudadas (ex. “type 1 diabetic mice”, “rats”), complicações associadas (ex. “arterial atherogenesis”, “cardiac hypertrophy”), e fatores de risco (ex. “fasting blood glucose”).
- Estrutura das Conexões: O grafo exibe uma estrutura em estrela, onde “diabetes” é o ponto central com muitas ramificações. Isso sugere que “diabetes” é o tema principal que conecta uma vasta gama de subtemas e termos relacionados.
- Especificidade das Conexões: Algumas conexões são bastante específicas, como “streptozotocin (stz 50 mg/kg)” e “streptozotocin (stz 60 mg/kg body wt)”, indicando detalhes experimentais que foram capturados nos textos analisados.
- Possíveis Problemas na Extração: Observa-se que algumas entidades podem estar fragmentadas ou mal agrupadas, como “only thermo - reversible insulin liquid suppository [insulinp407f4-2]”. Isso pode indicar a necessidade de refinamento nas etapas de pré-processamento e extração de entidades para evitar fragmentações e garantir que entidades compostas sejam reconhecidas como um único termo.

No grafo, também é visível a presença de arestas mais espessas, indicando um peso maior atribuído com base na frequência das ocorrências do relacionamento entre os nodos. Esse padrão sugere que determinadas conexões foram repetidamente observadas em várias sentenças, ressaltando a importância e a recorrência de certos relacionamentos no conjunto de dados analisado.

É importante mencionar que a qualidade do grafo pode ser influenciada por certas imprecisões, como erros ortográficos na nomeação de entidades e a

identificação de agrupamentos extensivos de termos pelos modelos. Essas inconsistências podem comprometer a coesão do grafo, visto que pequenas variações nos termos das entidades resultam em nodos distintos, impactando diretamente a estrutura e a interpretação do grafo de conhecimento.

A partir do grafo construído, foram geradas estatísticas descritivas básicas utilizando a biblioteca NetworkX. Essas métricas fornecem uma visão quantitativa da estrutura do grafo e estão detalhadas na Figura 65. Essas estatísticas incluem medidas como densidade, centralidade e conectividade, oferecendo uma compreensão mais profunda da organização e inter-relações das entidades no grafo, o que é essencial para avaliar sua validade e utilidade no contexto da pesquisa sobre diabetes mellitus.

Figura 65 - Métricas geradas para o KG construído.

	Métrica	Valor
0	Número de Nodos	4.948000e+03
1	Número de Arestas	2.763000e+03
2	Densidade do Grafo	1.128780e-04
3	Número de Componentes Conexas	2.187000e+03
4	Tamanho do Maior Componente Conexo	6.500000e+01
5	Grau Médio	1.116815e+00
6	In-degree Médio	5.584074e-01
7	Out-degree Médio	5.584074e-01
8	Coeficiente de Agrupamento	0.000000e+00
9	Centralidade de Grau Média	2.257560e-04
10	Centralidade de Proximidade Média	1.207327e-04
11	Centralidade de Intermediação Média	3.683911e-09
12	Diâmetro do Grafo	0.000000e+00

Fonte: Elaborado pelo autor (2024)

Neste segmento, apresento uma análise das métricas do grafo de conhecimento gerado a partir dos dados extraídos durante a etapa de representação de conhecimento. Esta análise é crucial para avaliar a qualidade e a estrutura do grafo, fornecendo insights sobre a conectividade, densidade e distribuição das entidades e suas relações.

O grafo, construído com informações extraídas de abstracts de documentos científicos sobre diabetes mellitus, foi avaliado quanto a diversas métricas estruturais. Os resultados dessas métricas estão detalhados na Tabela X e oferecem uma visão quantitativa das características do grafo. A análise destas métricas

permite identificar padrões e possíveis limitações, além de sugerir hipóteses sobre a organização dos dados e a representatividade das relações semânticas detectadas.

A seguir, exponho cada métrica em detalhe, interpretando seus valores e formulando hipóteses sobre a estrutura do grafo, com o objetivo de compreender melhor a organização das entidades biomédicas e as inter-relações identificadas:

- Número de Nodos (4,948) :Indica a quantidade de entidades distintas identificadas no grafo. O número elevado sugere uma diversidade significativa de termos extraídos dos abstracts.
- Número de Arestas (2,763): Representa a quantidade de relações identificadas entre as entidades. O número relativamente menor de arestas em comparação ao número de nodos sugere uma conectividade baixa.
- Densidade do Grafo (0.00011288): A baixa densidade indica que o grafo é esparso, com poucas conexões entre os nodos. Isso pode ser devido à natureza dos abstracts, que geralmente contêm menos informações detalhadas e conexões explícitas.
- Número de Componentes Conexos (2,187): O alto número de componentes conexos sugere a existência de muitos subgrafos desconectados, o que reforça a hipótese de que muitos vértices são isolados ou formam pequenos clusters.
- Tamanho do Maior Componente Conexo (65): Indica que o maior subgrafo conectado possui apenas 65 nodos. Este valor relativamente baixo implica que a maioria dos nodos não está interconectada em um único componente grande.
- Grau Médio (1.1168): O grau médio baixo indica que, em média, cada nodo está conectado a pouco mais de uma outra entidade. Isso pode ser um reflexo da natureza resumida dos abstracts.
- In-degree Médio (0.5584) e Out-degree Médio (0.5584): A simetria entre o in-degree e o out-degree médios sugere uma distribuição equilibrada de entradas e saídas de arestas para os nodos.
- Coeficiente de Agrupamento (0.0): O coeficiente de agrupamento zero indica que não há triângulos no grafo, reforçando a falta de clusters de nodos fortemente conectados.
- Centralidade de Grau Média (0.00022575): A centralidade de grau média muito baixa sugere que poucos nodos têm um número significativo de conexões.

- Centralidade de Proximidade Média (0.12073722): Indica que, em média, a distância entre os nodos é relativamente grande, refletindo a baixa conectividade do grafo.
- Centralidade de Intermediação Média (3.683911e-09): O valor extremamente baixo sugere que poucos nodos atuam como intermediários nas conexões entre outros nodos, o que é consistente com a baixa densidade do grafo.
- Diâmetro do Grafo (0): O diâmetro zero é um indicador de que há componentes isolados ou subgrafos desconectados, uma vez que o diâmetro normalmente se aplica ao maior componente conexo.

A partir da análise exposta podemos formular algumas hipóteses:

- Vértices Isolados em Duplas: A existência de muitos vértices isolados ou em pequenos clusters pode ser atribuída ao fato de que os dados foram extraídos de abstracts, que tendem a resumir informações e omitir conexões detalhadas presentes nos textos completos.
- Baixa Conectividade e Agrupamento: A baixa conectividade e a ausência de agrupamentos significativos sugerem que os abstracts fornecem uma visão fragmentada das relações semânticas, dificultando a formação de um grafo densamente conectado.
- Componentes Conexos: O alto número de componentes conexos indica que as entidades estão frequentemente isoladas ou apenas conectadas a uma ou duas outras entidades, refletindo a natureza sucinta dos abstracts.

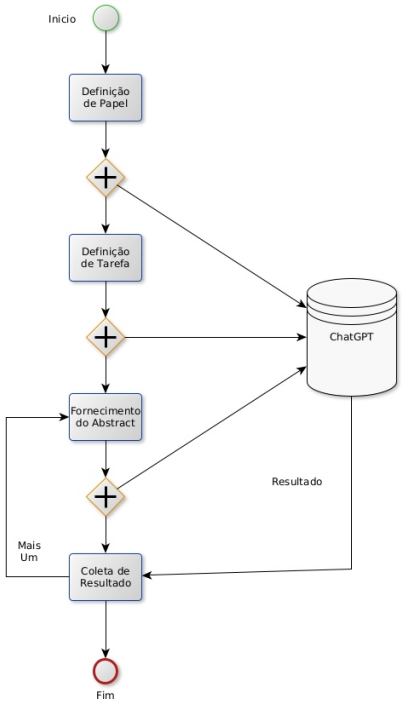
Algumas conclusões podem ser obtidas com base no exposto. No qual as métricas obtidas do grafo indicam uma estrutura esparsa e fragmentada, o que é consistente com as limitações inerentes à extração de dados de abstracts. A análise sugere a necessidade de considerar a inclusão de textos completos em futuras pesquisas para aumentar a densidade e a conectividade do grafo, proporcionando uma visão mais integrada das relações semânticas no domínio biomédico. Pela análise visual foi possível concluir que o grafo possui conhecimento válido, porém necessitando refinamento na extração, seja utilizando textos completos, seja no pré-processamento, ou, seja pelo refinamento na etapa de representação de conhecimento.

