CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

DETECÇÃO DE EVENTOS ADVERSOS EM MÍDIA SOCIAL: UMA REVISÃO DE LITERATURA E EXPERIMENTOS UTILIZANDO O METAMAP PARA O PORTUGUÊS DO BRASIL

Luiz Augusto de Souza Perciliano

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadora:
Kele Teixeira Belloze

Rio de Janeiro,

Janeiro de 2025

Detecção de Eventos Adversos em Mídia Social: uma revisão de literatura e experimentos utilizando o MetaMap para o Português do Brasil

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Luiz Augusto de Souza Perciliano

Aprovada por:

Presidente, Prof. Kele Teixeira Belloze, D.Sc. (orientadora)

Rafaelli de Carvalho Coutinho, D.Sc.

Jairo Francisco de Souza, D.Sc.

Rio de Janeiro,

Janeiro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

P429 Perciliano, Luiz Augusto de Souza
Detecção de eventos adversos em midia social: uma revisão de literatura e experimentos utilizando o metamap para o português do Brasil / Luiz Augusto de Souza Perciliano. — 2025.
132f. + apêndice : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 122-132
Orientador: Kele Teixeira Belloze

1. Processamento de linguagem natural (Computação). 2. Medicamentos - Utilização. 3. Medicamentos - Efeitos colaterais. 4. Twitter (Rede social on-line). I. Belloze, Kele Teixeira. (Orient.). II. Título.

CDD 006.312

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico este trabalho a Deus, pela força e sabedoria ao longo desta jornada, à minha esposa Luciana e à minha filha Rebeca, por serem minha inspiração diária, e à minha mãe, Maria Aparecida, pelo amor e valores transmitidos.

"Porque dele, por ele e para ele são todas as coisas. A ele seja a glória para sempre! Amém."

(Romanos 11:36)

AGRADECIMENTOS

A realização desta dissertação não seria possível sem a presença e o apoio de pessoas especiais em minha vida. Primeiramente, agradeço a Deus, pela força e sabedoria concedidas ao longo desta jornada.

À minha amada esposa, Luciana, por seu apoio incondicional, paciência e amor, que me inspiraram a perseverar mesmo nos momentos mais desafiadores. À minha filha, Rebeca, cuja presença iluminou meus dias e foi uma motivação constante para seguir em frente. À minha mãe, Maria Aparecida, pelo exemplo de determinação e dedicação, que sempre nortearam minha vida.

À minha orientadora, professora Kele Belloze, pelo seu conhecimento, orientação e incentivo ao longo de todo o processo. Aos excelentes professores do CEFET-RJ.

Agradeço também aos amigos de diferentes momentos da vida, que estiveram presentes e me apoiaram de maneiras diversas, em especial ao Antoine El Mallah com seu grande repertório a ajuda com o JAVA. A todos, deixo minha profunda gratidão por terem feito parte dessa caminhada, cada um contribuindo de forma única para a realização deste sonho.

RESUMO

Detecção de Eventos Adversos em Mídia Social: uma revisão de literatura e experimentos utilizando o MetaMap para o Português do Brasil

Luiz Augusto de Souza Perciliano

Orientadora:
Kele Teixeira Belloze

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Nos últimos anos, a área da saúde tem experimentado transformações tecnológicas substanciais, refletidas na expansão de bases de dados biomédicas em virtude de fatores como o aumento populacional, o surgimento de novas doenças e medicamentos. No entanto, no contexto da farmacovigilância, ainda há desafios significativos, como a complexidade e a limitação de acesso às bases de dados oficiais. Com o aumento da circulação global de pessoas e a crescente demanda por medicamentos, também cresce o número de Reações Adversas a Medicamentos (RAM). Apesar dos avanços, as autoridades responsáveis pela segurança dos medicamentos enfrentam dificuldades para registrar todos as RAMs reportadas, bem como para realizar buscas ativas que reduzam riscos para a população. A dificuldade em acessar e processar essas informações, somada à subnotificação de RAM, impulsiona a busca por novas fontes de dados. Devido às barreiras no acesso aos dados oficiais, as mídias sociais têm se mostrado uma alternativa relevante oferecendo informações valiosas para auxiliar na tomada de decisões, como: incluir notificações a fabricantes para ajustar bulas, recolhimento de lotes ou até mesmo suspensão de medicamentos. Diante deste cenário, este trabalho tem como objetivos: i) apresentar uma revisão bibliográfica sobre o tema; ii) explorar o Twitter para a detecção de RAMs; e iii) executar um experimento utilizando a ferramenta biomédica MetaMap, em conjunto com a base de dados UMLS adaptada para o português do Brasil. A pesquisa visa mapear termos e conceitos médicos presentes em tuítes, buscando identificar Sinais de Farmacovigilância (SF). O trabalho obteve sucesso na criação de um modelo UMLS em português do Brasil. Foram coletados e analisados milhares de tuítes sobre medicamentos para enxaqueca e sintomas de dengue. Por meio da ferramenta MetaMap, foi possível mapear os termos médicos presentes nos tuítes para conceitos da UMLS, evidenciado a viabilidade do uso da ferramenta para a detecção de (SF). Os resultados obtidos fornecem *insights* sobre o uso de medicamentos e a percepção da população sobre eles. O modelo UMLS e a metodologia desenvolvida neste trabalho abrem caminho para futuras pesquisas e aplicações de farmacovigilância com base em mídias sociais.

Palavras-chave:
Processamento de Linguagem Natural, Farmacovigilância, Reação Adversa a Medicamento, MetaMap, Twitter.

Rio de Janeiro,

Janeiro de 2025

ABSTRACT

Detecção de Eventos Adversos em Mídia Social: uma revisão de literatura e experimentos utilizando o MetaMap para o Português do Brasil

Luiz Augusto de Souza Perciliano

Advisor:
Kele Teixeira Belloze

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

In recent years, the healthcare sector has undergone substantial technological transformations, reflected in the expansion of biomedical databases due to factors such as population growth, the emergence of new diseases, and medications. However, in the context of pharmacovigilance, there are still significant challenges, such as the complexity and limited access to official databases. With the increase in global human mobility and the growing demand for medications, the number of Adverse Drug Reaction (ADR) also rises. Despite advancements, authorities responsible for medication safety struggle to record all reported (ADRs) as well as to conduct proactive searches to mitigate risks to the population. The difficulty in accessing and processing this information, coupled with under-reporting of (ADRs), drives the search for new data sources. Due to barriers in accessing official data, social media has proven to be a relevant alternative, offering valuable information to assist in decision-making, such as including notifications to manufacturers to adjust package inserts, recalling batches, or even suspending medications. In this scenario, this work aims to: i) present a literature review on the subject; ii) explore Twitter for the detection of (ADRs); and iii) conduct an experiment using the biomedical tool MetaMap, in conjunction with the UMLS database adapted for Brazilian Portuguese. The research seeks to map medical terms and concepts present in tweets to identify Pharmacovigilance Signal (PS). The work succeeded in creating a UMLS model in Brazilian Portuguese. Thousands of tweets about medications for migraines and dengue symptoms were collected and analyzed. Through the MetaMap tool, it was possible to map the medical terms present in the tweets to UMLS concepts, demonstrating the feasibility of using the tool for detecting (PS). The results provide insights into medication use and public perception of them. The UMLS model and methodology developed in this work pave the way for future research and applications in pharmacovigilance based on social media.

Key-words:
Natural Language Processing, Pharmacovigilance, Adverse Drug Reaction, MetaMap, Twitter

Rio de Janeiro,

Janeiro de 2025

Lista de Abreviações

ANVISA Agência Nacional de Vigilância Sanitária 120

API *Application Programming Interface* (Interface de Programação de Aplicativos) 81

ATC *Anatomical Therapeutic Chemical* (Classificação Anatômica, Terapêutica e Química) 74

BPC *Good Clinical Practices* (Boas Práticas Clínicas) 29

CED Critérios de Exclusão de Documentos 55

CEP Comitês de Ética em Pesquisa 29

CFQ *Council on Pharmacy and Chemistry* (Conselho de Farmácia e Química) . 30

CNS Conselho Nacional de Saúde 28

Conep Comissão Nacional de Ética em Pesquisa 29

CLAMP *Clinical Language Annotation, Modeling, and Processing Toolkit* (Kit de Ferramentas de Anotação, Modelagem e Processamento de Linguagem Clínica) 71

CRISP-DM *Cross-Industry Standard Process for Data Mining* 53

CSD Critérios de Seleção de Documentos 55

CNMM Centro Nacional de Monitorização de Medicamentos 31

cTAKES *clinical Text Analysis and Knowledge Extraction System* (Sistema de Extração de Conhecimento de Análise de Texto Clínico) 71

DATASUS Departamento de Informática do SUS 19

DFB *MetaMap Data File Builder* 84

DeCS Descritores em Ciências da Saúde 133

EA Evento Adverso 55

EHR *Electronic Health Record* (Registro Eletrônico em Saúde) 77

EMA *European Medicines Agency* (Agência Européia de Medicamentos) 19

EMEA *European Agency for the Evaluation of Medicinal Products* 28

EMR *Electronic Medical Record* (Registro Médico Eletrônico) 56

FDA *Food and Drug Administration* (Administração de Alimentos e Drogas) . . 19

GFARM Gerência de Farmacovigilância 33

ICH *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (Conselho Internacional para Harmonização de Requisitos Técnicos para Produtos Farmacêuticos para Uso Humano) . . . 135

ICSR *Individual Case Safety Reports* (Relatório de Segurança de Casos Individuais) 20

IR *Information Retrieval* (Recuperação de Informações) 71

IA Inteligência Artificial 79

KDD *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados (DCBD)) 35

KNIME *KNIME Analytics Platform* 24

LVG *Lexical Tools* (Gerador de Variante Lexical) 84

MaDIB Manutenção de Documentos de Bases Indexadas 53

MD *Data Mining* (Mineração de Dados) 53

MedDRA *Medical Dictionary for Regulatory Activities* (Dicionário Médico para Atividades Regulatórias) 74

MEDLINE *Medical Literature Analysis and Retrievel System Online* (Sistema Online de Análise e Recuperação de Literatura Médica) 73

MeSH *Medical Subject Heading* (Títulos de Assuntos Médicos) 97

ML *Machine Learning* (Aprendizado de Máquina) 73

MT *Text Mining* (Mineração de Texto) 77

MTI *Medical Text Indexer* (Indexador de Texto Médico) 39

NEME Núcleo de Estudos de Medicamentos 31

NERP-CRF NERP *Conditional Random Fields* 39

NLM *National Library of Medicine* (Biblioteca Nacional de Medicina) 71

OMS Organização Mundial da Sáude 106

PCLD Planilha de Controle de Leitura de Documentos 55

PIMM Programa Internacional de Monitorização de Medicamentos 20

PLN Processamento de Linguagem Natural 55

RAM Reação Adversa a Medicamento 54

REN Reconhecimento de Entidade Nomeada 73

SF Sinal de Farmacovigilância 61

SOBRAVIME Sociedade Brasileira de Vigilância de Medicamentos 31

SNOMED-CT *Systematized Nomenclature of Medicine-Clinical Terms* (Nomenclatura Sistematizada de Medicina - Termos Clínicos) 136

UMLS *Unified Medical Language System* (Sistema Unificado de Linguagem Médica) 56

Sumário

I Introdução **17**

I.1 Contextualização 18

I.2 Motivação 19

I.3 Objetivos 23

I.4 Método de Pesquisa 24

I.5 Organização 25

II Referencial Teórico **27**

II.1 Processo para Desenvolvimento de um Novo Fármaco 27

II.1.1 Descoberta de Novos Compostos 27

II.1.2 Estudo Pré-Clínico 28

II.1.3 Estudo Clínico 28

II.2 Um Breve Histórico da Farmacovigilância no Cenário Mundial 30

II.3 Reação Adversa a Medicamentos (RAM) 31

II.4 Evento Adverso (EA) 33

II.5 Sinal de Farmacovigilância (SF) 33

II.6 Metodologia CRISP-DM 34

II.7 Processamento de Linguagem Natural (PLN) 36

II.7.1 Coleta do *Corpus* 37

II.7.2 Nuvem de Palavras 38

II.7.3 Uniformização de Letras 38

II.7.4 Reconhecimento de Entidade Nomeada (REN) 39

II.7.5 Correção Textual 39

II.8 Ferramentas de PLN Aplicadas à Área Biomédica 39

II.8.1 MetaMap 39

II.8.2 MetaMap Data File Builder (DFB) 40

II.8.3 Sistema Unificado de Linguagem Médica (UMLS) 40

II.8.4 Modelo de Dados em Português do Brasil 49

II.8.5 Aplicação Java para Mapeamento e Serviços 50

II.9 Twitter como Fonte de Dados para Estudos de Farmacovigilância 51

III Método de Pesquisa Bibliográfica **53**

III.1 Análise e Processamento de Artigos Científicos 53

III.1.1 Manutenção de Documentos de Bases Indexadas (MaDIB) 54

III.1.2 Fase 1 - Entendimento do Negócio 54

III.1.3 Fase 2 e 3 - Entendimento e Preparação dos Dados 55

III.1.4 Fase 3 - Modelagem e Avaliação 56

III.1.5 Fase 4 - Implementação 57

III.1.6 Implementação da Pesquisa 61

III.2 Análise dos Trabalhos 61

III.2.1 Resultados Obtidos do Método de Pesquisa Bibliográfica 62

III.2.2 Panorama das Abordagens e Desafios em Farmacovigilância 71

IV MetaMap e o Processo de Anotação de Dados na Pesquisa Biomédica **81**

IV.1 Preparação dos Dados para Anotação 81

IV.1.1 Coleta e Seleção de Dados do Twitter 81

IV.1.2 Pré-processamento dos Textos 82

IV.1.3 Fluxo de Atividades do Mapeamento 82

IV.2 Configuração e Funcionamento do MetaMap 83

IV.2.1 Infraestrutura e Ambiente de Execução 83

IV.2.2 Instalação e Configuração Inicial da Ferramenta 84

IV.2.3 Adaptação do MetaMap para o Português do Brasil 84

IV.2.4 Configuração no Ambiente de Integração 87

IV.2.5 Limitações e Requisitos Específicos 87

IV.3 Processo de Anotação do MetaMap 88

IV.3.1 Etapas do Processo de Anotação 88

IV.4 Resultados da Anotação 89

IV.4.1 Armazenamento e Organização dos Resultados 89

IV.4.2 Detalhes do Mapeamento Biomédico 90

IV.4.3 Processamento Posterior no KNIME 91

IV.4.4 Considerações sobre os Resultados 91

V Experimentos **92**

V.1 Medicamentos da Classe Terapêutica de Enxaqueca 92

V.2 Dengue - Sinais e Sintomas 103

V.3 Modelagem e Inferência de Sinais de Farmacovigilância (SF) em Mídias Sociais . . . 111
V.3.1 Carga e Novo Pré-processamento do *Corpus* de Tuítes 111
V.3.2 Aplicação de Modelos 112
V.3.3 Inferência de Sinais e Sintomas (SF) 115