5.3.11 Experimento utilizando Generative Pre-trained Transformer

Adicionalmente às investigações tradicionais de NLP que foram desenvolvidas, foram realizados experimentos utilizando o Generative Pre-trained Transformer (GPT) através do prompt do ChatGPT 4, esta solução fazendo uso de LLMs não foi incluída no método proposto devido a possuir uma natureza bem diferente das técnicas mais comúns de NLP. Esta seção visa apresentar um resumo do teste que teve por objetivo explorar as capacidades de Compreensão de Linguagem Natural (NLU) do modelo, com o intuito de alcançar os objetivos da pesquisa de forma mais eficiente e com menor dispêndio de esforços.

Os resultados mais eficazes foram alcançados adotando premissas de clareza com instruções diretas e específicas, o que naturalmente minimiza o ruído na interpretação dos comandos. Verificou-se que simplificar o processo, evitando excesso de complexidade, constitui uma prática recomendável para prevenir ambiguidades. O procedimento foi delimitado em quatro etapas, utilizando apenas três prompts. A estrutura desse fluxo é ilustrada na Figura 66, cujas fases serão detalhadamente discutidas subsequente.

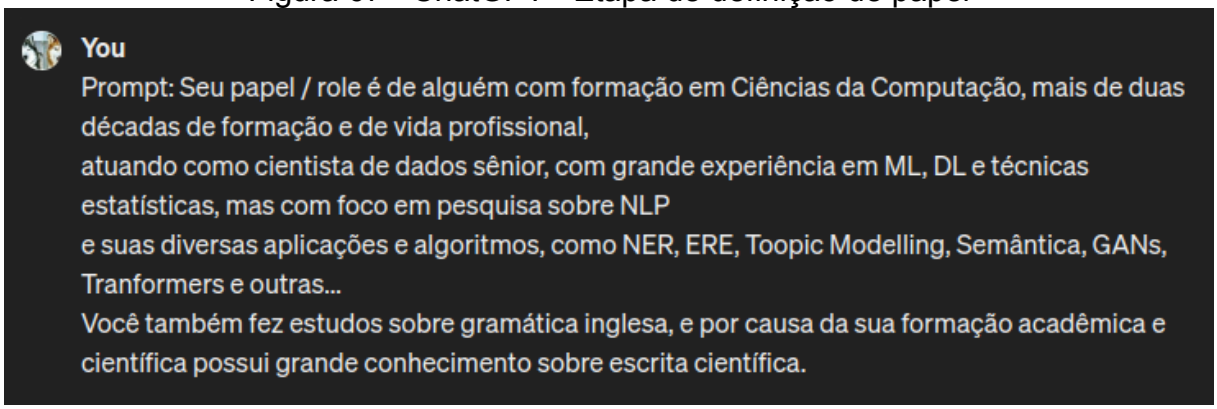
Figura 66 - Fluxo do uso do ChatGPT



Fonte: Elaborado pelo autor (2024)

A primeira etapa do processo envolveu a definição do papel e das capacidades do ChatGPT, estabelecendo um “personagem” com um escopo delimitado de funções. Essa configuração inicial foi crucial para determinar o tipo de conhecimento que o chat poderia gerenciar e aplicar, configurando a base para as interações subsequentes. Esse comando inicial é detalhado na Figura 67.

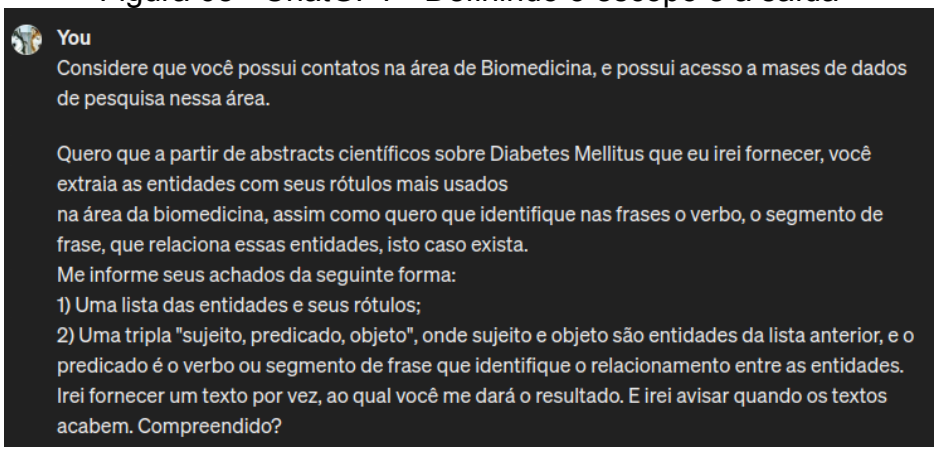
Figura 67 - ChatGPT - Etapa de definição de papel



Fonte: Elaborado pelo autor (2024)

Após a configuração inicial do papel do chat, a segunda etapa envolveu a definição clara da tarefa a ser realizada. Nesse estágio, foi estabelecido o escopo da atividade, especificando os tipos de dados a serem fornecidos (entradas) e os resultados esperados (saídas). Estabelecer um escopo bem definido é fundamental em qualquer processo, seja em um projeto, sistema ou pesquisa, pois promove a organização, sistematização e ordenação das atividades. Os detalhes dessa definição podem ser verificados na Figura 68.

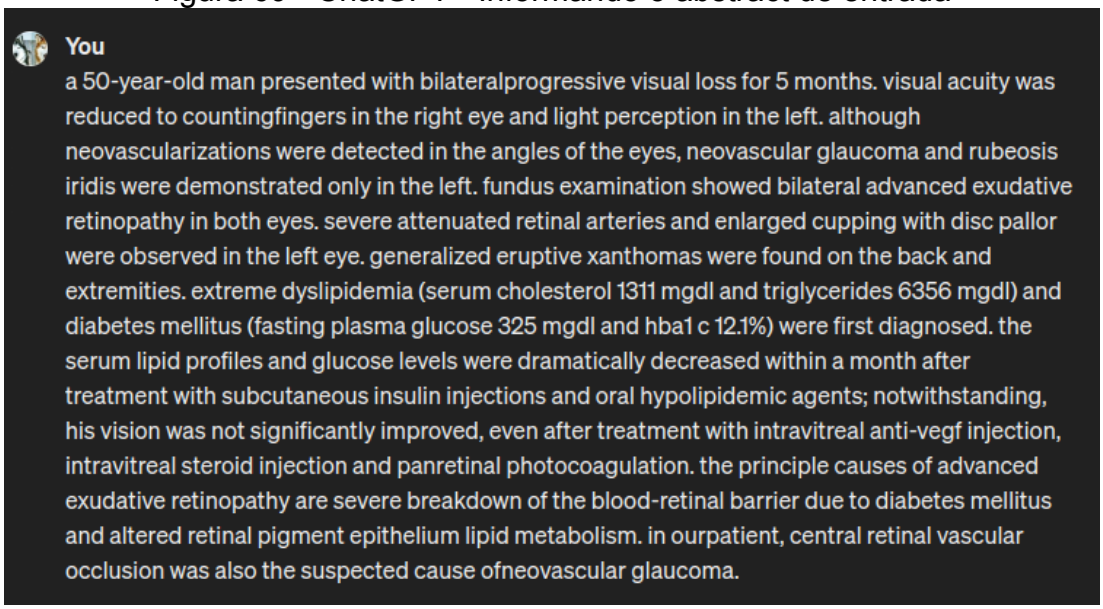
Figura 68 - ChatGPT - Definindo o escopo e a saída



Fonte: Elaborado pelo autor (2024)

Após definir o papel e o escopo das atividades, a terceira e última etapa envolveu o fornecimento dos dados necessários para os testes. Neste contexto, foi utilizado o resumo completo de um artigo científico, referenciado como A0. Esse procedimento permitiu uma comparação direta com os resultados obtidos nas fases anteriores dos experimentos. A Figura 69 ilustra o prompt utilizado para o envio desse texto ao sistema, destacando a estrutura e o conteúdo fornecido ao modelo para análise.

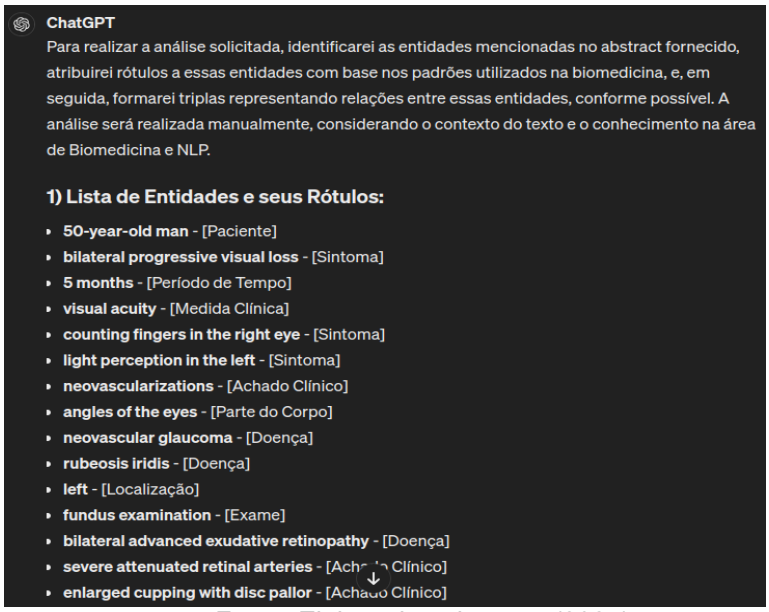
Figura 69 - ChatGPT - Informando o abstract de entrada



Fonte: Elaborado pelo autor (2024)

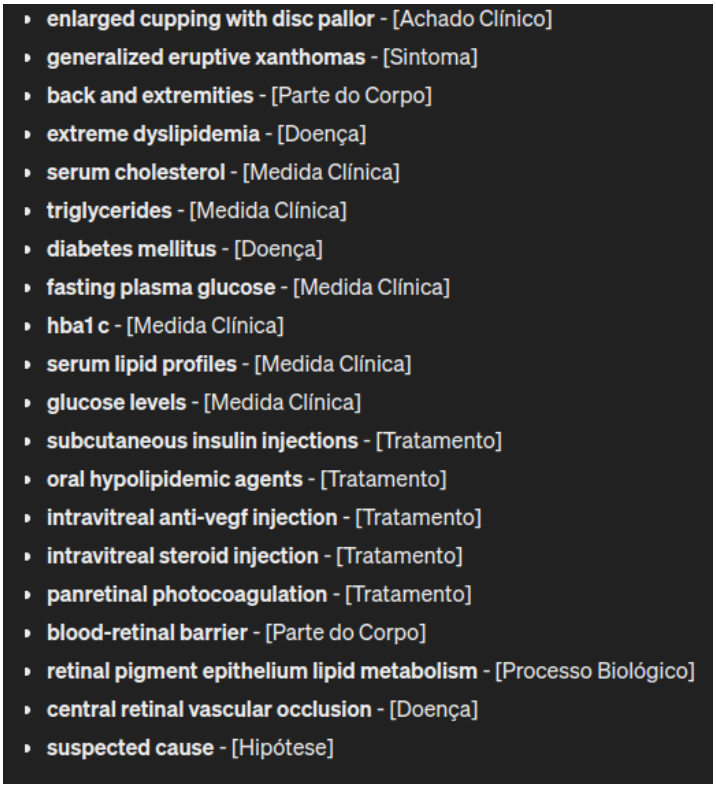
A resposta retornada foi extremamente concisa e atendeu às diretrizes especificadas no meu prompt, que definiu claramente a saída esperada. O sistema retornou dois conjuntos de dados. O primeiro conjunto é uma lista de entidades e suas respectivas classificações, conforme definido como “rótulo mais usado na área de Biomedicina”. O segundo conjunto consiste em uma lista de triplas “sujeito, predicado, objeto”, com a ressalva de que apenas valores encontrados deveriam ser apresentados. A lista de entidades e seus rótulos pode ser visualizada nas Figuras 70 e 71, enquanto a lista de triplas está apresentada na Figura 72.

Figura 70 - ChatGPT - Primeiras linhas da lista de entidades identificadas



Fonte: Elaborado pelo autor (2024)

Figura 71 - ChatGPT - Últimas linhas da lista de entidades identificadas

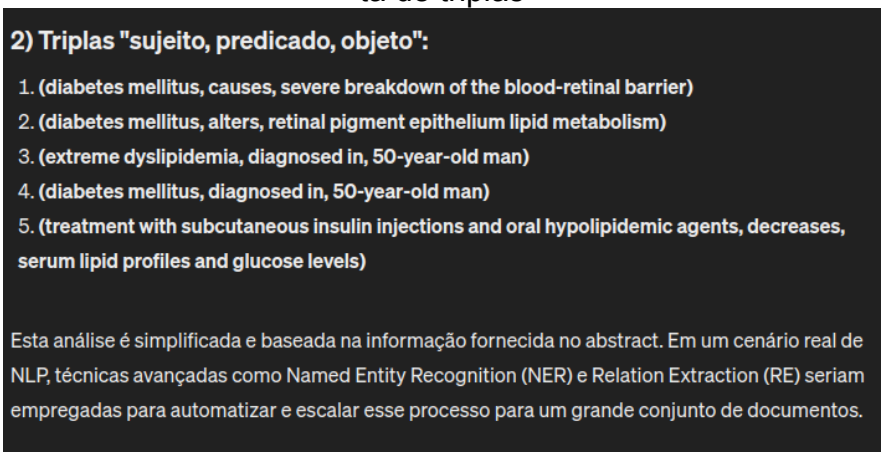


Fonte: Elaborado pelo autor (2024)

A análise da lista de entidades e rótulos revela uma notável diversidade, incluindo categorias como “Achado Clínico”, “Localização” e “Hipótese”. Embora

alguns desses rótulos tenham correspondentes nos modelos utilizados anteriormente, outros são exclusivos e não possuem equivalentes diretos. Diante dessa variedade, surgiu a ideia de utilizar o prompt para a rotulação de dados, com o objetivo de criar uma base de treinamento mais robusta e adaptada às necessidades específicas da pesquisa.

Figura 72 - ChatGPT - Parte final da resposta, contendo a lista de triplas



Fonte: Elaborado pelo autor (2024)

Ao observar-mos as triplas geradas revela-se que muitas das entidades identificadas não foram utilizadas na construção de triplas, evidenciando desafios na identificação do predicado que articula as relações entre entidades. De um total de 34 entidades detectadas, apenas 5 triplas foram efetivamente formadas. Esse fenômeno sugere deficiências no processo de captura das relações entre as entidades no texto examinado.

Por exemplo, na sentença “neovascularizations were detected in the angles of the eyes,” as entidades “neovascularizations” e “angles of the eyes” foram corretamente identificadas; a última entidade foi reconhecida não apenas como “eyes”, mas corretamente associada a “angles”. Contudo, o verbo “were detected” não foi identificado como o predicado que relaciona essas entidades. Embora “were detected” indique a ação de detecção e localização, essa relação pode não refletir a verdadeira conexão semântica entre “neovascularizations” e “angles of the eyes”, levantando questões sobre a validade desta como uma relação distinta em vez de uma única entidade complexa, como “neovascularizations on angles of the eyes”.

Esse exemplo ilustra a necessidade de uma modelagem cuidadosa dos dados, destacando a importância da colaboração entre engenheiros do conhecimento e especialistas do domínio, cuja presença é indispensável.

O exemplo também reflete o que ocorreu com outros casos, incluindo os experimentos manuais. Fica evidente a complexidade não apenas dos dados e das informações existentes nos textos, mas também da linguagem, onde nem sempre os relacionamentos entre entidades são claros e diretos.

6 CONCLUSÕES E TRABALHOS FUTUROS

O crescimento constante das publicações científicas reflete o avanço incessante das pesquisas e o compromisso da comunidade acadêmica com a inovação e a descoberta de novos conhecimentos. No entanto, esse aumento exponencial de documentos também apresenta desafios significativos na extração e no acesso ao conhecimento gerado. A quantidade maciça de informações, muitas vezes não estruturadas, dificulta a identificação e a análise de dados relevantes, exigindo o desenvolvimento de técnicas avançadas de mineração de dados e processamento de linguagem natural para transformar esse vasto repositório de textos em conhecimento acessível e útil.

A análise de palavras e documentos revelou duas questões centrais: a predominância do inglês nas publicações relevantes e a distribuição de palavras-chave relacionadas ao diabetes. Foi observado um aumento significativo nas publicações em inglês, evidenciando um crescente interesse na temática. Essa tendência sugere uma predominância do inglês como língua de pesquisa, acompanhada por um aumento das publicações em outros idiomas que seguem essa tendência.