VI Conclusões e Trabalhos Futuros **119**

A Vocabulários Biomédicos **133**

B Processo de Mapeamento do MetaMap **138**

Lista de Figuras

Figura II.1 - Painel de Notificações de Farmacovigilância. 34
Figura II.2 - Fases do Ciclo da Metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) adaptado de [Chapman et al., 2000]. 36
Figura II.3 - Quantidade de Vocabulários por Língua. 41
Figura II.4 - Processo de Construção do *Unified Medical Language System* (Sistema Unificado de Linguagem Médica) (UMLS) *Metathesaurus* [Nguyen et al., 2021]. 42
Figura II.5 - Estrutura de Extração de Informações do *Metathesaurus* UMLS [Abbas et al., 2019]. 43
Figura II.6 - Quantidade de Tipos Semânticos por Grupo. 44
Figura II.7 - Uma Parte da Rede Semântica UMLS: Hierarquia de "Funções Biológicas" [National Library of Medicine (US), 2009a]. 45
Figura II.8 - Uma Parte da Rede Semântica UMLS: Relações [National Library of Medicine (US), 2009a]. 46

Figura III.1 - Diagrama de Atividades da Implementação do programa Manutenção de Documentos de Bases Indexadas (MaDIB). 57
Figura III.2 - Quantidade de Documentos por Níveis de Acessos. 63
Figura III.3 - Quantidade de Tipos de Documentos. 63
Figura III.4 - Veículo de Publicação. 64
Figura III.5 - Quantidade de Documentos por Ano de Publicação - Base Scopus. 65
Figura III.6 - Quantidade de Documentos por Ano de Publicação - Base PubMed. 65
Figura III.7 - Nuvem de Palavras nos Títulos dos Documentos - Base Scopus. 66
Figura III.8 - Nuvem de Palavras nos Títulos dos Documentos - Base PubMed. 67
Figura III.9 - Nuvem de Palavras nos Resumos dos Documentos - Base Scopus. 67
Figura III.10 - Nuvem de Palavras nos Títulos dos Documentos em Inglês. 68
Figura III.11 - Documentos por *Status*. 68
Figura III.12 - Correlação do Documento 70

Figura V.1 - Coleta e Processamento de Tuítes para Detecção de Sinal de Farmacovigilância (SF) 95

Figura V.2 - Formato Original do Tuíte na Plataforma Twitter. 96
Figura V.3 - Registros Mapeados por Vocabulário UMLS. 98
Figura V.4 - Tuítes Coletados X Mapeados por Grupo Semântico. 99
Figura V.5 - Nuvem de Palavras Coletadas. 103
Figura V.6 - Nuvem de Palavras Mapeadas Coletadas. 103
Figura V.7 - Painel Interativo: Coleta e Processamento de Tuítes sobre Dengue 105
Figura V.8 - Tuítes Coletados X Mapeados (Dengue). 106
Figura V.9 - Tuítes Coletados X Mapeados por Tipo Semântico (Dengue). 107
Figura V.10 - Nuvem de Palavras dos Termos Mais Referenciados da Coleta de Dengue. . 108
Figura V.11 - Nuvem de Palavras dos Termos Mais Mapeados. 109
Figura V.12 - Dengue: Formato Original do Tuíte na Plataforma Twitter. 110
Figura V.13 - Comparativo dos Corretores Ortográficos. 116
Figura V.14 - Normalização dos Textos. 116
Figura V.15 - Entidades Identificadas. 117
Figura V.16 - Dependências e Relacionamentos Identificados. 117

Lista de Tabelas

II.1 Diferenças entre Processamento de Linguagem Natural (PLN) e Linguística Computacional 37
II.2 Publicações no Twitter Acerca de EA. 52

IV.1 Descrição dos Campos Usados no Mapeamento do MetaMap. 90

V.1 Tuíte Mapeado. 97
V.2 Dez Tipos Semânticos com Mais Termos Mapeados. 100
V.3 Os Termos Mais Mapeados. 101
V.4 Tuítes Mapeados por Ano e Linguagem. 101
V.5 Os Termos Mais Referenciados na Coleta. 108
V.6 Os Termos Mais Mapeados. 109
V.7 Tuíte Mapeado referente a Figura V.12. 110
V.8 Resultado dos modelos *spaCy* aplicados no tuíte de Id Mapeado 29865. 118

Capítulo I Introdução

Em escala global, observa-se uma busca incessante por um bem-estar integral, sobretudo por meio da administração de medicamentos e suplementos vitamínicos. Esta busca alinha-se com a visão da Organização Mundial da Sáude (OMS), que define saúde como "um estado de pleno bem-estar físico, mental e social, e não meramente a ausência de doenças ou enfermidades" [Gaino et al., 2018].

Fundada em 7 de abril de 1948, a OMS consolidou-se como a principal entidade global especializada em saúde. Ela desempenha um papel vital ao normatizar e fornecer diretrizes para diversas agências reguladoras internacionais. Mais do que isso, a OMS está continuamente oferecendo recomendações e diretrizes focadas na preservação da saúde humana e na orientação farmacológica [Segre and Carvalho, 1997].

As agências reguladoras desempenham um papel fundamental na farmacovigilância, buscando constantemente entender e monitorar o Evento Adverso (EA) e a Reação Adversa a Medicamento (RAM). A farmacovigilância é a área que se atenta sobre o quanto os EAs estão relacionados ou são a causa do uso dos medicamentos. Além disso, o Sistema Nacional de Farmacovigilância (SINAF) no Brasil tem como objetivo aprimorar a captação e qualidade das notificações sobre esses eventos [Pepe and Novaes, 2020]. A Agência Nacional de Vigilância Sanitária (ANVISA) enfatiza que "a farmacovigilância visa garantir que os benefícios dos medicamentos superem os riscos"[1], refletindo a importância dessa vigilância contínua para a segurança dos pacientes.

A segurança farmacológica é uma preocupação, pois estimativas indicam sua presença entre as dez causas mais frequentes de óbito [Joshi et al., 2022]. Vale ressaltar que tais estimativas referem-se exclusivamente aos registros formais realizados por pacientes e profissionais da saúde. Uma RAM refere-se a um efeito colateral, não intencional e inesperado, decorrente da administração de uma dose padrão de um medicamento [Assurance and of Medicines Team, 2002]. A vasta maioria dos países estabeleceu sistemas de vigilância sanitária, nos quais são reportadas tanto as RAMs confirmadas quanto as suspeitas de EA. Um EA é caracterizado como qualquer evento médico adverso ocorrido durante o tratamento farmacológico, sem que haja, necessariamente, uma relação causal com o tratamento em questão [Assurance and of Medicines Team, 2002]. Comumente, esses registros são efetuados por médicos, podendo ser de natureza semi-estruturada ou em formato de

[1] `https://www.gov.br/anvisa/pt-br/assuntos/fiscalizacao-e-monitoramento/farmacovigilancia`

texto livre [Shah et al., 2012].

I.1 Contextualização

A elaboração de um medicamento inovador é um processo extenso, meticuloso e oneroso, permeado por riscos e incertezas. Antes de um medicamento ser disponibilizado ao mercado, é imperativo atender a um conjunto rigoroso de normas técnicas, éticas e regulatórias[2]. O ciclo de desenvolvimento de um medicamento, atualmente, pode se estender por aproximadamente uma década [Figueiredo et al., 2021].

Existem três etapas fundamentais para a disponibilização de um novo fármaco ao mercado:

- Descoberta de Novos Compostos;
- Estudo Pré-Clínico;
- Estudo Clínico.

A primeira etapa, Descoberta de Novos Compostos, objetiva identificar uma molécula, preferencialmente de origem natural, com potencial terapêutico para a condição alvo. A subsequente, o Estudo Pré-Clínico, avalia a eficácia da molécula em modelos animais [Nasciutti and Carvalho, 2012]. Caso as etapas anteriores sejam bem-sucedidas, o Estudo Clínico engloba experimentos rigorosos subdivididos em quatro fases. Após a conclusão bem-sucedida destas fases, o medicamento é disponibilizado ao mercado [Cardoso and Amorim, 2013].

As fases do Estudo Clínico são sequenciais. Na fase inicial, o fármaco é avaliado em indivíduos saudáveis, que não apresentam a condição alvo do tratamento. O foco primordial é garantir a segurança da dosagem proposta. Na segunda fase, o medicamento é administrado a pacientes portadores da condição alvo, com o intuito de avaliar sua eficácia terapêutica. Após a validação da eficácia e identificação das RAMs nas fases anteriores, o medicamento avança para a terceira fase, onde se busca a chancela das agências reguladoras para sua comercialização. Na quarta e última fase, a farmacovigilância assume um papel preponderante, monitorando a utilização do medicamento na população e identificando sua eficácia e potenciais EAs [Nasciutti and Carvalho, 2012].

Deste modo, a farmacovigilância é a disciplina encarregada de monitorar e analisar Sinais de Farmacovigilância (SF) associado aos medicamentos comercializados, buscando assegurar a segurança dos pacientes e a eficácia dos tratamentos [Joshi et al., 2022]. Um SF, frequentemente referido também como Sinal de Segurança, na área de farmacologia, é caracterizado como qualquer evidência ou dado que sugira a possibilidade de uma nova RAM ou EA [Meyboom et al.,

[2]https://www.gov.br/pt-br/servicos/obter-registro-de-medicamentos-genericos-similares-novos-e-inovadores

1997]. Logo, a atividade da farmacovigilância é fundamental não apenas para identificar possíveis riscos após a comercialização dos medicamentos, mas também para gerar dados que embasem ações corretivas ou preventivas. Quando identificado um SF, profissionais especializados, como médicos, farmacêuticos e outros profissionais da saúde, devem reportar tais ocorrências de forma sistemática, garantindo que a informação alcance os sistemas de monitoramento adequados.

Em geral, cada nação possui sua agência reguladora, que disponibiliza sistemas de notificação [Melo et al., 2021]. No Brasil, a ANVISA é encarregada da autorização e monitoramento de medicamentos, enquanto as Vigilâncias Sanitárias municipais reportam RAMs e EAs à ANVISA. Em âmbito internacional, a ANVISA reporta RAMs, EAs e demais informações pertinentes à OMS. Os sistemas de notificação disponíveis incluem VigiMed[3], Notivisa[4] e o e-SUS Notifica[5].

Historicamente, as notificações no Brasil eram realizadas manualmente. Em 2008, a ANVISA instituiu o sistema Notivisa, que requer cadastro prévio e possui restrições de acesso. O VigiMed, estabelecido em 2018, é acessível tanto a profissionais da saúde quanto ao público em geral, porém é restrito a notificações de EA e vacinas. Em 27 de março de 2020, o Ministério da Saúde, por meio do Departamento de Informática do SUS (DATASUS), lançou o e-SUS Notifica. A utilização frequente destes sistemas é essencial, visto que as informações são encaminhadas à OMS, que conduz análises e estudos sobre a eficácia dos medicamentos, podendo resultar na retirada de um medicamento do mercado em caso de numerosos relatos de RAMs [Cardoso and Amorim, 2013].

Embora haja um aumento nas notificações, este número ainda é considerado subestimado por especialistas e agências reguladoras, fenômeno conhecido como subnotificação [Melo et al., 2021]. Com a ascensão das mídias sociais, muitos indivíduos compartilham suas experiências terapêuticas nesses canais, seja pela simplicidade ou pelo desconhecimento de sistemas oficiais de notificação. Assim, estudos de farmacovigilância em mídias sociais têm identificado RAMs não registradas em sistemas oficiais [Duval and da Silva, 2019].

A análise desses relatos em mídias sociais é de suma importância. É relevante compreender as percepções e experiências dos indivíduos em relação aos medicamentos e seus EAs. Agências como a *European Medicines Agency* (Agência Européia de Medicamentos) (EMA) e a *Food and Drug Administration* (Administração de Alimentos e Drogas) (FDA) reconhecem as mídias sociais como uma fonte valiosa de dados para pesquisas em segurança farmacológica [Chen et al., 2018].

I.2 Motivação

Ampliar a conscientização e esclarecimento sobre a natureza e propósito das notificações faz-se necessário, considerando que estas são espontâneas e não compulsórias. A quantidade de relatos

[3] `https://www.gov.br/anvisa/pt-br/assuntos/fiscalizacao-e-monitoramento/notificacoes/vigimed/`
[4] `https://www8.anvisa.gov.br/notivisa/frmlogin.asp`
[5] `https://notifica.saude.gov.br/login`

e notificações de EA relacionados a medicamentos ou vacinas é insuficiente globalmente. Há uma expectativa de notificações por milhão de habitantes [Duval and da Silva, 2019]. Conforme os estudos de Maigetter et al. [2015], Vaz et al. [2016] e Avong et al. [2018], a OMS estipula que um centro nacional de farmacovigilância deve registrar, no mínimo, 200 notificações de suspeitas de RAM por milhão de habitantes anualmente.

No contexto brasileiro, as notificações de EA são escassas. A promoção à população sobre a capacidade de realizar notificações é limitada. Muitos profissionais de saúde, por desconhecimento ou receio, evitam esse processo, frequentemente motivados por temores relacionados à incerteza da causalidade entre o medicamento e o efeito colateral, ou por preocupações reputacionais [Cardoso and Amorim, 2013]. Este receio pode ser exacerbado pela possibilidade de serem contatados pela ANVISA ou outros órgãos de saúde. Tal fenômeno não é exclusivo do Brasil, sendo observado em diversas nações [Lima et al., 2020].

Notificações de EA benéficos também são importantes. Em algumas ocasiões, um medicamento pode beneficiar pacientes com condições diferentes daquelas para as quais foi originalmente desenvolvido. O uso de medicamentos para tratar condições diferentes das indicadas é denominado "uso *off label*". Um exemplo notório é o Viagra, inicialmente desenvolvido pela Pfizer para tratamentos cardiovasculares. Durante os ensaios clínicos, observou-se que os pacientes frequentemente experimentavam um EA específico – ereção peniana. Como resultado, o foco do medicamento foi reorientado. Embora o caso do Viagra tenha sido identificado durante ensaios clínicos, outros benefícios podem ser descobertos após a comercialização do medicamento. Isso reforça a importância das notificações, mesmo em situações onde os efeitos são benéficos ou quando o medicamento não produz o efeito esperado [Donato et al., 2013].

Embora não seja mandatório, é recomendável que os países encaminhem seus dados à OMS, que mantém a VigiBase[6] como o principal repositório global de *Individual Case Safety Reports* (Relatório de Segurança de Casos Individuais) (ICSR). Esta base de dados, estabelecida em 1968, contém mais de 30 milhões de notificações de suspeitas de RAMs, recebidas de países membros do Programa Internacional de Monitorização de Medicamentos (PIMM) da OMS para Monitoramento Internacional de Medicamentos [Melo et al., 2021].

Com a integração do Brasil ao VigiMed em 2018 e a incorporação da base médica *Medical Dictionary for Regulatory Activities* (Dicionário Médico para Atividades Regulatórias) (MedDRA), houve uma expansão do vocabulário em português do Brasil nesta base. Aproximadamente metade dos registros do VigiBase são originários dos EUA, e cerca de 20% da União Europeia (UE). O público pode acessar parcialmente esses dados por meio do VigiAccess[7].

A ANVISA, assim como outras autoridades sanitárias internacionais, faz um apelo contínuo

[6] `https://www.who-umc.org/vigibase/`
[7] `https://www.vigiaccess.org/`

para que profissionais de saúde e a população em geral realizem notificações. O objetivo é reduzir o número de mortes e hospitalizações decorrentes de RAM e EA. Quanto mais notificações forem feitas, mais ágil será a retirada de um medicamento ou lote potencialmente perigoso do mercado, minimizando danos à população e reduzindo custos hospitalares associados aos EAs [Cardoso and Amorim, 2013; Melo et al., 2021].

Os registros clínicos são notoriamente heterogêneos, variando em formatação, extensão e terminologia. Frequentemente, contêm abreviações, jargões, expressões coloquiais, erros ortográficos e termos específicos do contexto médico. No Brasil, devido à sua vasta extensão territorial e diversidade cultural, é comum que diferentes regiões utilizem terminologias distintas para se referir a uma mesma patologia ou conceito [Meystre et al., 2007].

Embora uma parcela significativa das informações clínicas dos pacientes esteja armazenada em formatos estruturados, uma vasta quantidade de dados clínicos permanece em formato narrativo, dispersa em diversos sites e plataformas de mídias sociais. Textos não estruturados proporcionam aos profissionais de saúde a liberdade de registrar variadas informações sobre os pacientes, incluindo históricos médicos, sintomas, diagnósticos, patologias, tratamentos e medicamentos, além de hipóteses, prescrições e recomendações [Luo et al., 2017].

Com a emergência da COVID-19 e a rápida disponibilização de vacinas, o Twitter (atualmente referido por X)[8] desempenhou um papel vital na disseminação e recepção de informações sobre campanhas de vacinação. Esta plataforma, devido à sua natureza concisa e imediata, foi amplamente utilizada tanto para a divulgação de informações por autoridades de saúde quanto para o compartilhamento de experiências pessoais relacionadas à vacinação [Li et al., 2018].

Textos oriundos de mídias sociais apresentam desafios significativos para análise e *Information Retrieval* (Recuperação de Informações) (IR) automatizadas destinadas a pesquisas clínicas. Diversas tarefas são necessárias para extrair informações dessas fontes, incluindo a normalização dos dados, anotação semântica de conceitos relevantes e classificação desses conceitos em categorias semânticas apropriadas [Chiaramello et al., 2016].

A anotação semântica exige dois componentes principais: uma ferramenta de Processamento de Linguagem Natural (PLN) para identificar sintagmas nominais e uma base de conhecimento para classificar os conceitos detectados nas categorias semânticas pertinentes. As técnicas de PLN são essenciais para identificar termos e frases em textos, e muitas dessas ferramentas foram inicialmente desenvolvidas para IR da literatura biomédica. Na literatura é possível observar trabalhos com adaptações dessas ferramentas para a análise de mídias sociais [Luo et al., 2017; Duval and da Silva, 2019]. Exemplos de ferramentas biomédicas que incorporam técnicas de PLN englobam: *clinical Text Analysis and Knowledge Extraction System* (Sistema de Extração de Conhecimento de Análise

[8] `https://help.twitter.com/pt/new-user-faq`

de Texto Clínico) (cTAKES)[9]; *Clinical Language Annotation, Modeling, and Processing Toolkit* (Kit de Ferramentas de Anotação, Modelagem e Processamento de Linguagem Clínica) (CLAMP)[10]; MagiCoder PRO[11]; MeaningCloud[12]; MedLEE; MetaMap[13] e SemRep[14].

Particularmente, a MetaMap é uma ferramenta de código aberto e gratuito que suporta diversos idiomas, incluindo o português. Ela utiliza o UMLS e é configurável para mapear texto biomédico, identificando conceitos em diferentes tipos de texto [Rogers et al., 2012]. Dada a escolha desta ferramenta para os experimentos deste trabalho, mais detalhes são apresentados na Subseção II.8.1. Diversos estudos têm explorado combinações dessas ferramentas [Bompelli et al., 2020].

Estas ferramentas operam com um ou mais vocabulários médicos. Um vocabulário médico é um conjunto padronizado de terminologia médica, projetado para facilitar o compartilhamento internacional de informações regulatórias sobre produtos médicos para humanos. Exemplos incluem o UMLS[15], MedDRA[16] e WHODrug[17] [Meystre and Haug, 2005].

Ao longo do tempo, RAMs e EAs têm sido objeto de estudo na farmacovigilância. Muitos desses estudos baseiam-se em estatísticas que não consideram o conhecimento consolidado na literatura biomédica. No entanto, a tendência atual é a crescente utilização de ferramentas de PLN voltadas para textos biomédicos [Shang et al., 2014; Golder et al., 2023; Sloane et al., 2015] e bases médicas com conceitos e semântica estabelecidos [Luo et al., 2017].

Nesse sentido, há estudos que detectam RAMs e EAs em mídias sociais utilizando ferramentas de PLN, como o trabalho de Lardon et al. [2015]; Duval and da Silva [2019]; Korkontzelos et al. [2016]. Outro exemplo é o estudo de Nguyen et al. [2017], que estimou a prevalência de RAMs em mídias sociais. Adicionalmente, De Rosa et al. [2021] exploraram a farmacovigilância na era das mídias sociais, descobrindo SF ao relacionar discussões no Twitter com textos do PubMed.

Novos algoritmos de IR têm sido desenvolvidos, focando na identificação de conceitos médicos em diversas línguas, além do inglês, como chinês [Zhang et al., 2018], espanhol [Oronoz et al., 2013; Campillos-Llanos, 2019; Duval and da Silva, 2019], italiano [Chiaramello et al., 2016], francês [Allam et al., 2021; Delbecque and Zweigenbaum, 2007; Deléger et al., 2010], português de Portugal [Ferreira et al., 2013], sueco [Skeppstedt et al., 2014] e japonês [Yokotsuka et al., 2000].

Compreender textos em diferentes línguas é desafiador. O português, por exemplo, é uma língua complexa e ambígua, repleta de regionalismos e expressões populares. É uma língua mais flexionada que o inglês e possui um léxico vasto, especialmente na terminologia médica. O estudo

[9] `https://ctakes.apache.org/`
[10] `https://clamp.uth.edu`
[11] `https://www.medbrains.it/en/`
[12] `https://www.meaningcloud.com/`
[13] `https://metamap.nlm.nih.gov/`
[14] `https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/SemRep.html`
[15] `https://www.nlm.nih.gov/research/umls/index.html`
[16] `https://www.meddra.org/`
[17] `https://who-umc.org/whodrug/`

de Campillos-Llanos [2019] também identificou desafios com o espanhol e trabalhou na expansão do léxico médico nesse idioma.

Muitas ferramentas biomédicas apresentam desafios de uso e interoperabilidade com diferentes vocabulários médicos. Algumas são mais específicas para determinados vocabulários, e a maioria desses vocabulários estão disponíveis principalmente em inglês. O MetaMap, apesar de sua complexidade de instalação e uso, é compatível com o UMLS, que suporta várias línguas, incluindo o português do Brasil.

Diversos estudos têm sido desenvolvidos para auxiliar profissionais da saúde e pacientes, com abordagens como: uso de *Machine Learning* (Aprendizado de Máquina) (ML) para reconhecimento de conceito e função de distância no banco de dados da *National Library of Medicine* (Biblioteca Nacional de Medicina) (NLM) usando MetaMap e cogência ontológica inversa [Shannon et al., 2021]; exames pré-clínicos [Jung et al., 2021]; detecção antecipada ou inferência de RAM [Guo et al., 2020]; PLN em vocabulários médicos: UMLS [Meystre and Haug, 2005] e MedDRA [Chen et al., 2018]; aprimoramento da alfabetização em saúde [Pradhan et al., 2014]; sistemas de apoio à decisão [Meystre and Haug, 2005; Wang et al., 2009; Iyer et al., 2013]; descoberta de correlações desconhecidas entre doenças [Hanauer et al., 2014; Jensen et al., 2012]; uso de ontologias para classificar termos biomédicos [Carrero et al., 2008] e PLN para coleta de texto em imagens [Ruiz, 2006].

Na revisão de literatura deste trabalho, não foi identificado nenhum trabalho sobre farmacovigilância voltado especificamente para o vocabulário brasileiro. Esta pesquisa explora essa lacuna, utilizando o MetaMap e o UMLS para coletar dados do Twitter a fim de realizar um mapeamento conceitual com um modelo UMLS em português do Brasil e, na sequencia, criar um experimento inicial para inferência de RAM. A escolha dessas ferramentas e bases de dados deve-se ao fato de serem de código aberto, consolidadas e amplamente reconhecidas na área biomédica. Conforme Shannon et al. [2021], o UMLS é uma base de dados anotada manualmente com conceitos rotulados, e em conjunto com o MetaMap, oferece resultados eficazes.

I.3 Objetivos

O objetivo deste trabalho é desenvolver um *pipeline* PLN adaptado ao português do Brasil, capaz de mapear conceitos biomédicos e detectar SF em mídias sociais. Os objetivos específicos incluem: (i) conduzir uma revisão da literatura sobre o tema; (ii) criar um modelo UMLS no idioma português do Brasil; (iii) compreender e integrar ferramentas biomédicas voltadas à farmacovigilância; (iv) utilizar dados do Twitter para realizar mapeamentos conceituais com o MetaMap; e (v) validar um experimento para identificar SF nos textos analisados.

Inicialmente, foi conduzida a revisão da literatura, com o objetivo de mapear estudos existentes

e compreender como as pesquisas nessa área têm sido realizadas em idiomas distintos do inglês. Essa análise visou identificar tendências e abordagens relevantes para subsidiar as formas de detecção de EAs e RAMs. Com base nessa revisão, criou-se o modelo UMLS. Este trabalho buscou adaptar a estrutura do UMLS para o contexto brasileiro, oferecendo suporte à área de farmacovigilância. Para validar o modelo, foram conduzidos experimentos de detecção de EA em textos extraídos do Twitter sobre medicamentos para enxaqueca e sintomas de dengue, utilizando o modelo criado.

O experimento proposto envolve a coleta, pré-processamento e análise de textos provenientes do Twitter, considerando o uso integrado de ferramentas biomédicas. Este estudo é inovador no Brasil, ao combinar mídias sociais com tecnologias de mapeamento conceitual biomédico, como o MetaMap, para identificar tendências relacionadas ao uso de medicamentos pela população.

I.4 Método de Pesquisa

O método de pesquisa proposto é estruturado em cinco etapas principais:

1. **Revisão da Literatura:** definição de *string* de busca para localizar documentos nas bases indexadas SCOPUS[18] e PubMed[19]. Foi desenvolvido um programa para facilitar o levantamento bibliográfico, gerenciar os resultados das buscas e definição dos trabalhos que fazem parte do escopo desta pesquisa.

2. **Coleta de Dados no Twitter:** foi criado um *pipeline* no Google Colab[20], onde dada uma *string* de busca, o sistema coleta tuítes, filtra, realiza um pré-processamento básico e salva os dados em arquivos (Excel e JSON).

3. **Pré-processamento de Dados:** utilização da ferramenta *KNIME Analytics Platform* (KNIME)[21], um sistema *open source* voltado para ciência de dados, que oferece facilidades como processos pré-definidos de PLN, integrações com diversos sistemas e bancos de dados. No KNIME são processadas diversas bases de dados inclusive os arquivos coletados da etapa anterior.

4. **Mapeamento Conceitual de Textos:** uso da ferramenta MetaMap em conjunto com o vocabulário médico UMLS e diversas outras ferramentas da NLM para mapear os textos pré-processados.

5. **Análise do Conteúdo:** avaliação quantitativa dos termos coletados e dos termos processados e análise das frases que indicam um possível SF conforme discutido por Duval and da Silva [2019].

[18] `https://www.scopus.com/`
[19] `https://pubmed.ncbi.nlm.nih.gov/`
[20] `https://colab.research.google.com/`
[21] `https://www.knime.com/`

Os detalhes específicos de cada etapa, bem como os experimentos realizados, são discutidos com mais profundidade nos Capítulos III e V respectivamente. É importante destacar as principais contribuições deste trabalho, que se configuram como um marco na área de farmacovigilância em português do Brasil.

1. **Pioneirismo no Idioma:** este estudo é pioneiro ao abordar a farmacovigilância utilizando ferramentas biomédicas, como o MetaMap, especificamente no contexto do português do Brasil, preenchendo uma lacuna significativa na literatura existente.

2. **Desenvolvimento de um Modelo UMLS:** a criação de um modelo UMLS adaptado para o português do Brasil representa uma inovação importante, permitindo o mapeamento de conceitos biomédicos de maneira mais eficaz e relevante para a população local.