A técnica de Semantic Relation mostrou-se promissora para enfrentar os desafios da variação terminológica em pesquisas sobre Diabetes Mellitus. Algoritmos de aprendizado de máquina permitiram representar e agrupar termos equivalentes, como “Diabetes Mellitus Type 2” e “Type 2 Diabetes”, melhorando a consistência e a clareza das informações, além de aprimorar a recuperação de dados e a extração de insights. Diferenciações criteriosas entre termos, como “Type 1 Diabetes” e “Type 2 Diabetes,” são cruciais, ressaltando a relevância de técnicas de pré-processamento adequadas.

Essas descobertas reforçam a necessidade de uma abordagem estruturada no tratamento de dados linguísticos em pesquisas biomédicas, enfatizando a importância da engenharia do conhecimento e da linguística computacional para a compreensão e o manejo de condições complexas como o diabetes mellitus. Essa etapa do projeto preparou o terreno para análises mais aprofundadas, revelando a interconexão entre linguagem, conhecimento e tecnologia, essencial para a eficiência da mineração de dados em contextos científicos.

A modelagem de tópicos no estudo do diabetes mellitus forneceu informações valiosas sobre a heterogeneidade e a complexidade do tema, sugerindo

a necessidade de explorar um maior número de tópicos para captar nuances mais profundas. Técnicas de modelagem como LDA e NMF mostraram diferenças significativas na distribuição de termos e na definição de tópicos. O LDA mostrou uma distribuição mais uniforme de palavras-chave, enquanto o NMF concentrou-se em menos palavras com maior intensidade.

O método BERTopic se destacou por sua capacidade de discernir tópicos com grande precisão, capturando melhor o contexto e nuances semânticas dos dados em comparação ao LDA. Essa técnica permite a criação de mapas de distância entre tópicos, facilitando a compreensão das relações e proximidade entre eles, o que é fundamental para a criação de um corpo de conhecimento mais sólido e detalhado.

O EDA mostrou-se essencial no estudo bibliométrico e cienciométrico, destacando o inglês como o idioma predominante na disseminação do conhecimento científico em biomedicina. Essa predominância reforça a decisão estratégica de se concentrar na análise de documentos em inglês, permitindo o acesso a uma variedade de pesquisas globais e a compreensão de tendências editoriais. A análise também revelou um aumento sustentado na produção científica, demonstrando o crescimento e diversificação na área biomédica.

A análise de NER em textos biomédicos sobre diabetes mellitus revelou a aplicabilidade de diversos modelos, destacando a variedade e profundidade na identificação de entidades biomédicas. Modelos genéricos apresentaram maior cobertura de entidades, enquanto modelos específicos forneceram informações detalhadas sobre categorias biomédicas. O modelo biomedical-ner-all mostrou-se robusto, identificando uma grande variedade de entidades médicas com precisão, reforçando a importância de escolher e configurar modelos de NER de acordo com os objetivos específicos de pesquisa biomédica.

A análise das funções SRL mostrou um bom desempenho na identificação e extração de verbos. No entanto, para aumentar a precisão e a abrangência da análise semântica, é sugerida a avaliação de outros modelos especializados em SRL, como variantes do BERT. Essa abordagem comparativa é crucial para aperfeiçoar as ferramentas de NLP usadas em contextos que requerem uma compreensão semântica aprofundada e precisão no tratamento de dados textuais.

A análise do Grafo de Conhecimento revelou dados relevantes sobre a estrutura e as conexões entre as entidades relacionadas à diabetes mellitus. A presença de arestas mais espessas no grafo indica uma maior frequência de certas

conexões, demonstrando a recorrência e a relevância desses relacionamentos. No entanto, erros ortográficos e agrupamentos extensivos de termos possivelmente contribuíram para a formação de noções distintas, afetando a coesão e a utilidade do grafo.

A análise quantitativa realizada com a biblioteca NetworkX permitiu uma compreensão mais aprofundada da topologia do grafo. A densidade do grafo é baixa, indicando uma conexão limitada, possivelmente devido à natureza especializada dos textos e às imprecisões nas designações das entidades. A fragmentação dos componentes conexos indica uma grande quantidade de informações com muitas entidades isoladas ou pouco conectadas. Para aperfeiçoar o grafo, recomenda-se a implementação de novas estratégias de normalização e padronização durante a extração de entidades e relações, além da expansão do corpus analisado.

A aplicação do ChatGPT-4 demonstrou as capacidades avançadas de Compreensão de Linguagem Natural (NLU) do modelo, alcançando os objetivos da pesquisa de forma eficiente e com menor esforço. As instruções claras e diretas reduziram os ruídos na interpretação dos comandos e simplificaram o processo de experimentação, gerando uma lista de entidades com rótulos relevantes à biomedicina e uma lista de triplas sujeito-predicado-objeto. No entanto, a formação limitada de triplas revelou dificuldades na captura das relações entre entidades, destacando a necessidade de uma modelagem de dados mais cuidadosa e a colaboração indispensável de especialistas no campo.

A pesquisa alcançou seus objetivos ao realizar com sucesso todas as etapas, desde a aquisição de dados até a geração e visualização de triplas como um KG, proporcionando descobertas e insights sobre o tratamento de dados, a modelagem do processo e o aprendizado sobre a área estudada. O valor do embasamento teórico, do DSRM e das investigações apoiadas por modelos específicos revelou-se fundamental.

Desafios significativos foram enfrentados durante o processamento dos dados, com análises que demandaram grandes quantidades de memória e tempo. No entanto, os resultados obtidos em cada fase do projeto foram inspiradores. Estou confiante de que o modelo desenvolvido é aplicável em ambientes produtivos, embora requeira ajustes e experimentações adicionais para ampliar sua eficácia.

O processo de melhoria é contínuo, e a geração de dados, conhecimento e publicações não cessa. Essa ferramenta será de grande utilidade para disseminar

conhecimento publicado e fomentar novas descobertas. A mineração de dados se mostra uma técnica indispensável em um mundo que demanda incessantemente por avanços na pesquisa.

6.1 LIMITAÇÕES

Durante a realização desta pesquisa, várias limitações foram identificadas, destacando-se a disponibilidade de poder computacional. Processar grandes quantidades de textos requer não apenas uma CPU de alto desempenho, mas também uma quantidade substancial de memória RAM. A criação e a análise de grafos de conhecimento, que envolvem a manipulação de grandes conjuntos de dados, demandam ainda mais recursos computacionais. Além disso, muitas técnicas avançadas de NLP e modelagem de tópicos se beneficiam significativamente do uso de GPUs para acelerar o processamento. A falta de acesso a hardware adequado pode, portanto, limitar a complexidade e a escala das análises possíveis, impactando a profundidade e a abrangência dos resultados obtidos.

Além das limitações computacionais, a pesquisa em NLP enfrenta desafios inerentes à complexidade da linguagem natural. Embora tenham sido feitos avanços significativos na área, NLP ainda não é um problema totalmente resolvido. As ambiguidades linguísticas, as variações terminológicas e as sutilezas contextuais presentes nos textos biomédicos, em particular, representam obstáculos significativos. Modelos de NLP, mesmo os mais avançados, podem apresentar dificuldades na compreensão e na interpretação precisa das relações semânticas. A necessidade de contínua pesquisa e desenvolvimento na área é evidente, com esforços voltados para aprimorar a precisão e a robustez dos modelos, visando superar as limitações atuais e melhorar a eficácia das técnicas de extração de conhecimento.

6.2 PERSPECTIVAS E TRABALHOS FUTUROS

As perspectivas para trabalhos futuros incluem um refinamento significativo do pré-processamento dos textos. Melhorias neste estágio inicial podem resultar em uma limpeza mais eficaz dos dados, eliminação de ambiguidades e correção de

erros ortográficos, garantindo que as entidades sejam reconhecidas de forma consistente. O uso de técnicas avançadas de normalização e padronização pode ajudar a consolidar variações terminológicas e melhorar a precisão na identificação de entidades. Além disso, a integração de métodos de desambiguação de entidades pode aumentar a coesão e a utilidade do grafo de conhecimento, permitindo uma representação mais fiel das relações presentes nos textos.