3. **Aplicação de PLN:** a implementação de um *pipeline* de PLN para a detecção de SF em textos extraídos do Twitter demonstra a viabilidade do uso de mídias sociais como fonte de dados para monitoramento da segurança dos medicamentos.

4. **Revisão da Literatura:** uma revisão da literatura foi conduzida, proporcionando um panorama das abordagens internacionais em PLN para farmacovigilância em idiomas não-ingleses, contribuindo para o corpo de conhecimento na área e guiando a metodologia adotada.

5. **Apoio à Identificação de Tendências:** o trabalho contribui para a comunidade de interesse de saúde ao oferecer um arcabouço de possibilidades e soluções que podem auxiliar na identificação de tendências no uso de medicamentos pela população, promovendo uma vigilância mais ativa e informada.

I.5 Organização

Este documento está estruturado da seguinte forma, além desta introdução:

- **Capítulo II Referencial Teórico**: apresenta os conceitos fundamentais necessários para a compreensão deste trabalho;

- **Capítulo III Método de Pesquisa Bibliográfica**: detalha o método de pesquisa adotado para a revisão da literatura e composição dos documentos (artigos) analisados neste estudo e descreve análises de trabalhos relacionados ao tema de pesquisa;

- **Capítulo V Experimentos**: elucida um método proposto para identificação de RAM e EA e a utilização da ferramenta MetaMap adaptada para o português do Brasil, além de um processo inicial para modelagem e inferência de sinais de farmacovigilância;

- **Capítulo VI Conclusões e Trabalhos Futuros**: apresenta as considerações finais e direcionamentos para trabalhos futuros.

Capítulo II Referencial Teórico

Nesse capítulo, são apresentados alguns conceitos básicos sobre a criação de um fármaco, farmacovigilância, bases de dados médicas e ferramentas utilizadas nesse trabalho de pesquisa. O conceito computacional que norteia a pesquisa, Processamento de Linguagem Natural (PLN) é apresentado. Ainda, são descritos também a ferramenta de PLN da área biomédica - MetaMap e o vocabulário Sistema Unificado de Linguagem Médica (UMLS), importantes para o domínio em questão. E por fim, é apresentado o Twitter, mídia social de enfoque do trabalho para coleta de informações.

II.1 Processo para Desenvolvimento de um Novo Fármaco

O desenvolvimento de um novo medicamento tem um longo processo, inicia na pesquisa básica, passa pela descoberta do fármaco, tem os testes de desenvolvimento nas etapas clínicas e Estudo Pré-Clínico e, por fim, a submissão da pesquisa documentada com o novo medicamento para aprovação da agência reguladora [Klauck et al., 2022]. Aqui são detalhados apenas os três macros processos para o desenvolvimento de um novo fármaco: o primeiro, a descoberta de novos compostos que são as moléculas; o segundo é o Estudo Pré-Clínico onde já há testagem do produto e, por fim, a fase clínica. Esta última compreende quatro fases que avaliam a segurança do medicamento ao longo de um tempo pré-determinado até que na quarta fase, este possa ser disponibilizado para o mercado [Nasciutti and Carvalho, 2012; Cardoso and Amorim, 2013].

II.1.1 Descoberta de Novos Compostos

Este processo tem o objetivo de encontrar uma molécula com potencial para tratar doenças. Inicialmente procura-se alguma coisa similar na natureza, que possa ser convertido de maneira fácil na molécula idealizada, por uma síntese farmacêutica ou algo da própria natureza [Cardoso and Amorim, 2013; Barsi, 2005]. De cada 10.000 substâncias submetidas à pesquisa do Estudo Pré-Clínico, apenas uma segue para a pesquisa clínica com probabilidade de se transformar em um fármaco [Zucchetti and Morrone, 2012].

II.1.2 Estudo Pré-Clínico

Após a descoberta e seleção do novo composto candidato a um novo fármaco, este é cuidadosamente avaliado e monitorado nos testes do Estudo Pré-Clínico. Esta fase com testes primordialmente em apenas animais, objetiva principalmente a avaliação farmacológica em sistemas *in vitro*, *ex vivo* e *in vivo* [Zucchetti and Morrone, 2012]. O teste *in vitro* são ensaios biológicos conduzidos em sistemas não vivos. É realizado em ambiente controlado em um laboratório e geralmente feito em recipientes de vidro. Dentre os diversos propósitos destacados por Stefanelli [2011] são alguns como: cultura de células, de órgãos e de tecidos, podendo ser utilizados na genética, imunologia, produção de vacinas, pesquisa sobre vírus, dentre outros. Um exemplo é realizar cultura de células cancerígenas para identificar substâncias anticancerígenas. Com simulações computadorizadas e modelos matemáticos é possível analisar como aconteceria em um organismo vivo.

O teste *ex vivo* são ensaios conduzidos na extensão de um organismo em um ambiente externo com o mínimo de mudanças nas condições naturais. O trabalho de Monteiro [2012] mostra que este tipo de ensaio é geralmente empregado para testar a permeabilidade de um fármaco em um organismo animal. Usualmente são utilizados tecido intestinal de diferentes animais, o mais comum é o da rã, devido ser um alimento humano e não é preciso sacrificar o animal. Com este modelo de estudo é possível predizer a extensão da absorção intestinal. Com uma absorção de 90%, segundo Niazi [2019] um medicamento é considerado altamente permeável, contudo, a *European Agency for the Evaluation of Medicinal Products* (EMEA) diz que só é altamente permeável quando absorvido completamente.

O teste *in vivo* é utilizado em animais para identificar possíveis atividades farmacológicas, obter mais conhecimentos sobre as propriedades do fármaco e identificar Evento Adverso (EA) no fármaco em desenvolvimento [Nasciutti and Carvalho, 2012]. Os pesquisadores têm investido cada vez mais nos testes *ex vivo*, com o intuito de diminuir a exposição de voluntários sadios, os custos e tempo dispensado na criação de novos produtos farmacêuticos [Monteiro, 2012].

Nos ensaios do Estudo Pré-Clínico avalia-se tanto a toxicidade dos candidatos (animais) a fármacos à estabilidade do composto e possibilidade de produção em grande escala. Em todas as fases já são realizados acompanhamento dos testes clínicos, contudo, nesta fase a vigilância é intensificada. Existem diversas exigências legais da pesquisa pré-clínica conforme preconiza o Conselho Nacional de Saúde (CNS) de 1988[1] que também devem ser seguidas [Nasciutti and Carvalho, 2012].

II.1.3 Estudo Clínico

Após concluída a fase do Estudo Pré-Clínico com sucesso, inicia-se a Fase de Estudo Clínico. Essa também conhecida como Pesquisa Clínica, precisa de investimentos altos. A indústria far-

[1] `https://conselho.saude.gov.br/resolucoes/reso_88.htm`

macêutica predomina nesta fase e tem pressionado governos com as burocracias [Zucchetti and Morrone, 2012] e liberação de medicamentos de forma mais acelerada [Cardoso and Amorim, 2013]. Contudo, os experimentos devem atender às *Good Clinical Practices* (Boas Práticas Clínicas) (BPC) [Barsi, 2005].

As BPC foram publicadas em 2005 pela Organização Pan-Americana de Saúde, disponibilizando o Documento de Boas Práticas Clínicas das Américas[2]. Contudo, na *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (Conselho Internacional para Harmonização de Requisitos Técnicos para Produtos Farmacêuticos para Uso Humano) (ICH) em 1996, foi publicado padrões de BPC para União Europeia, Japão, EUA e Mercosul. As pesquisas clínicas realizadas no Brasil, além destas, também são submetidas às exigências da ANVISA[3], da Comissão Nacional de Ética em Pesquisa (Conep)[4] e dos Comitês de Ética em Pesquisa (CEP) locais [Zucchetti and Morrone, 2012].

O Estudo Clínico é dependente de todas estas entidades citadas até o momento e precisa cumprir suas respectivas exigências legais e éticas. Esta etapa possui quatro importantes fases, conduzidas de forma encadeada. Da primeira até a terceira é avaliada a segurança do medicamento. A quantidade de voluntários ao uso do medicamento aumenta quando se muda de fase. Na última fase, a avaliação é em escala global e o medicamento já se encontra disponível para o mercado em geral [Cardoso and Amorim, 2013].

Os estudos clínicos podem ser divididos em quatro fases principais. Na **fase 1**, o medicamento é testado em um pequeno grupo de 20 a 100 voluntários saudáveis, com o objetivo de avaliar a segurança, tolerância e posologia segura, além de investigar seu perfil farmacocinético e farmacodinâmico [Klauck et al., 2022]. A **fase 2** envolve até 300 voluntários com a condição ou doença para a qual o medicamento foi desenvolvido, focando na eficácia terapêutica e na relação dose-resposta, enquanto continua a monitorar a segurança [Klauck et al., 2022; Nasciutti and Carvalho, 2012].

Na **fase 3**, são incluídos até 3.000 voluntários doentes, com a finalidade de obter aprovação regulatória junto às autoridades, como a ANVISA no Brasil. Esta etapa avalia a segurança e eficácia a curto e longo prazo, incluindo possíveis RAMs associadas a fatores dos pacientes [Klauck et al., 2022; Zucchetti and Morrone, 2012]. Após a aprovação, ocorre a **fase 4**, que compreende estudos pós-comercialização realizados em condições reais de uso. Essa etapa é essencial para monitorar a segurança em larga escala, detectar EAs raros e investigar interações medicamentosas, além de englobar estudos de farmacovigilância, farmacoepidemiologia e bioequivalência [Zucchetti and Morrone, 2012; Nasciutti and Carvalho, 2012].

[2] `https://bvsms.saude.gov.br/bvs/publicacoes/boas_praticas_clinicas_opas.pdf`
[3] `https://www.gov.br/anvisa/pt-br`
[4] `http://conselho.saude.gov.br/comissoes-cns/conep/`

II.2 Um Breve Histórico da Farmacovigilância no Cenário Mundial

A farmacovigilância é considerada uma ciência farmacológica, relacionada com a coleta, identificação, análise, rastreamento e prevenção de EAs e RAMs que podem estar relacionados com: produtos fitoterápicos; medicamentos tradicionais e complementares; hemoterápicos; produtos biológicos; produtos para a saúde e vacinas [Barsi, 2005]. Além das RAMs, os seguintes assuntos são relevantes para a farmacovigilância: EAs causados por desvios da qualidade de medicamentos, inefetividade terapêutica, erros de medicação, uso de medicamentos para indicações não aprovadas no registro, uso abusivo, intoxicações e interações medicamentosas [ANVISA, 2022].

Seguem alguns eventos históricos que corroboraram para a necessidade de intervenções para garantir a qualidade e a segurança de medicamentos:

- Em 1906, nos Estados Unidos da América, foi criada a lei *"Pure Food and Drug Act"*, em resposta ao volume de falsificações e adulterações de medicamentos e alimentos [Brasil, 2022];

- Em 1937, mais de 100 pessoas faleceram por falência renal por consequência do uso de elixir de sulfanilamida dissolvida em dietilenoglicol. O FDA, agência reguladora dos EUA, passou a exigir testes de toxicidade e dados clínicos sobre segurança das drogas a serem comercializadas [Brasil, 2022; Rozenfeld and Rangel, 1988];

- Na década de 50, descobriu-se que o cloranfenicol poderia causar anemia aplástica. Então, o *Council on Pharmacy and Chemistry* (Conselho de Farmácia e Química) (CFQ) estabeleceu o primeiro registro oficial de RAM para detectar discrasias (alterações) sanguíneas induzidas por medicamentos [Brasil, 2022];

- Em 1962, o marco histórico da farmacovigilância mundial é conhecido como o "desastre da talidomida" no qual aconteceram milhares de casos de focomelia (rara malformação congênita) causados por seu uso para prevenção de enjoos em mulheres grávidas nas décadas de 50 e 60. Após a retirada do produto do mercado, em 1962, foram registrados aproximadamente 4.000 casos, dos quais, 498 mortes durante ou após o parto [Brasil, 2022; ANVISA, 2022; Rozenfeld and Rangel, 1988];

- Em 1963 é realizada a décima-sexta Assembleia Mundial da Saúde, na qual foi adotada uma resolução que reafirma a necessidade de ações imediatas em relação à disseminação rápida de informações sobre RAM [ANVISA, 2022];

- Em 1968, foi formado pela OMS, o PIMM para viabilizar a disseminação rápida de informações sobre RAM [Brasil, 2022];

- No final da década de 90 é criada a ANVISA, com a responsabilidade de estabelecer, coordenar e monitorar os sistemas de vigilância toxicológica e farmacológica, além de regular, controlar e fiscalizar medicamentos de uso humano que representam risco à saúde pública. O Brasil possui uma longa tradição na produção de medicamentos, que se iniciou com a fabricação artesanal de remédios a partir de extratos vegetais durante o período colonial. Após a Segunda Guerra Mundial, o setor evoluiu com a criação de indústrias farmacêuticas que produzem desde medicamentos à base de plantas até antibióticos e psicotrópicos em larga escala. Apesar do avanço na produção de medicamentos, a pesquisa farmacológica no Brasil permaneceu limitada até o surgimento de iniciativas como o Núcleo de Estudos de Medicamentos (NEME), criado em 1987, e a Sociedade Brasileira de Vigilância de Medicamentos (SOBRAVIME), instituída em 1990, que impulsionaram estudos sobre vigilância de medicamentos. Em 1998, o Programa Estadual de Redução de Iatrogenias foi implementado em São Paulo, tornando obrigatória a notificação de EAs e a elaboração de Planos de Farmacovigilância [ANVISA, 2022].;

- Em 2001, o Centro Nacional de Monitorização de Medicamentos (CNMM) foi aceito como 62º membro efetivo do PIMM da OMS [ANVISA, 2022].

- Em 2009, a ANVISA publicou a primeira norma específica de farmacovigilância destinada aos detentores de registro de medicamentos. Foram publicados os Guias de Farmacovigilância para Detentores de registro de Medicamentos. A partir de então, a notificação de EA pertinentes ao uso de medicamentos passou a ser compulsória a todos os detentores do registro de medicamentos em território nacional [ANVISA, 2022].

A partir do evento do "desastre da talidomida", houve uma mobilização mundial para vigiar os medicamentos de forma mais controlada e sistêmica [Brasil, 2022]. Uma revisão realizada nos estudos na fase de pré-comercialização do medicamento talidomida indicou que os dados haviam sido mal interpretados [ANVISA, 2022]. Grande parte dos países europeus aderiram esta mobilização, criando leis para medicamentos, a fim de garantir a segurança dos medicamentos e principalmente no Estudo Pré-Clínico baseados em animais. A OMS é a responsável por leis e monitoração de RAMs através dos sistemas de notificação [Rozenfeld and Rangel, 1988].

II.3 Reação Adversa a Medicamentos (RAM)

Uma Reação Adversa a Medicamentos (RAM), é um efeito colateral nocivo e não intencional inesperado causado em um ser humano ao tomar uma dose normal de um medicamento [Assurance and of Medicines Team, 2002]. Ainda que a natureza ou gravidade da RAM manifestada no paciente esteja na bula, ainda assim, é considerado uma RAM. Do mesmo modo, caso a reação não consista

com a rotulagem do medicamento ou a autorização dada por uma entidade, ou ainda, aquilo que não é esperado pelas características do medicamento, também é uma RAM.

A OMS desempenha um papel central no monitoramento das RAMs [Rozenfeld and Rangel, 1988], que são classificadas em leves, moderadas, graves e letais. As RAMs leves incluem sintomas como cefaleia, dores musculares e distúrbios digestivos, geralmente não exigindo a interrupção do medicamento. As moderadas podem apresentar distúrbios visuais, erupções cutâneas e alterações no humor, podendo requerer tratamento específico ou ajustes na dosagem, mas sem necessidade de suspensão do fármaco [Aguiar et al., 2015; Scripcaru, 2018].

As RAMs graves são raras e podem ser fatais, como reações alérgicas severas, insuficiências hepáticas e arritmias cardíacas, frequentemente necessitando da descontinuação imediata do medicamento. As reações letais resultam em óbito e podem levar à retirada do fármaco do mercado; exemplos incluem a `troglitazona, terfenadina e a talidomida` [Scripcaru, 2018; DeLucia et al., 2016].

A RAM é uma das principais preocupações com a segurança em medicamentos, as estimativas apontam que está entre as 10 principais causas de morte [Joshi et al., 2022]. Segundo a OMS [2005], as reações estão entre a quarta e sexta maior causa de mortalidade nos EUA. Esse número de mortes é relativo apenas aos relatos de pacientes e profissionais da área de saúde na forma oficial. Grande parte dos países possuem sistemas de vigilância sanitária para que nestes sejam relatados uma RAM ou até mesmo a suspeita de um EA. Na maioria das vezes estes relatos são realizados por médicos de forma semi-estruturada ou escrito em texto livre [Shah et al., 2012]. Alguns países têm acima de 10% de internação hospitalar devido às RAMs, a Noruega tem 11,5%, a França 13% e o Reino Unido com 16% [OMS, 2005].

A pesquisa de Costa et al. [2016] destaca a importância de aprimorar estratégias regulatórias e assistenciais para prevenir e reduzir os desfechos fatais relacionados a RAM. O estudo mostrou que, entre 2010 e 2017, ocorreram 650 notificações de RAM fatais no Brasil, predominantemente em mulheres, adultos e na região Sudeste. Os medicamentos mais envolvidos foram antimicrobianos, antineoplásicos e analgésicos.

Os achados revelam a necessidade de monitoramento prioritário de certas classes medicamentosas para detectar possíveis casos de superdosagem, que podem levar a RAM graves e fatais. O aprimoramento dos sistemas de farmacovigilância e a adoção de medidas regulatórias mais efetivas, somadas a estratégias assistenciais como protocolos de administração segura de medicamentos, podem contribuir para a prevenção e redução dessas RAMs potencialmente fatais. Dados como os apresentados no estudo são fundamentais para nortear as autoridades a implementar ou buscar protocolos mais eficientes para melhorar a segurança do uso de medicamentos [Costa et al., 2016].

II.4 Evento Adverso (EA)

A definição de EA é, qualquer ocorrência médica prejudicial ocorrida durante o tratamento com um medicamento em um usuário ou paciente. Contudo, não tem necessariamente uma relação causal com este tratamento. Um EA, também é chamado de experiência adversa [OMS, 2005].

Um EA, não está necessariamente associado ao uso de um medicamento. Ao ingerir um medicamento em dose normal e acontecer um evento inesperado como, por exemplo, muito sono, isso é uma RAM, mesmo que na bula fale dessa reação. O medicamento `Polaramine` mesmo na dose normal pode causar essa RAM, porém, se em uma crise, a dose ingerida for duplicada e acontecer um evento inesperado, isso não é RAM e sim uma EA causada por super dose.

Ao acessar o site de notificações de farmacovigilância, é exibido um painel sobre EA oriundas do sistema de notificação VigiMed, como mostra a Figura II.1. Os dados do painel referem-se as suspeitas de EAs, ou seja, apesar dos eventos terem sido observados após a administração de um medicamento ou vacina, não é possível ter confiança em sua relação, pois, cabe ao notificador o preenchimento dos dados e cabe somente a Gerência de Farmacovigilância (GFARM) os ajustes na notificação de medicamentos ou substâncias casos estas existam na base de dados doVigiMed.

O painel representado na Figura II.1 foi disponibilizado em 2018 e até 02/04/2023 possui 155.401 notificações, sendo o público feminino maior notificador, com 59%. Dentre os quatro tipos de notificadores, os serviços de saúde é o maior com 68% e empresas farmacêuticas com 49%. A faixa etária entre 18 a 44 é a maior notificadora com 49%.

O painel é navegável e pode ser ajustado com diversos tipos de filtros e visualizações. Pode ser visualizado desde notificações por ano, por EA, por nome do medicamento e também a combinação destes filtros, atualizando em tempo real os mapas e outros gráficos. A ANVISA mantém o painel[5] atualizado periodicamente.

II.5 Sinal de Farmacovigilância (SF)

Um Sinal de Farmacovigilância (SF), mais conhecido como Sinal de Segurança, em farmacologia, é definido como qualquer informação que sugira uma nova RAM ou um EA previamente conhecido. Esse SF deve estar associado a um medicamento, e que requer investigação adicional [Meyboom et al., 1997]. Essas informações podem ser obtidas de várias fontes, incluindo relatos de pacientes, profissionais de saúde, literatura científica e mídias sociais [Sarker et al., 2015].

O monitoramento de SF a partir de mídias sociais tem se tornado uma área de pesquisa crescente, especialmente devido à proliferação de informações sobre experiências individuais com medicamentos. O Processamento de Linguagem Natural (PLN) tem sido amplamente utilizado para

[5]Painel de Notificações da ANVISA

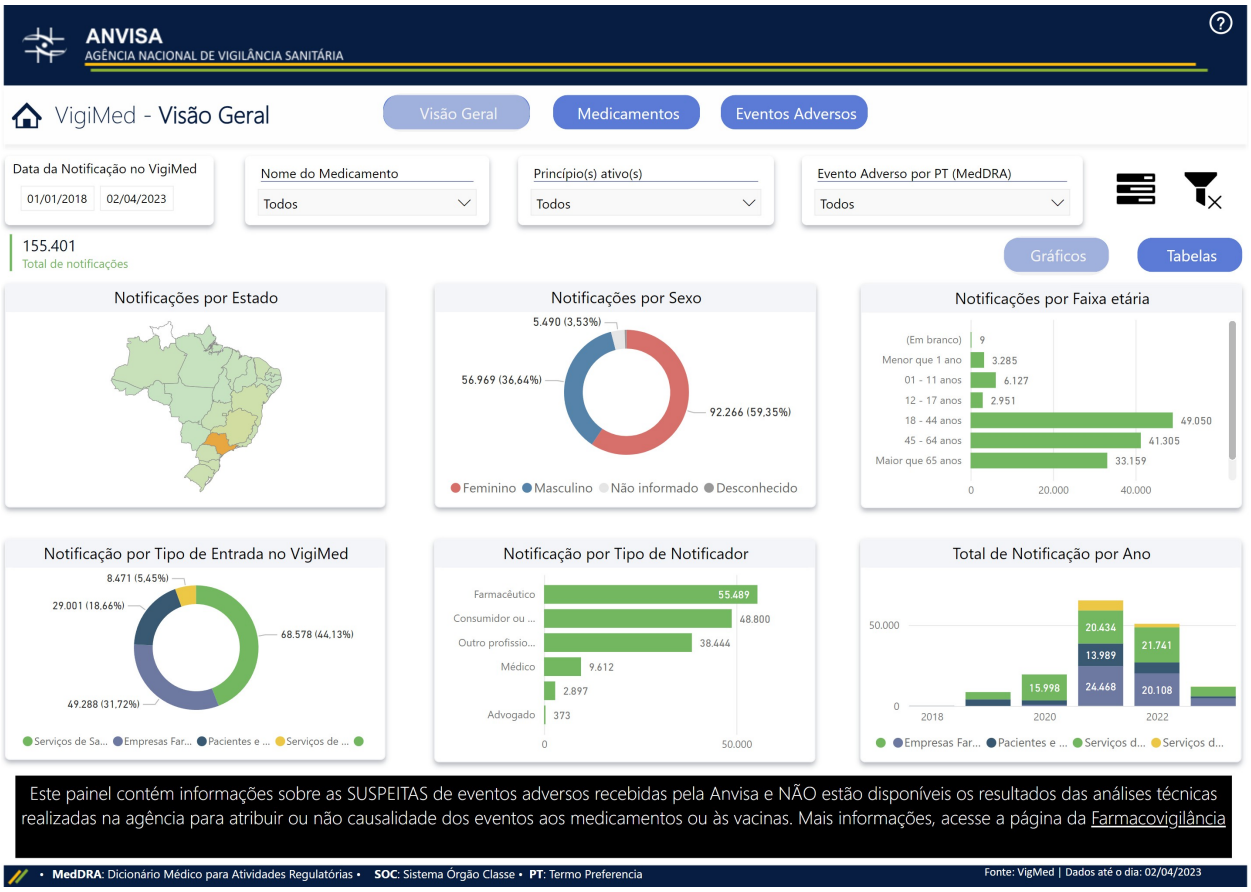


Figura II.1: Painel de Notificações de Farmacovigilância.

extrair e analisar esses dados de forma automatizada, permitindo a identificação de potenciais RAMs e EAs de forma mais ágil e abrangente [Pappa and Stergioulas, 2019].

A detecção de SF em textos provenientes de redes sociais como o Twitter, representa uma tarefa complexa e de grande relevância no âmbito do PLN, sobretudo no contexto da saúde pública [Bhattacharya et al., 2017]. Este desafio envolve não apenas o Reconhecimento de Entidade Nomeada (REN) e IR relevantes, mas também a interpretação e classificação adequadas desses dados em um espectro que vai além do textual.

No contexto da detecção de SFs, o Modelo Baseado em Regras e o Reconhecimento de Entidade Nomeada (REN) são essenciais. Esta primeira aplicação possibilita uma identificação precisa de EA, como evidenciado pelo trabalho de Mahendran and McInnes [2021]. Este estudo demonstra como o modelo extrai eficientemente EA de notas clínicas. O detalhamento destas abordagens se encontra na Subseção V.3.2.

II.6 Metodologia CRISP-DM

A metodologia CRISP-DM é um modelo de processo amplamente utilizado para *Data Mining* (Mineração de Dados) (MD). Ela fornece um guia abrangente para as fases deste processo, desde

a compreensão do negócio até a implementação dos resultados. A metodologia é composta por seis fases inter-relacionadas: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação [Chapman et al., 2000].

A metodologia CRISP-DM foi desenvolvida em meados da década de 1990 por um consórcio de empresas líderes em MD, incluindo a Daimler-Benz, Integral Solutions, NCR e OHRA. O objetivo era criar um processo genérico e aplicável em qualquer setor, de onde surgiu o termo *"Cross-Industry"* no nome. A Figura II.2 ilustra as fases do CRISP-DM, destacando suas interações e iterações. No centro estão os dados cercados pelas fases detalhadas nos itens abaixo:

1. **Entendimento do Negócio:** Esta fase inicial envolve a compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócio. Aqui, é fundamental transformar o conhecimento do negócio em um problema de MD, definindo um plano preliminar para alcançar os objetivos.

2. **Entendimento dos Dados:** Nesta fase, os dados são coletados e explorados. O objetivo é familiarizar-se com os dados, identificar problemas de qualidade, descobrir primeiros *insights* e formar hipóteses para testes posteriores.

3. **Preparação dos Dados:** Esta fase abrange todas as atividades necessárias para construir o conjunto de dados final a partir dos dados brutos. Tarefas como seleção de tabelas, limpeza de dados, construção de novos atributos e transformações são realizadas nesta fase.

4. **Modelagem:** Nesta fase, são aplicadas técnicas de modelagem com o objetivo de descobrir padrões nos dados. A seleção e aplicação de técnicas de modelagem adequadas dependem do problema de negócio e da natureza dos dados.

5. **Avaliação:** Os modelos são avaliados com respeito aos objetivos do negócio. Esta fase é crucial para determinar se o modelo atende aos critérios de sucesso do negócio e para decidir como proceder no projeto.

6. **Implementação:** A última fase do processo CRISP-DM envolve a implementação do modelo ou dos modelos escolhidos. Isso pode significar a realização de ações baseadas nas descobertas do modelo ou a integração do modelo em sistemas de tomada de decisão.