Melhorias no processo de NER e ER também são essenciais. A adoção de modelos mais sofisticados e específicos para o domínio biomédico pode aumentar a quantidade e a qualidade das entidades identificadas. Focar em técnicas que aprimorem a identificação de predicados que ligam sujeito e objeto é particularmente importante, uma vez que isso fortalece a construção das triplas semânticas. Investigações futuras podem explorar a combinação de diferentes modelos de NER e ER, bem como a utilização de técnicas de ensemble learning para otimizar o desempenho e a cobertura das entidades e relações extraídas. A colaboração com especialistas no domínio também pode ser valiosa para a validação e refinamento dos resultados obtidos.

Uma vez obtido um grafo de conhecimento robusto e bem estruturado, há um grande potencial para a aplicação de técnicas de Graph Neural Networks (GNN). Essas técnicas podem ser utilizadas para prever conexões não identificadas nos documentos originais, gerando novos insights e contribuindo para a expansão do conhecimento no campo da biomedicina. A capacidade das GNNs de capturar e modelar relações complexas entre entidades pode abrir novas possibilidades para a descoberta de interações e padrões ocultos, promovendo avanços significativos na pesquisa científica. Esse processo de geração de novo conhecimento a partir de dados existentes exemplifica a importância de uma abordagem integrada e contínua para a melhoria das técnicas de mineração de dados e análise semântica.

REFERÊNCIAS

ABU-SALIH, Bilal *et al*. **Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities**. *Journal of Big Data*, v. 10, n. 1, 28 maio 2023.

AGUSTIN , H.; ARIFIANTO; WINARTI, R. The relationship of the incident of diabetes distress and self care in diabetes mellitus patients in Semarang. **Ilmu dan Teknologi Kesehatan STIKES Widya Husada**., v. 15, n. 2, p. 50–54, 3 jul. 2024.

AKBIK, Alan; BERGMANN, Tanja; VOLLGRAF, Roland. Pooled Contextualized Embeddings for Named Entity Recognition. **Proceedings Of The 2019 Conference Of The North**, [S.L.], v. 1, p. 724-728, 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/n19-1078.

ALBERT, R.; BARABÁSI, A. L. **Statistical mechanics of complex networks**. *Reviews of Modern Physics*, v. 74, n. 1, p. 47-97, 2002.

ALLEN, B. P.; STORK, L.; GROTH, P. **Knowledge Engineering using Large Language Models**. 2023.

AMARATUNGA, Thimira. **Understanding Large Language Models: Learning Their Underlying Concepts and Technologies**. Nugegoda, Sri Lanka: Apress Media LLC, 2023. ISBN 979-8-8688-0016-0 (impresso); ISBN 979-8-8688-0017-7 (eletrônico). Disponível em: https://doi.org/10.1007/979-8-8688-0017-7.

AMERICAN DIABETES ASSOCIATION. **Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2019**. *Diabetes Care*, v. 42, supl. 1, p. S13–S28, 1 jan. 2019. Disponível em: https://doi.org/10.2337/dc19-S002. Acesso em: 18 de mai. 2021.

ANDROUTSOPOULOS, I.; PALIOURAS, G.; KARKALETSIS, V.; SAKKIS, G.; SPYROPOULOS, C. D.; STAMATOPOULOS, P. **Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach**. arXiv preprint cs/0009009, 2000.

BABIĆ, K.; MARTINČIĆ-IPŠIĆ, S.; MEŠTROVIĆ, A. **Survey of neural text representation** modelsInformation (Switzerland)Multidisciplinary Digital Publishing Institute, , 30 out. 2020. Disponível em: <https://www.mdpi.com/2078-2489/11/11/511/htm>. Acesso em: 12 mar. 2023

BACON, F. **Novum Organum**. Londres: [s.n.], 1620.

BARBER, D. **Bayesian reasoning and machine learning**. Cambridge ; New York: Cambridge University Press, 2012.

BARNES, J. **Aristotle: A Very Short Introduction**. Oxford: Oxford University Press, 1995.

BERISHA, B. SC Blend; MËZIU, B. SC Endrit. **Big Data Analytics in Cloud Computing: An overview**. Unpublished, 2021. Disponível em: http://dx.doi.org/10.13140/RG.2.2.26606.95048. Acesso em: 23 de jun. 2024.

BESTA, M. *et al*. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 38, n. 16, p. 17682–17690, 24 mar. 2024.

BIRUNDA, S. Selva; DEVI, R. Kanniga. A Review on Word Embedding Techniques for Text Classification. **Innovative Data Communication Technologies and Application**, [S.L.], p. 267-281, 2022. Springer Singapore. http://dx.doi.org/10.1007/978-981-15-9651-3_23.

BLEI, D.; NG, A.; JORDAN, M. **Latent Dirichl*et Al*location**. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003.

BOSSELUT, Antoine *et al*. **Comet: Commonsense transformers for automatic knowledge graph construction**. arXiv preprint arXiv:1906.05317, 2019.

BRACHMAN, R.; ANAND, T. The process of knowledge discovery in databases. In: FAYYAD, U. *et al*. (Eds.). ***Advances in knowledge discovery and data mining*. AAAI Press**, 1996. p. 37-57.

BRATANIC, T. **Graph algorithms for data science**. New York, NY: Manning Publications, 2024.

BRATE R, DANG M-H, HOPPE F *et al*. Improving language model predictions via prompts enriched with knowledge graphs. In: **CEUR Workshop Proceedings**. 2022. https://doi.org/10.5445/IR/1000151291.

BRUCHES, E. *et al*. Entity Recognition and Relation Extraction from Scientific and Technical Texts in Russian. **2020 Science and Artificial Intelligence conference (S.A.I.ence)**, 14 nov. 2020.

BUNGE, M. **Philosophy of Science: From Problem to Theory**. Transaction Publishers, 1998.

BUNGE, M. **La investigación científica: su estrategia y su filosofía**. México: Siglo XXI, 2004.

BUNGE, Mario. **The Scientific Approach**. In: **Philosophy of Science. Vol. 1, From Problem to Theory**. Revised ed. New York: Routledge, 1998. p. 3–50. ISBN 978-0-7658-0413-6.

BUNGE, Mario Augusto. **Philosophy of Science: From Problem to Theory**. Transaction Publishers, 1998. p. 24. ISBN 978-0-7658-0413-6.

CAMACHO-COLLADOS, J.; PILEHVAR, M. T. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. **Journal of Artificial Intelligence Research**, v. 63, p. 743–788, 6 dez. 2018.

CHANG, T. *et al*. Accelerating mixed methods research with natural language processing of big text data. ***Journal of Mixed Methods Research***, v. 15, n. 3, p. 398–412, 2021.

CHAO, M. H. *et al*. **Technology mining for intelligent chatbot development. Advances in Transdisciplinary Engineering. Anais**...IOS Press, 20 out. 2021 Disponível em: <https://ebooks.iospress.nl/doi/10.3233/ATDE210090>. Acesso em: 13 jun. 2023

CHEN, Xiaojun; JIA, Shengbin; XIANG, Yang. A review: knowledge reasoning over knowledge graph. **Expert Systems With Applications**, [S.L.], v. 141, p. 112948, mar.
2020. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2019.112948.

CHEN, Xieling *et al*. Topic analysis and development in knowledge graph research: a bibliometric review on three decades. **Neurocomputing**, [S.L.], v. 461, p. 497-515, out. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.neucom.2022.02.098.

CHEN, Ya *et al*. Collaborative filtering grounded on knowledge graphs. **Pattern Recognition Letters**, [S.L.], v. 151, p. 55-61, nov. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.patrec.2022.07.022.

COHEN, P. R. **Empirical Methods for Artificial Intelligence**. MIT Press, 1995.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to algorithms**. MIT Press, 2022.

CUPANI, Alberto. La peculiaridad del conocimiento tecnológico. **ScientiaeStudia**, São Paulo, v. 4, n. 3, p. 353-71, 2006.

DAHL, D. A. **Natural Language Understanding with Python: Combine natural language technology, deep learning, and large language models to create human-like comprehension**. Packt Publishing, 2023.

DEVARAKONDA, M. V.; RAJA, K.; XU, H. Named Entity Recognition. In: XU, H.; FUSHMAN, D. D. (Eds.). **Natural Language Processing in Biomedicine - a Practical Guide**. [s.l.] Springer International Publishing, 2024. p. 79–99.

DEVARAKONDA, M. V.; RAJA, K.; XU, H. Relation Extraction. In: XU, H.; FUSHMAN, D. D. (Eds.). **Natural Language Processing in Biomedicine - a Practical Guide**. [s.l.] Springer International Publishing, 2024b. p. 101–135.