Além do CRISP-DM, existem diversas outras metodologias empregadas em projetos de MD, como, por exemplo, o SEMMA (*Sample, Explore, Modify, Model, Assess*) e o *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados (DCBD)) (KDD). O SEMMA foi desenvolvida pelo *SAS Institute* e se destaca por sua abordagem estruturada e sequencial, o trabalho de Azevedo and Santos [2008] faz uma comparação entre o KDD, SEMMA e faz um bom

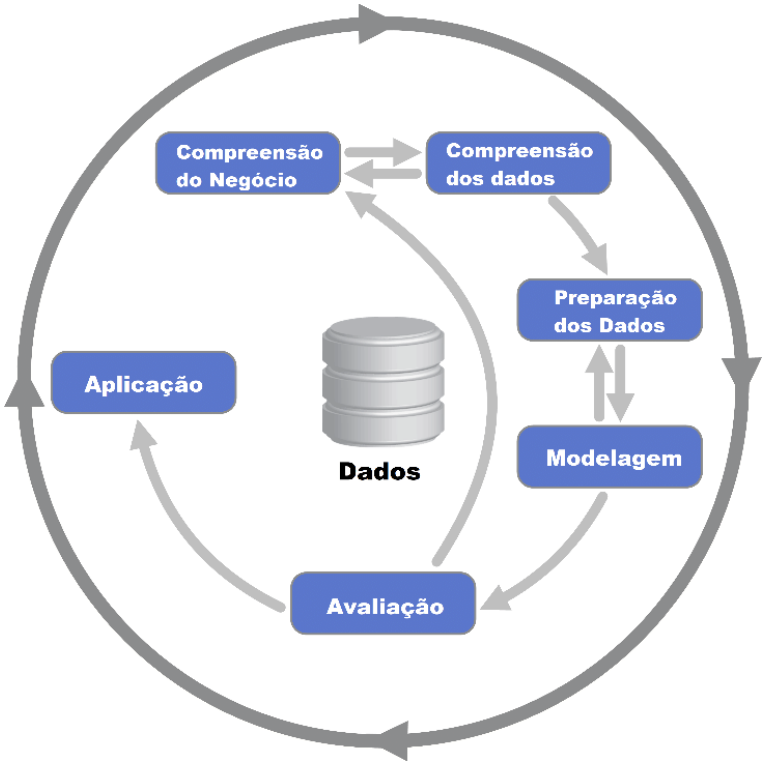


Figura II.2: Fases do Ciclo da Metodologia CRISP-DM adaptado de [Chapman et al., 2000].

destaque para o CRISP-DM. Dada a ampla aceitação e utilização do CRISP-DM tanto no meio acadêmico quanto no empresarial, esta metodologia é adotada na presente pesquisa [Piatetsky, 2014]. A aplicação do CRISP-DM se deu na mineração de tuítes e de trabalhos pesquisados nas bases de dados Scopus e PubMed.

II.7 Processamento de Linguagem Natural (PLN)

O PLN é uma subárea da Ciência da Computação e da Linguística que se dedica a explorar como as máquinas podem ser utilizadas para compreender e manipular textos ou discursos em linguagem natural. Os pesquisadores em PLN buscam entender como os seres humanos utilizam e compreendem a linguagem, com o objetivo de desenvolver técnicas e ferramentas que permitam aos computadores processar idiomas de forma natural [Chowdhury, 2005]. O intuito é extrair informações relevantes e realizar tarefas pré-determinadas.

O PLN e a Linguística Computacional diferem em seus objetivos e abordagens. O PLN utiliza métodos baseados em aprendizado de máquina e modelos estatísticos, enquanto a Linguística Computacional se concentra em modelagem formal e análise linguística. Apesar das diferenças mostradas na Tabela II.1, ambas as áreas se complementam, sendo a Linguística Computacional essencial para a base teórica do PLN [Opitz et al., 2024; Di Felippo and Dias-da Silva, 2009].

Aspecto	Processamento de Linguagem Natural (PLN)	Linguística Computacional
Objetivo Principal	Resolver problemas práticos de interação entre humanos e máquinas.	Explorar e entender a linguagem natural de forma teórica e computacional.
Abordagem	Centrada em algoritmos, modelos estatísticos e aprendizado de máquina.	Baseada em linguística teórica, modelagem formal e análise de linguagem.
Natureza	Foco aplicado, visando a criação de sistemas utilizáveis.	Foco científico, com ênfase na pesquisa e análise linguística.
Exemplo de Ferramentas	TensorFlow, spaCy, NLTK, Hugging Face.	Modelos gramaticais, corpus linguísticos, software de análise sintática.
Disciplinas Relacionadas	IA, Estatística e Ciência da Computação.	Linguística, Semiótica e Computação.

Tabela II.1: Diferenças entre PLN e Linguística Computacional

A essência do PLN perpassa diversas disciplinas, como Ciências da Informação, Linguística, Matemática, Inteligência Natural e Psicologia, entre outras. As aplicações do PLN abrangem uma ampla gama de pesquisas e tipos de *Text Mining* (Mineração de Texto) (MT), incluindo: análise de sentimentos, *chatbots*, classificação e clusterização de documentos, modelagem de tópicos, reconhecimento de fala, IR multilíngue e entre idiomas, similaridades de documentos, sistemas de perguntas e respostas, sistemas especialistas, sumarização de texto e tradução automática de texto [Chowdhury, 2005; Liddy, 2001].

Quatro etapas de análises abrangem o PLN: morfológica, sintática, semântica e pragmática, ambas realizadas nesta ordem. A morfológica é encarregada em definir artigos, adjetivos, substantivos e verbos em uma espécie de dicionário. Após a construção de um dicionário, a análise sintática se utiliza deste, mostrando o relacionamento entre as palavras e em seguida examina o sujeito, o predicado, apostos, adjuntos e complementos verbais e nominais. Na sequência, na semântica, é descoberto os termos ambíguos, os afixos e sufixos. Por fim, para entendimento das etapas anteriores, a pragmática, mostra visualmente todos os resultados. Para estas análises, há algoritmos apresentando em forma de árvore o texto, mostrando todas as etapas até o final do processamento [Santos et al., 2015; Liddy, 2001]. A seguir são apresentadas as abordagens de PLN utilizados neste trabalho.

II.7.1 Coleta do *Corpus*

Para iniciar as tarefas de MT, se faz necessário criar primeiramente um processo para coletar e armazenar os dados que serão trabalhados. Esta base é o *corpus* do projeto em análise. Um *corpus*

é composto de inúmeros documentos [Christian and Emmanuel, 2009]. Segundo Silva et al. [2021], cada projeto tem os seus corpora desenvolvido, bem definido, alinhado e escrito conforme o projeto proposto.

II.7.2 Nuvem de Palavras

A nuvem de palavras, também conhecida por *"word cloud"* é uma representação visual do conteúdo textual, na qual as palavras são exibidas com tamanhos proporcionais à sua frequência no texto analisado. Essa técnica permite identificar rapidamente os termos mais proeminentes e relevantes em um conjunto de dados textual, facilitando a análise exploratória e obtenção de *insights* [Grus, 2016].

Ao gerar uma nuvem de palavras, as palavras que ocorrem com maior frequência são plotadas em tamanho maior e de forma mais destacada. Já termos menos frequentes aparecem menor. Dessa forma, obtém-se de maneira rápida e intuitiva uma noção dos tópicos e conceitos dominantes contidos em um determinado texto ou um *corpus*. Estas são utilizadas para analisar tendências, ênfases e padrões linguísticos em conteúdos textuais das mais diversas naturezas, incluindo transcrições de discursos, entrevistas, relatórios, tuítes, respostas de questionários e outros. Por prover uma representação visual das palavras-chave, essa técnica se aplica a uma grande variedade de casos de uso, como análise de sentimentos, monitoramento de tendências, descoberta de tópicos frequentes e pesquisa textual [McNaught and Lam, 2010].

II.7.3 Uniformização de Letras

A uniformização de palavras em textos, seja convertendo todas para maiúsculas ou minúsculas, é uma etapa crucial no pré-processamento de dados em PLN. Essa técnica, frequentemente referida como "normalização de caixa", visa reduzir a variabilidade e a complexidade dos dados, facilitando a análise subsequente. Ao converter todas as palavras de um texto para um formato uniforme, seja ele todo em maiúsculas ou todo em minúsculas, elimina-se uma fonte comum de variação nos dados, permitindo que algoritmos de PLN tratem palavras como "Brasil", "BRASIL" e "brasil" como a mesma entidade [Pustejovsky and Stubbs, 2012].

Isso é especialmente útil em tarefas como a tokenização, a análise morfológica e a extração de entidades nomeadas, onde a consistência na representação das palavras pode melhorar significativamente a precisão dos modelos. Além disso, a normalização de caixa pode ser essencial quando se trabalha com fontes de dados heterogêneas, onde diferentes convenções de nomenclatura ou estilos de escrita podem ser usados. Algumas ferramentas de PLN biomédicas já possui esse procedimento para melhor realizar um mapeamento de conceitos [Kang et al., 2013].

II.7.4 Reconhecimento de Entidade Nomeada (REN)

O Reconhecimento de Entidade Nomeada (REN) é uma subárea de estudo pertencente ao domínio de IR, cuja tarefa é reconhecer as entidades a partir de textos de forma livre. Uma vez reconhecido, estes textos são classificados em um conjunto de regras e categorias pré-definidas, tendo como resultado uma representação estruturada desse texto com suas respectivas anotações [Oliveira et al., 2014]. Para isso há diversas soluções, para o português do Brasil como o NERP *Conditional Random Fields* (NERP-CRF) [Oliveira et al., 2014] e OpenNLP [Fonseca et al., 2015]. Atualmente, a ferramenta mais utilizada multilíngue e inclusive em português, é a biblioteca em Python *spaCy*.

II.7.5 Correção Textual

A correção textual é uma etapa fundamental no pré-processamento de dados em PLN. Esta etapa visa identificar e corrigir erros ortográficos, gramaticais ou de estilo presentes nos textos, garantindo assim uma maior precisão e eficácia nas análises subsequentes. Erros textuais podem surgir de diversas fontes, como entradas manuais, transcrições automáticas ou traduções imperfeitas, e podem comprometer a qualidade dos resultados obtidos em tarefas de PLN.

II.8 Ferramentas de PLN Aplicadas à Área Biomédica

Nesta seção são descritos os principais conceitos, vocabulários e ferramentas dentro da PLN aplicadas à área biomédica e utilizados nesse trabalho de pesquisa. São abordados o MetaMap e suas ferramentas, o Vocabulário Sistema Unificado de Linguagem Médica (UMLS), seus componentes, o Modelo de Dados em Português do Brasil, e por fim, a Aplicação Java para Mapeamento e Serviços que realiza o mapeamento conceitual de termos.

II.8.1 MetaMap

O MetaMap é um programa configurável de PLN da NLM[6]. Sua função principal é mapear texto biomédico para o *Metathesaurus* UMLS ou para descobrir os conceitos do *Metathesaurus* referidos em diversos tipos de texto. Ela possui as seguintes características [Rogers et al., 2012]:

- usa uma abordagem intensiva em conhecimento baseada em técnicas simbólicas, PLN e técnica de linguística computacional. Além de ser aplicado para IR e mineração de dados.
- é uma das bases do *Medical Text Indexer* (Indexador de Texto Médico) (MTI) da NLM, que está sendo usado para indexação semiautomática e totalmente automática da literatura biomédica com conceitos rotulados na NLM [Shannon et al., 2021].

[6] `https://www.ncbi.nlm.nih.gov/`

- é uma ferramenta para reconhecer conceitos UMLS em texto livre. Apesar do UMLS se apoiar em fontes de informação biomédica, os algoritmos da MetaMap são independentes do domínio e podem ser usados com qualquer domínio que forneça fontes de conhecimento adequadas.

Técnicas simbólicas são caracterizadas por avaliação das palavras servindo-se de recursos léxicos, determinando se o seu sentido é positivo ou negativo. A entrada pode ser texto de comprimento variável e sua saída é uma lista classificada de conceitos de *Metathesaurus* associados a cada parte do texto. Este serviço é possível por causa do *Lexical Tools* (Gerador de Variante Lexical) (LVG) [Olivier, 2004].

II.8.2 MetaMap Data File Builder (DFB)

O *MetaMap Data File Builder* (DFB)[7] é um módulo que permite a criação de modelos de dados UMLS personalizados. Por padrão o MetaMap sempre usa um modelo em inglês, o *USABase*. Com o DFB, podem ser gerados modelos em outras linguagens, por exemplo, para o português do Brasil [Rogers et al., 2012].

Existem três formas de criar uma base de conhecimento MetaMap: o *Metathesaurus* UMLS completo que mapeia texto bastando instalar o MetaMap de forma padrão. A segunda forma é um subconjunto do *Metathesaurus* UMLS, limitando vocabulários, ora por desempenho, ora por direitos autorais de algumas fontes ou uma língua diferente do inglês, para isso é necessário usar a ferramenta MetamorphoSys contida no *Specialist Lexicon & Lexical Tools* (Léxico ESPECIALISTA). Por fim, um banco de dados não UMLS, a fim de identificar conceitos em fontes de conhecimento não baseados no *Metathesaurus* UMLS, é necessário criar um conjunto de dados do zero.

O trabalho de Perciliano et al. [2023] gerou um modelo UMLS personalizado para o português do Brasil usando o MetaMap Data File Builder (DFB). Este modelo em conjunto com o MetaMap realizou mapeamentos conceituais de termos em português do Brasil sem a utilização da *Application Programming Interface* (Interface de Programação de Aplicativos) (API) do MetaMap. Além disso, o trabalho mostrou desafios e dificuldades no uso do MetaMap.

II.8.3 Sistema Unificado de Linguagem Médica (UMLS)

O UMLS é um conjunto de arquivos e aplicações que reúnem termos sinônimos de 214 vocabulários biomédicos[8] anotados manualmente. A instalação usando a versão UMLS2023AB possui 187 vocabulários controlados, é possível escolher quais vocabulários se deseja trabalhar. Eles estão disponíveis em 28 línguas, sendo a língua inglesa a maior em número com 105 vocabulários, em

[7] `https://metamap.nlm.nih.gov/Docs/datafilebuilder.pdf`
[8] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html`

seguida o espanhol com dez, o português do Brasil têm apenas dois, mas existem mais quatro especificamente "Portuguese", conforme a Figura II.3.

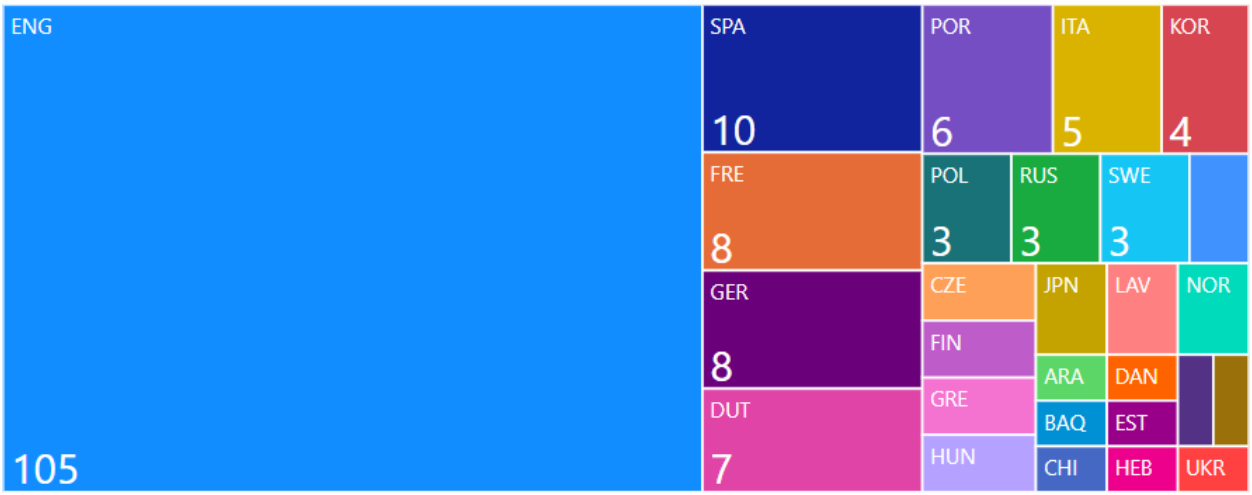


Figura II.3: Quantidade de Vocabulários por Língua.

Vocabulário controlado ou estruturado é uma lista de termos que tem a função de desenvolver uma linguagem estruturada. Linguagem estruturada é uma linguagem organizada e classificada, seguindo uma estrutura com seus termos ou conceitos listados por categorias e hierarquizadas. Os termos listados nestes vocabulários são chamados de descritores [Davanzo and Moreira, 2014].

Um vocabulário controlado ajuda o pesquisador ou sistema computacional na IR através do termo exato usado ou utilizando descritores. O UMLS utiliza tesauros da área biomédica que garantem o processamento e a IR.

Tesauro, conhecido também por *thesaurus* ou dicionário de ideias afins, é uma lista de palavras com significados semelhantes em um domínio específico de conhecimento. Por definição, um tesauro é restrito. Seu objetivo é mostrar as diferenças mínimas entre as palavras e auxiliar na escolha da palavra exata ou a "preferida" em uma área específica. Os tesauros não incluem definições. Logo, *Metathesaurus* é um conjunto de tesauros [Laan, 2002].

Sinônimos nesse contexto é o mesmo que *Concept Unique Identifier* (CUI). Cada vocabulário contido em uma fonte, tem um único identificador atômico (AUI). Entretanto, um mesmo vocábulo ou sua variação pode estar em uma fonte diferente, que terá seu AUI, no entanto, o CUI será o mesmo, podendo os mesmos pertencerem ao mesmo grupo semântico.

A Figura II.4 apresenta um quadro que mostra o seguinte entendimento: o grande bloco na cor laranja apresenta duas fontes, o SNOMECT_US e o MSH, as *strings* "Headache", "Headaches" e "Cranial Pains" são distintas, porém a fonte MSH mantém os vocábulos "Headaches" e "Cranial Pains" com o mesmo código SCUI. O UMLS interpreta que cada vocábulo distinto, tem seu AUI específico e único, porém, semanticamente todos possuem o mesmo conceito (CUI) e o mesmo grupo semântico "Disorders".

Source Vocabulary				UMLS		
Source	SCUI	String	Semantic Group	AUI	LUI	CUI
SNOMEDCT_US	139490008	Headache	Disorders	A22825519	L0018681	C0018681
MSH	M0009824	Headaches	Disorders	A0066008	L0018681	C0018681
MSH	M0009824	Cranial Pains	Disorders	A1641924	L1406212	C0018681

Figura II.4: Processo de Construção do UMLS *Metathesaurus* [Nguyen et al., 2021].

O UMLS utiliza por padrão um modelo de dados em língua inglesa, o *USABase*, contudo, existem os modelos *Base* e NLM. A diferença entre os três modelos são a quantidade de vocabulários e algumas restrições. Na língua inglesa, são mais de 10 milhões de termos, agrupados em aproximadamente 3.85 milhões de conceitos [Olivier, 2004; Nguyen et al., 2021].

A Figura II.5 ilustra a arquitetura do *Metathesaurus* UMLS. A caixa da figura com o nome "*Metathesaurus* UMLS" possui três componentes, também chamados de fontes de conhecimentos: o *Metathesaurus*, a Rede Semântica e o *Specialist Lexicon & Lexical Tools* (Léxico ESPECIALISTA), que podem ser usados separados ou em conjunto. O trabalho de Abbas et al. [2019], extraiu informações pertinentes de textos clínicos buscando a melhoria da tomada de decisões no âmbito médico. Além do trabalho de Abbas et al. [2019] apontar os desafios enfrentados nesse processo, principalmente devido a complexidade dos registros clínicos, delineia um *framework* meticulosamente estruturado para a IR. O processo inicia-se com a "Entrada de Texto Clínico", onde os registros clínicos são introduzidos no sistema, estabelecendo assim a fundação para a subsequente extração de informações (IR). Segue-se o "Pré-processamento", uma fase na qual o texto é rigorosamente refinado. Esta etapa é vital para assegurar a precisão na extração subsequente, envolvendo atividades como correções ortográficas, remoção de palavras supérfluas e a adequada preparação do texto para o mapeamento.

O cerne do *framework* reside no "Mapeamento com o UMLS *Metathesaurus*". Neste estágio, o texto é meticulosamente comparado e mapeado com o UMLS *Metathesaurus*. Posteriormente, o "Pós-processamento" entra em ação como uma etapa de aprimoramento, onde as informações extraídas são otimizadas, garantindo sua relevância e eliminando redundâncias. O processo culmina na "Saída de Informações Extraídas", onde as informações, agora refinadas e pertinentes, são apresentadas com: os termos, os conceitos para cada termo, os tipos semânticos identificados e os tipos de entidades, ambas, prontas para auxiliar os profissionais de saúde na tomada de decisões informadas [Abbas et al., 2019].

II.8.3.1 *Metathesaurus*

O *Metathesaurus* é o principal componente do UMLS e possui um repositório de conceitos biomédicos inter-relacionados [Olivier, 2004]. É desenvolvido a partir de uma vasta lista de ter-

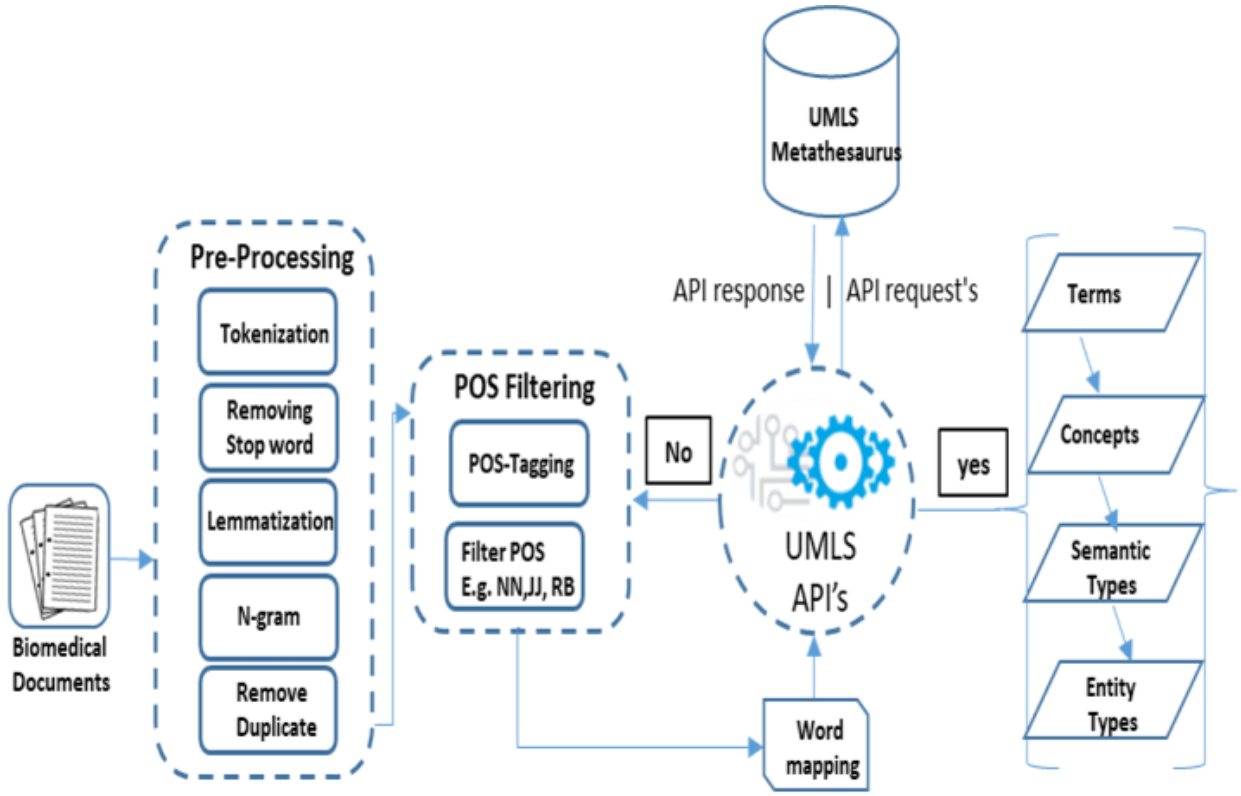


Figura II.5: Estrutura de Extração de Informações do *Metathesaurus* UMLS [Abbas et al., 2019].

mos controlados utilizados na assistência de pacientes, em estatísticas na área de saúde pública e diversas fontes de vocabulário biomédico. Cada termo do *Metathesaurus* é chamado de *string* do *Metathesaurus* e possui um identificador único que deve constar em pelo menos uma fonte de vocabulário. Caso este termo conste em mais de um vocabulário, um dos termos deve ser escolhido como preferencial pela UMLS. Além disso, um termo pode se relacionar com outro termo. Exemplificando o termo "Dor": está no vocabulário A como "dor", no B como "Dor" e no C como "DOR", a partir destes termos, a ferramenta escolhe como conceito o termo "DOR" do vocabulário C. No momento do mapeamento pode se optar para apresentar todos os termos caso existam, o preferido e os "não preferidos" com seus respectivos vocabulários.

O MetamorphoSys[9] é o assistente de instalação do UMLS, ferramenta de parametrização incluída em cada versão UMLS. Este instala uma ou mais fontes de conhecimento UMLS e permite a criação de subconjuntos personalizados do *Metathesaurus*, é recomendado utilizar sempre a última versão, com esta é possível instalar uma base em português.

Os vocabulários biomédicos utilizados nesta pesquisa foram os: LOINC Linguistic Variant, ICPC Portuguese, WHOART Portuguese, MeSH Portuguese, MedDRA Brazilian Portuguese e LOINC Linguistic Variant - Portuguese, Brazil. A descrição de cada vocabulário é apresentado no Apêndice A.

[9] `https://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/help.html`

II.8.3.2 Rede Semântica

A Rede Semântica (*Semantic Network*) é uma das três fontes de conhecimento pertencente ao UMLS. Esta fonte fornece categorias de alto nível usadas para categorizar cada conceito de *Metathesaurus* e recursos lexicais. São chamados de grupos semânticos dos conceitos do *Metathesaurus*, fornecendo um conjunto de relações úteis entre esses grupos semânticos [Olivier, 2004].

A Rede Semântica UMLS diminui a complexidade do *Metathesaurus* agrupando os conceitos em tipos semânticos, grupos semânticos e relações semânticas. A regra é que todo termo pertença a um tipo semântico e cada termo pertença a um grupo semântico. Os grupos semânticos contemplam 99,5% dos conceitos do *Metathesaurus*. Atualmente são 133 tipos semânticos distribuídos em 15 grupos e 54 relações semânticas.

A Figura II.6 apresenta nas colunas os grupos semânticos, no eixo "y" é informado quantos tipos semânticos existem em cada grupo. A coluna sem nome no meio da figura, se dá devido a falta de mapeamento interno da UMLS, onde alguns termos ainda não foram categorizados.

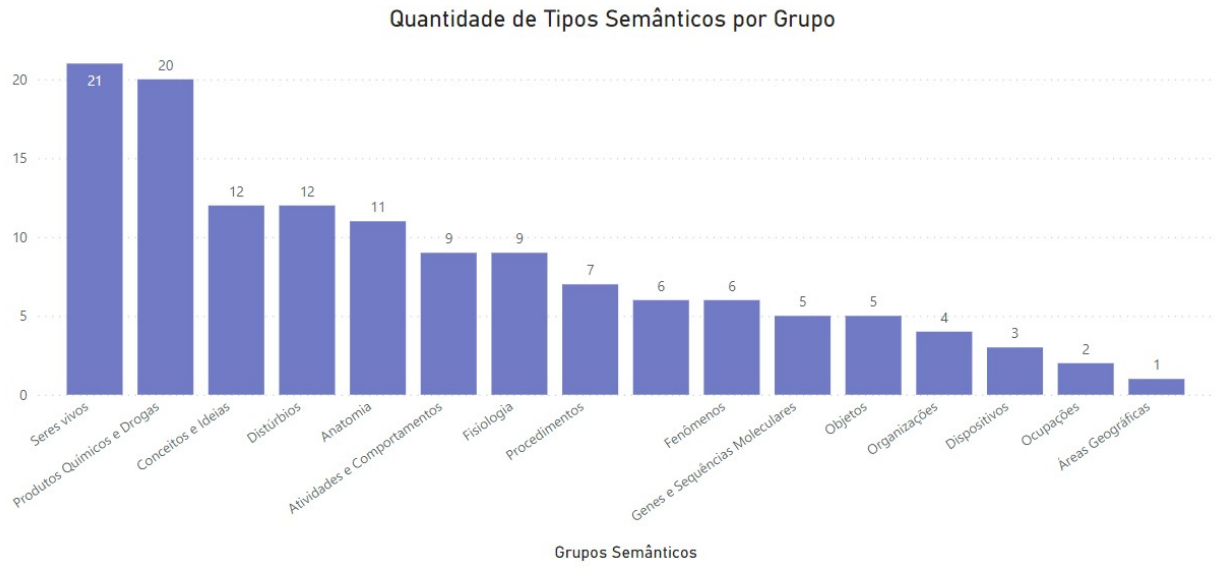


Figura II.6: Quantidade de Tipos Semânticos por Grupo.

O *Metathesaurus* consiste em termos de suas respectivas fontes de vocabulário. O significado de cada termo é definido pela sua fonte, seja por meio de definição explícita ou anotação, pelo contexto (sua posição na hierarquia), por sinônimos e outras relações declaradas entre os termos, e pelo seu uso em descrição, classificação ou indexação. Cada conceito do *Metathesaurus* é associado a pelo menos um tipo conceitual [National Library of Medicine (US), 2009a].