DEVLIN, J. *et al*. **BERT: Pre-training of deep bidirectional transformers for language understanding** arXiv, 2018. Disponível em: <https://paperswithcode.com/paper/bert-pre-training-of-deep-bidirectional>. Acesso em: 16 mar. 2022

DIAZ GONZALEZ, A. D. *et al*. **Applying BioBERT to extract germline gene-disease associations for building a knowledge graph from the biomedical literature**. In: 2023 THE 7TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEM AND DATA MINING (ICISDM). Anais... New York, NY, USA: ACM, 2023.

EKMAN, M. **Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, NLP, and Transformers using TensorFlow**, 2021. Disponível em: <https://www.oreilly.com/library/view/learning-deep-learning/9780137470198/>. Acesso em: 18 dez. 2022

EMELE, Martin; DORNA, Michael. **Ambiguity preserving machine translation using packed representations**. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998. p. 365-371.

ESTEVA, A. *et al*. **Dermatologist-level classification of skin cancer with deep neural networks**. *Nature*, v. 542, n. 7639, p. 115-118, 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **Knowledge Discovery and Data Mining: Towards a Unifying Framework**. [s.l: s.n.]. Disponível em: https://cdn.aaai.org/KDD/1996/KDD96-014.pdf. Acesso em: 15 de ago. 2022.

FELDMAN, R.; DAGAN, I. **Knowledge discovery in textual databases (KDT)**. Disponível em: https://cdn.aaai.org/KDD/1995/KDD95-012.pdf. Acesso em: 3 abr. 2024.

FIRE, M.; GUESTRIN, C. **Over-optimization of academic publishing metrics: observing Goodhart's Law in action**. *GigaScience*, v. 8, n. 6, 2019.

FLOR, Luisa Sorio; CAMPOS, Monica Rodrigues. **Prevalência de diabetes mellitus e fatores associados na população adulta brasileira: evidências de um inquérito de base populacional**. *Revista Brasileira de Epidemiologia*, v. 20, p. 16-29, 2017.

FREITAS JUNIOR, Vanderlei *et al*. A pesquisa científica e tecnológica. **Espacios**, [s. l], v. 35, n. 9, p. 0-0, jul. 2014.

FREITAS JUNIOR, Vanderlei ; CECI, F. ; WOSZEZENKI, C. ; GONÇALVES, Alexandre L. **Design Science Research Methodology Enquanto Estratégia Metodológica para a Pesquisa Tecnológica**. Espacios (Caracas), v. 38, p. 25-34, 2017.

FRIEDMAN, C.; RINDFLESCH, T. C.; CORN, M. **Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine**. *Journal of Biomedical Informatics*, v. 46, n. 5, p. 765–773, 2013.

GERHARDT, T.; SILVEIRA, D. **Métodos de Pesquisa**. 1. ed. Porto Alegre: Editora da UFRGS, 2009.

GHASEMI, A. *et al*. **Scientific Publishing in Biomedicine: A Brief History of Scientific Journals**. *International Journal of Endocrinology and Metabolism*, v. 21, n. 1, 31 dez. 2022.

GLASGOW, B.; MANDELL, A.; BINNEY, D.; GHEMRI, L.; FISHER, D. **MITA: an information-extraction approach to the analysis of free-form text in life insurance applications**. AI Magazine, v. 19, n. 1, p. 59, 1998.

GOLDBERG, Y. **Neural Network Methods in Natural Language Processing**. [s.l.]: Morgan & Claypool Publishers, 2017.

GONÇALVES, Hortência de Abreu. **Manual de metodologia da pesquisa científica**. [s.l.]: São Paulo Avercamp, 2005.

GRAY, David E. **Pesquisa no mundo real**. Porto Alegre: Penso, 2012.

GREGÓRIO, J. *et al*. The role of Design Science Research Methodology in developing pharmacy eHealth services. **Research in Social and Administrative Pharmacy**, v. 17, n. 12, p. 2089–2096, 1 dez. 2021.

GRIFFITHS, T. L.; STEYVERS, M. **Finding scientific topics**. Proceedings of the National Academy of Sciences, v. 101, n. Suppl 1, p. 5228–5235, 2004.

GUNTURU, V. *et al*. Development of Language Model on Biomedical Domain to Pretrain Natural Language Processing. **2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)**, 9 maio 2024.

HAJ-YAHIA, Zied; SIEG, Adrien; DELERIS, Léa A. Towards Unsupervised Text Classification Leveraging Experts and Word Embeddings. **Proceedings of The 57Th Annual Meeting Of The Association For Computational Linguistics**, [S.L.], p. 0-0, 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/p19-1036.

HAN, J.; PEI, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2011.

HAREL, D.; FELDMAN, Y. **Algorithmics: The Spirit of Computing**. Addison-Wesley, 2004.

HAYES, P. J. **Intelligent high-volume text processing using shallow, domain-specific techniques**. In: TEXT-BASED INTELLIGENT SYSTEMS: CURRENT RESEARCH AND PRACTICE IN INFORMATION EXTRACTION AND RETRIEVAL. Anais... 1992. p. 227-242.

HEVNER *et al*. Design Science in Information Systems Research. **MIS Quarterly**, v. 28, n. 1, p. 75, 2004.

HEVNER, A. R. Design Research in Food Science: Keynote Introduction. **2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)**. Anais...IEEE, 1 abr. 2020Disponível em: <https://ieeexplore.ieee.org/document/9094108/>. Acesso em: 4 ago. 2022

HOLZINGER, A., Dehmer, M., & Jurisica, I. **Knowledge Discovery and Data Mining in Biomedical Informatics: The Future is in Integrative, Interactive Machine Learning Solutions**. In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics (pp. 1-18). Springer, Berlin, Heidelberg, 2014.

Horita, F. E. A., Neto, V. V. G., Santos, R. P. (2018). Design Science Research em Sistemas de Informação e Engenharia de Software: Conceitos, Aplicações e

Trabalhos Futuros. In: André Luiz Satoshi Kawamoto; Ana Grasielle Dionísio Corrêa; Valéria Farinazzo Martins. (Org.). **I Jornada Latino-Americana de Atualização em Informática**, p. 191-210.

IBM. **What is topic modeling?** Disponível em: https://www.ibm.com/topics/topic-modeling. Acesso em: 25 abr. 2024.
INTERNATIONAL DIABETES FEDERATION (IDF). **Diabetes Atlas**. 10th ed. Brussels, Belgium: IDF, 2021. Disponível em: http://www.diabetesatlas.org. Acesso em: 20 maio 2024.

INTERNATIONAL DIABETES FEDERATION (IDF). **Diabetes Atlas**. 9th ed. Brussels, Belgium: IDF, 2019. Disponível em: http://www.diabetesatlas.org. Acesso em: 1 fev. 2021.

JANSEN, Axel; ROESCH, Claudia. **Introduction: Biomedicine in Contemporary History**. Journal of Contemporary History, v. 57, n. 4, p. 843-858, set. 2022. DOI: 10.1177/00220094211039547.

JAYASEKARA, P. K.; ABU, K. S. **Text Mining of Highly Cited Publications in Data Mining**. In: INTERNATIONAL SYMPOSIUM ON EMERGING TRENDS AND TECHNOLOGIES IN LIBRARIES AND INFORMATION SERVICES (ETTLIS), 5., 2018. Anais [...]. [S.l: s.n.], 2018. DOI: 10.1109/ettlis.2018.8485261.

JI, Z. *et al*. Survey of Hallucination in Natural Language Generation. **ACM Computing Surveys**, v. 55, n. 12, 17 nov. 2022.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition**. Upper Saddle River, N.J: Pearson Education, 2008.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3rd ed. Pearson, 2020.

KALYAN, K. S.; SANGEETHA, S. **SECNLP: A survey of embeddings in clinical natural language processing**. Journal of Biomedical Informatics, v. 101, p. 103323, 1 jan. 2020.

KHARROUBI, A. T. **Diabetes mellitus: The epidemic of the century**. *World Journal of Diabetes*, v. 6, n. 6, p. 850, 2015.

KNIGHT, Kevin; LANGKILDE, Irene. **Using probabilistic models for generation in the Pangloss system**. In: Proceedings of the third conference on Applied Natural Language Processing. 2000. p. 170-179.

KOJIMA, Takeshi *et al*. **Large Language Models are Zero-Shot Reasoners**. ArXiv, [S.l.], v. abs/2205.11916, 2022. Disponível em: https://api.semanticscholar.org/CorpusID:249017743. Acesso em: 12 de Jun. de 2024.

KOTHARI, C. R. **Research Methodology: Methods and Techniques**. New Age International, 2004.

LAI, Po-Ting; LU, Zhiyong. **BERT-GT: Cross-sentence n-ary relation extraction with BERT and graph transformer**. arXiv preprint arXiv:2101.04158, 2021.

LAKATOS, E. M.; MARCONI, M. DE A. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2003.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. *Nature*, v. 521, n. 7553, p. 436-444, 2015.

LI, M.; YANG, H.; LIU, Y. **Biomedical named entity recognition based on fusion multi-features embedding**. *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine*, v. 31, p. 111–121, 2023.

LI, Xiaoxiao *et al*. **Explain Graph Neural Networks to Understand Weighted Graph Features in Node Classification**. In: INTERNATIONAL CROSS-DOMAIN CONFERENCE FOR MACHINE LEARNING AND KNOWLEDGE EXTRACTION. Springer, Cham, 2020. p. 57-76.

LI, S. *et al*. DeepPatent: patent classification with convolutional neural networks and word embedding. **Scientometrics**, v. 117, n. 2, p. 721–744, 1 nov. 2018

LIDDY, Elizabeth D. **Natural Language Processing: Overview**. In: Encyclopedia of Library and Information Science. New York: Marcel Decker, Inc., 2001. p. 212-220.

LIU, Y. **Python machine learning by example : implement machine learning algorithms and techniques to build intelligent systems**. Second ed. [s.l.] Birmingham Packt, 2019.

LOWY, Ilana. **Historiography of Biomedicine "Bio," "Medicine," and In Between**. Isis, v. 102, n. 1, p. 116-122, mar. 2011. DOI: 10.1086/658661.

LV, S. *et al*. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 05, p. 8449–8456, 3 abr. 2020.

MALTA, Deborah Carvalho *et al*. **Prevalência de diabetes mellitus determinada pela hemoglobina glicada na população adulta brasileira, Pesquisa Nacional de Saúde**. *Revista Brasileira de Epidemiologia*, v. 22, p. E190006. SUPL. 2, 2019.

MATURANA, H. R. **Biology of cognition**. *Biological Computer Laboratory Research Report BCL 9.0*, 1970.

MATURANA, H. R.; VARELA, F. J. **Autopoiesis and cognition: The realization of the living**. Springer, 1980.

MATURANA, Humberto R.; VARELA, Francisco J. **A árvore do conhecimento: as bases biológicas da compreensão humana**. Tradução de Humberto Mariotti e Lia Diskin. 9. ed. São Paulo: Palas Athena, 2011.

MATSUMOTO, N. *et al*. KRAGEN: a knowledge Graph-Enhanced RAG framework for biomedical problem solving using large language models. **Bioinformatics**, 3 jun. 2024.

MCCALLUM, A.; NIGAM, K. **A comparison of event models for naive bayes text classification**. In: AAAI-98 Workshop on Learning for Text Categorization. Anais... Menlo Park: AAAI, 1998. v. 752, p. 41-48.

MIKOLOV, T. *et al*. **Distributed Representations of Words and Phrases and their Compositionality**. Nips (2013), 1–9. DOI: h pDx. Doi. Org/10.1162/Jmlr. [s.l: s.n.]. Disponível em: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-com.pdf>. Acesso em: 21 mar. 2022.

MIKOLOV, T. *et al*. Efficient estimation of word representations in vector space. **1st International Conference on Learning Representations, ICLR 2013** - Workshop Track Proceedings. Anais...16 jan. 2013b Disponível em: <http://arxiv.org/abs/1301.3781>. Acesso em: 21 mar. 2022.

MORGAN, Gareth. **Paradigms, metaphors, and puzzle solving in organization theory**. *Administrative Science Quarterly*, v. 25, n. 4, p. 605-622, 1980.

NEEDHAM, M.; HODLER, A. E. **Graph Algorithms**. Sebastopol, CA: O'Reilly Media, Inc., 2019.

NIAN, Y. *et al*. **Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing**. *BMC Bioinformatics*, v. 23, n. S6, 2022.

NICHOLSON, D. N.; GREENE, C. S. **Constructing knowledge graphs and their biomedical applications**. *Computational and Structural Biotechnology Journal*, v. 18, p. 1414–1428, 2020.

NOVIKOV, A. M.; NOVIKOV, D. A. **Research Methodology: From Philosophy of Science to Research Design**. Boca Raton, Fl: CRC Press Taylor & Francis Group, 2013.

OLULEYE, A. **Exploratory Data Analysis with Python Cookbook**. [s.l.] Packt Publishing Ltd, 2023.

OMRANI, Pouria; HOSSEINI, Alireza; HOOSHANFAR, Kiana; EBRAHIMIAN, Zahra; TOOSI, Ramin; ALI AKHAEE, Mohammad. **Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement**. In: 2024 10th International Conference on Web Research (ICWR), 2024. p. 22-26. DOI: 10.1109/ICWR61162.2024.10533345.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Diabetes**. 2021. Disponível em: https://www.who.int/health-topics/diabetes. Acesso em: 01 de dez. 2021

PACHECO, R. C. D. S. *et al*. **Plataforma de Gestão Estratégica à Governança Pública em CT&I**. In: CONGRESSO ASSOCIAÇÃO BRASILEIRA DAS INSTITUIÇÕES DE PESQUISA TECNOLÓGICA E INOVAÇÃO, 2012.

PACHECO, Roberto Carlos dos S. **Dados e Governo Abertos na Sociedade do Conhecimento**. LOD BRASIL Linked Open Data Brasil. Florianópolis, 2014. Disponível em:

http://www.inf.ufsc.br/~jose.todesco/LODBrasil/Abertura/DadosEGovernoAbertoNaSocConh.pdf. Acesso em: 07 maio 2022.

PAN, Shirui; LUO, Linhao; WANG, Yufei; CHEN, Chen; WANG, Jiapu; WU, Xindong. **Unifying Large Language Models and Knowledge Graphs: A Roadmap**. IEEE Transactions on Knowledge and Data Engineering, v. 36, n. 7, p. 3580-3599, 2024. DOI: 10.1109/TKDE.2024.3352100.

PATTON, M. Q. **Qualitative Research & Evaluation Methods: Integrating Theory and Practice**. Sage Publications, 2015.
POPPER, K. **A lógica da pesquisa científica**. São Paulo: Cultrix, [s.d.].

PEFFERS, K. *et al*. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45–77, 1 dez. 2007.

PULUGU, D. *et al*. **Empowering healthcare with NLP-driven deep learning unveiling biomedical materials through text mining**. *The Scientific Temper*, v. 15, n. 02, p. 1966–1972, 15 jun. 2024.

PURI, A.; AGRAWAL, G.; DUKARE, A.; JAWALE, M. **A Survey and Analysis of Textual Content Based on Exploratory Data Analysis Technique and Opinion Analysis**. In: INTERNATIONAL CONFERENCE ON COMPUTATION, AUTOMATION AND KNOWLEDGE MANAGEMENT (ICCAKM), 4., 2023. Proceedings […]. [S. l.: s. n.], 2023. p. 1-6. DOI: 10.1109/ICCAKM58659.2023.10449608.

RAMOS, H. DE S. C.; BRÄSCHER, M. **Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T**. *Ciência da Informação*, v. 38, n. 2, p. 56–68, 2009.

RINALDI, F.; KALJURAND, K.; SÆTRE, R. Terminological resources for text mining over biomedical scientific literature. ***Artificial Intelligence in Medicine***, v. 52, n. 2, p. 107–114, 2011.

RIZKALLAH, Sandra; ATIYA, Amir F.; SHAHEEN, Samir. New Vector-Space Embeddings for Recommender Systems. **Applied Sciences**, [S.L.], v. 11, n. 14, p. 6477, 13 jul. 2022. MDPI AG. http://dx.doi.org/10.3390/app11146477.

ROSSANEZ, A. *et al*. **KGen: a knowledge graph generator from biomedical scientific literature**. *BMC Medical Informatics and Decision Making*, v. 20, n. S4, 2020.

ROSSANEZ, Anderson; DOS REIS, Julio. **Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases**. p. 12–23, 1 jan. 2019.

ROTHMAN, Denis. **Transformers for Natural Language Processing**. Packt Publishing, 2021. ISBN: 9781800565791.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: A modern approach**. 4. ed. Upper Saddle River, NJ, USA: Pearson, 2020.

SAHA, A. *et al*. BIOINTMED: integrated biomedical knowledge base with ontologies and clinical trials. ***Medical & Biological Engineering & Computing***, v. 58, n. 10, p. 2339–2354, 2020.

SAMVELYAN, A.; SHAPTALA, R.; KYSELOV, G. Exploratory data analysis of Kyiv city petitions. In: I**NTERNATIONAL CONFERENCE ON SYSTEM ANALYSIS & INTELLIGENT COMPUTING (SAIC)**, 2., 2020, Kyiv. Anais [...]. [S.l: s.n.], 2020. p. 1-4. DOI: 10.1109/SAIC51296.2020.9239185.

SARKER, I. H. **Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions**. SN Computer Science, v. 2, n. 6, p. 1–20, 18 ago. 2021.

SERLES, U.; FENSEL, D. Knowledge Graphs. In: **An Introduction to Knowledge Graphs**. [s.l.] Springer, Cham, 2024. p. 85–87.

SHEMTOV, Haim. Ambiguity management in natural language generation. In: **Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics**. 1997. p. 284-292.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. **Improving language understanding by generative pre-training**. 2018. Disponível em: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf. Acesso em: 23 jul. 2024.

RAO, Rama; TEJOMURTULA, GN. **Diabetes Mellitus: A Review**. International Journal of Medical Sciences & Pharma Research, v. 10, n. 2, p. 5-9, 2024. DOI: http://dx.doi.org/10.22270/ijmspr.v10i2.97.

RIZKALLAH, Sandra; ATIYA, Amir F.; SHAHEEN, Samir. New Vector-Space Embeddings for Recommender Systems. **Applied Sciences**, [S.L.], v. 11, n. 14, p. 6477, 13 jul. 2022. MDPI AG. http://dx.doi.org/10.3390/app11146477.

ROSPOCHER, Marco; *et al*. Cross-lingual Event Extraction for English, Dutch, and Italian Texts. In: **Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)**, 2012.

SRINIVASA-DESIKAN, B. **Natural language processing and computational linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras**. Birmingham, England: Packt Publishing, 2023.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. **Knowledge engineering: Principles and methods**. *Data & Knowledge Engineering*, v. 25, n. 1–2, p. 161–197, 1998.

TAVARES DE SOUZA, M.; DIAS DA SILVA, M.; DE CARVALHO, R. **Revisão integrativa: o que é e como fazer**. *Einstein*, v. 8, n. 1, p. 102–108, 2010.

TILLMANN, C.; VOGEL, S.; NEY, H.; ZUBIAGA, A.; SAWAF, H. **Accelerated DP based search for statistical translation**. In: EUROSPEECH. Anais... 1997.

URRESTARAZU, Hugo. **Autopoietic Systems: A Generalized Explanatory Approach - Part 1**. *Constructivist Foundations*, v. 6, p. 307-324, jul. 2011.

VASWANI, A. *et al*. **Attention is all you need**. arXiv preprint arXiv:1706.03762, 2017.

VENZIN, Markus; KROGH, George von; ROOS, Johan. **Future Research into Knowledge Management**. In: KROGH, George von; ROOS, Johan; KLEINE, Dirk (Eds.). *Knowing in Firms–Understanding, Managing and Measuring Knowledge*. London: SAGE Publications, 1998.

VYAS, R. *et al*. **Building and analysis of protein-protein interactions related to diabetes mellitus using support vector machine, biomedical text mining and network analysis**. *Computational Biology and Chemistry*, v. 65, p. 37-44, 2016. Disponível em: https://doi.org/10.1016/j.compbiolchem.2016.09.011. Acesso em: 27 de out. 2022

WALTON, Douglas N. **Argumentation Schemes for Presumptive Reasoning**. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.

WANG, Xiaoyang *et al*. Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. In: **AAAI Conference on Artificial Intelligence, 2018**. Disponível em: https://api.semanticscholar.org/CorpusID:52291548. Acesso em: 15 de jun. 2024.

WEI, Jason *et al*. **Chain of Thought Prompting Elicits Reasoning in Large Language Models**. ArXiv, [S.l.], v. abs/2201.11903, 2022. Disponível em: https://api.semanticscholar.org/CorpusID:246411621. Acesso em: 12 de Jun. de 2024.

WILLIAMS, L. *et al*. Topic Modelling: Going beyond Token Outputs. **Big data and cognitive computing**, v. 8, n. 5, p. 44–44, 25 abr. 2024.

YANG, Hao; ZHANG, Min; WEI, Daimeng. **IRAG: Iterative Retrieval Augmented Generation for SLU**. In: 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 2024. p. 30-34. DOI: 10.1109/CSPA60979.2024.10525270.

YANG, X. *et al*. **Mining a stroke knowledge graph from literature**. *BMC Bioinformatics*, v. 22, n. S10, 2021.

YE, Jiancheng *et al*. **Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes**. *BMC Medical Informatics and Decision Making*, 2020. Disponível em: https://dx.doi.org/10.1186/s12911-020-01318-4. Acesso em: 18 jul. 2023.

YU, Chuanming *et al*. Research on knowledge graph alignment model based on deep learning. **Expert Systems With Applications**, [S.L.], v. 186, p. 115768, dez. 2022. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2022.115768.

ZAJIC, D. M.; DORR, B. J.; LIN, J. **Single-document and multi-document summarization techniques for email threads using sentence compression**. Information Processing & Management, v. 44, n. 4, p. 1600-1610, 2008.

ZHANG, Fan *et al*. A Review of Knowledge Graph Technology in the field of Automatic Question Answering. **2020 International Signal Processing, Communications And Engineering Management Conference (Ispcem)**, [S.L.], p. 0-0, nov. 2020. IEEE. http://dx.doi.org/10.1109/ispcem52197.2020.00042.

ZHANG, Y.; CHEN, M.; LIU, L. **A review on text mining**. In: IEEE INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCE (ICSESS), 6., 2015. Anais [...]. [S.l: s.n.], 2015. DOI: 10.1109/icsess.2015.7339149.

ZHAO, W. X. *et al*. **A Survey of Large Language Models**. 31 mar. 2023.

ZHOU, J. *et al*. **Graph Neural Networks: A Review of Methods and Applications**. arXiv preprint arXiv:1812.08434, 2018.

APÊNDICE A – PROTOCOLO PARA REVISÃO INTEGRATIVA

A estrutura deste protocolo para revisão integrativa da literatura foi adaptado de Kitchenham (2007) e está organizado da seguinte forma:

1. Data: 05/06/2023
2. Fundamentos teóricos da pesquisa:
    1. A descrição da fundamentação teórica se encontra no Capítulo 2.
3. Questão de pesquisa:
    1. “Quais são os métodos e abordagens de Data Mining, aplicados a dados textuais não estruturados, para a construção de Knowledge Graphs?”
4. Bases de dados consultadas: liste as bases de dados que serão pesquisadas. Liste revistas ou websites que serão pesquisados:
    1. ResearchGate®
    2. Science Direct®
    3. PubMed®
    4. arXiv®
5. Critérios de inclusão e exclusão:
    1. Critérios de inclusão:
        1. Os termos de busca devem constar no título, resumo ou palavra-chave;
        2. Os documentos devem estar preferencialmente em inglês;
        3. Os documentos publicados devem respeitar o período de 2018 a 2023;
        4. Os documentos devem estar disponível para download.
    2. Critérios de exclusão:
        1. Os documentos com inconsistências como: falta de título, autor, resumo ou palavras-chave;
        2. Os documentos duplicados.
6. Estratégias de busca:
    1. Estratégias, *queries* de busca, distintas para as bases de dados conforme descrito na seção 3.3 Revisão Integrativa da Literatura.
7. Critérios de qualidade para seleção dos artigos:
    1. Serão considerados somente os artigos que estão de acordo com a busca realizada e em concordância com a questão de pesquisa.
8. Estratégias de extração dos dados, como os dados serão extraídos dos artigos:

    1. Será utilizada uma matriz de síntese onde, após a seleção dos artigos, será realizada a leitura dos títulos, resumos e palavras-chave de todas as publicações completas. Em seguida, os artigos selecionados passarão por uma nova seleção com base nos abstracts para depois serem relacionados na matriz para posterior análise e adequação aos critérios para inclusão definitiva na pesquisa.
9. Estratégias de análise dos dados:
    1. Os dados serão analisados para levantar as técnicas e métodos utilizados referentes a Data Mining em dados textuais não estruturados.
10. Estratégia de disseminação do conhecimento:
    1. O objetivo é gerar uma seção para a dissertação levando em consideração os artigos mais recentes na área sobre classificação de patentes.