Em todos os casos, o tipo conceitual mais específico disponível na hierarquia é atribuído ao conceito. Por exemplo, o conceito "Macaca" recebe o tipo conceitual "Mamífero" porque não existe um tipo mais específico como "Primata" disponível na rede. O nível de detalhamento varia na rede, o que tem implicações importantes na interpretação do significado atribuído a um conceito

do *Metathesaurus*. Por exemplo, uma subárvore abaixo do nó "Objeto Físico" é "Objeto Manufaturado", que possui apenas dois nós filhos, "Dispositivo médico" e "Dispositivo de pesquisa". Existem objetos manufaturados além destes dispositivos. Ao invés de aumentar o número de tipos conceituais para abranger várias subcategorias adicionais para esses objetos, os conceitos que não são dispositivos médicos ou dispositivos de pesquisa são atribuídos ao tipo conceitual mais abrangente "Objeto fabricado" [National Library of Medicine (US), 2009a].

A Figura II.7 apresenta uma seção da rede. O tipo conceitual "Atividade Biológica" possui dois nós filhos, "Função Fisiológica" e "Função Patológica", e cada um deles por sua vez possui vários filhos e descendentes. Cada filho na hierarquia está ligado ao seu pai por meio do *link* "isa" [National Library of Medicine (US), 2009a].

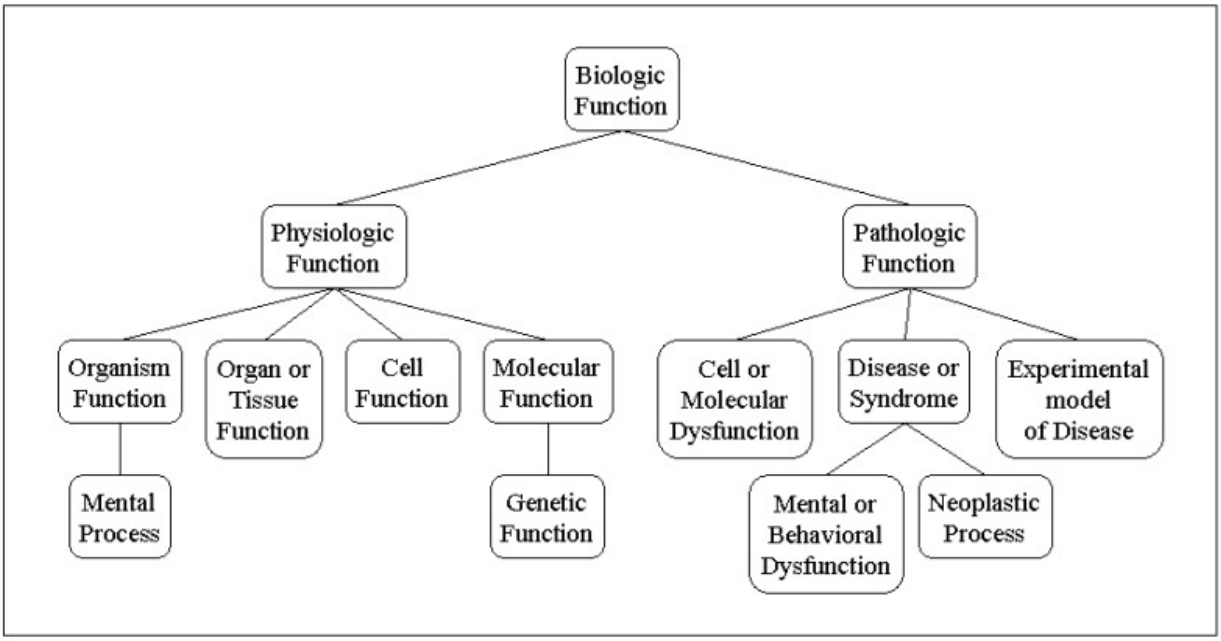


Figura II.7: Uma Parte da Rede Semântica UMLS: Hierarquia de "Funções Biológicas" [National Library of Medicine (US), 2009a].

As conexões são estabelecidas entre "tipos conceituais" de alto nível na Rede Semântica sempre que possível e são geralmente herdadas através do *link* "isa" por todos os descendentes desses tipos. Assim, por exemplo, é afirmado que a relação "processo de" existe entre os "tipos semânticos" de "Função Biológica" e "Organismo". Portanto, também é válido entre "Função de Órgão ou Tecido" (que é uma "Função Fisiológica", que por sua vez é uma "Função Biológica") e "Animal" (que é um "Organismo") [National Library of Medicine (US), 2009a].

As relações são estabelecidas entre os tipos conceituais e não necessariamente se aplicam a todas as instâncias de conceitos atribuídos a esses tipos. Em outras palavras, a relação pode ou não existir entre um par específico de conceitos. Portanto, embora a relação "avaliação de" exista entre os tipos semânticos de "Sinal" e "Atributo do Organismo", um sinal ou atributo específico pode não

estar vinculado por essa relação. Por exemplo, “excesso de peso” e “febre” são avaliações dos atributos do organismo “peso corporal” e “temperatura corporal”, respectivamente. No entanto, “excesso de peso” não é uma avaliação da “temperatura corporal” e “febre” não é uma avaliação do “peso corporal” [National Library of Medicine (US), 2009a].

Em alguns casos, pode haver um conflito entre a colocação dos tipos na Rede e o *link* a ser herdado. Nesses casos, diz-se que a herança do *link* é bloqueada. Por exemplo, por herança, o tipo “Processo Mental” seria “processo de” “Planta”. No entanto, como as plantas não são seres sencientes, esse *link* é explicitamente bloqueado. Em outros casos, a natureza da relação é tal que não deve ser herdada pelos descendentes dos tipos vinculados. Nesses casos, a relação é explicitamente definida para os dois tipos semânticos vinculados, mas bloqueada para todos os descendentes desses tipos. Por exemplo, “parte conceitual de” vincula “Sistema corporal” e “Estrutura anatômica totalmente formada” [National Library of Medicine (US), 2009a].

A Figura II.8 ilustra uma seção da Rede Semântica, mostrando as relações, hierárquicas ou associativas, existentes entre os tipos semânticos.

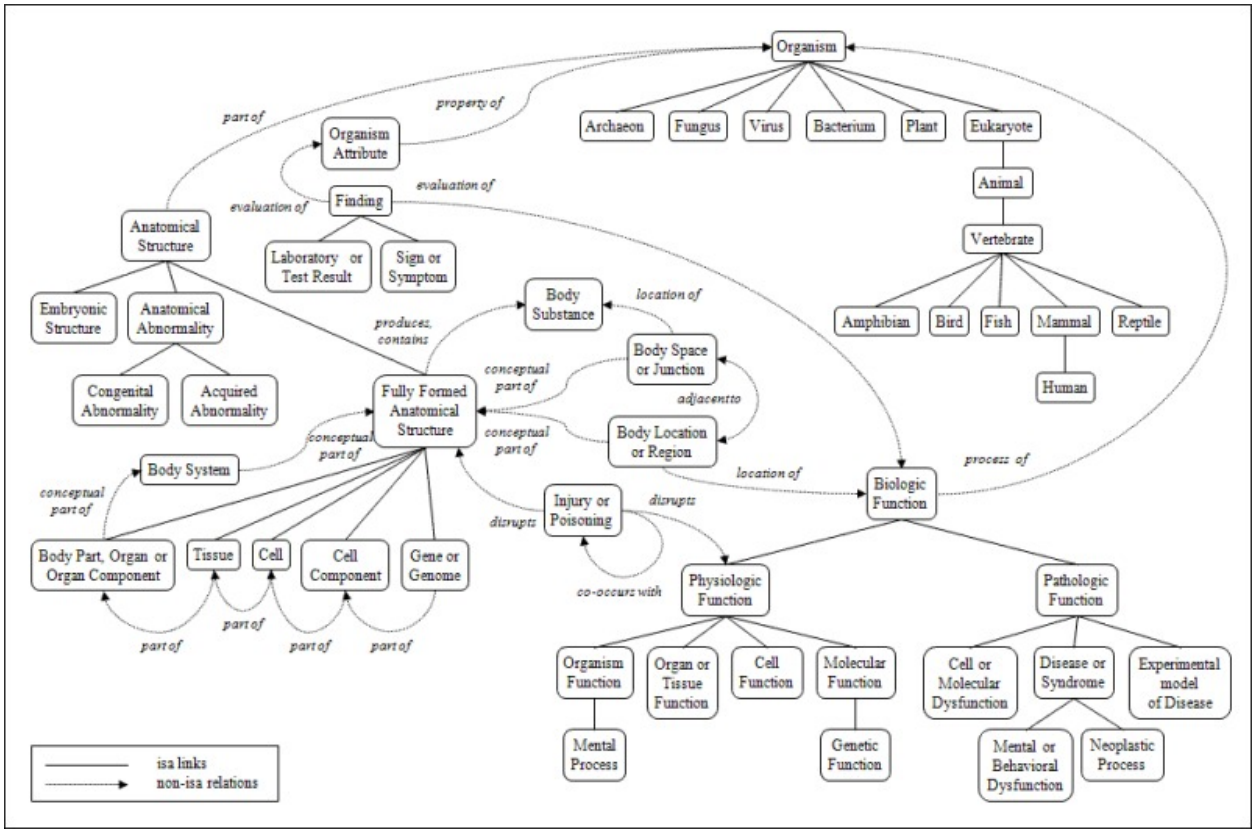


Figura II.8: Uma Parte da Rede Semântica UMLS: Relações [National Library of Medicine (US), 2009a].

II.8.3.3 *Specialist Lexicon & Lexical Tools* (Léxico ESPECIALISTA)

A versão 2023 é uma versão 21 da *Java Lexical Tools* e inclui sete ferramentas: Lvg, Norm, LuiNorm, wordInd, toAscii, fields e lgt. Os três programas principais são: **Norm** - um normalizador, **wordInd** - um gerador de índice de palavras e o **Lvg** - um gerador de variante léxica. O *Specialist Lexicon & Lexical Tools* (Léxico ESPECIALISTA) foi desenvolvido para fornecer as informações lexicais necessárias ao Sistema de PLN ESPECIALISTA. Destina-se a ser um léxico inglês geral que contém muitos termos biomédicos. Cobre palavras comuns em inglês e vocabulário biomédico. A entrada do léxico para cada palavra ou termo registra as informações sintáticas, morfológicas e ortográficas necessárias ao Sistema de PLN ESPECIALISTA [National Library of Medicine (US), 2009b].

O LVG[10] é um conjunto de ferramentas básicas de PLN fundamentais para recuperar variantes lexicais. É baseado no Léxico ESPECIALISTA e regras codificadas manualmente. Este programa permite que os usuários gerem variantes lexicais baseadas na flexão (plural e singular) e derivação (adjetivo e substantivo), bem como executar tarefas úteis para mapeamento de termos e IR, como a remoção de palavras semanticamente sem importâncias (*stop words*) ou abstraindo da variação do hífen. O LVG ainda inclui algumas funções como: recuperação de variantes flexionais, variantes ortográficas, variantes derivacionais, variantes lexicais, criação de índices, sinônimos, normalização de caracteres, conversão de UTF-8 para ASCII, conversão de caracteres para minúsculas, etc. São 64 componentes de fluxo e 37 operações de comando oferecidas neste conjunto de ferramentas [Olivier, 2004].

As principais ferramentas léxicas compactadas são:

Normalização (Norm) Este programa gera as *strings* normalizadas utilizadas no índice de *string* normalizado MRXNS. Portanto, é necessário realizar a normalização antes de pesquisar no MRXNS. O processo de normalização envolve a remoção de possessivos, substituição da pontuação por espaços, eliminação de palavras de parada, conversão de todas as palavras para caixa baixa, quebra da *string* em suas palavras constituintes e ordenação alfabética das palavras. As formas não flexionadas são geradas utilizando o *Specialist Lexicon & Lexical Tools* (Léxico ESPECIALISTA) quando as palavras estão presentes no mesmo, caso contrário, são geradas algoritmicamente. Caso uma forma possa ser uma flexão de mais de uma forma base, o processo de normalização retorna várias formas não flexionadas. Se uma *string* ao ser normalizada contiver múltiplas formas ambíguas e a permutação dessas formas oferecer mais de 10 formas de saída, a forma de entrada em minúsculas, com a pontuação substituída e a ordem das palavras classificada, mas não flexionada, será retornada. O limite superior do número de permutações (10) é configurável mediante modificação no arquivo de configuração.

[10] `https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/web/index.html`

Índice de palavras (wordInd) Este programa divide *strings* em palavras para uso com o índice de palavras em MRXW. Os usuários do índice de palavras devem utilizar o wordInd para quebrar as *strings* em palavras antes de pesquisar no índice de palavras. Isso assegura a correspondência entre as palavras a serem pesquisadas e o índice de palavras. A palavra, para este fim, é definida como um *token* contendo apenas caracteres alfanuméricos, com comprimento igual ou superior a um. O programa wordInd converte as palavras de saída para letras minúsculas.

Geração de variante léxica (lvg) O lvg gera variantes léxicas a partir de palavras de entrada. Ele consiste em vários componentes de fluxo diferentes que podem ser combinados de várias maneiras para produzir as variantes. O usuário do lvg seleciona as combinações de componentes de fluxo e os combina em um fluxo (*pipeline*).

As ferramentas lexicais são projetadas para lidar com o alto grau de variabilidade nos termos e palavras da linguagem natural. As palavras geralmente têm várias formas flexionadas que podem ser consideradas instâncias da mesma palavra. Por exemplo, o verbo "tratar" tem três flexões: "trata" forma do presente do indicativo da terceira pessoa do singular, "tratou" do pretérito e do particípio passado e "trata" do particípio do presente. Termos com várias palavras em *Metathesaurus* e outros vocabulários controlados podem ter variantes de ordem de palavras além de suas variantes flexionadas e alfabéticas. As ferramentas lexicais permitem ao usuário abstrair desse tipo de variação.

O Léxico consiste em um conjunto de entradas lexicais com uma entrada para cada ortografia ou conjunto de variantes ortográficas em uma parte específica do discurso. Itens lexicais podem ser termos de "várias palavras" compostos de outras palavras se o termo de várias palavras for determinado como um item lexical por sua presença como um termo em inglês geral ou dicionários médicos, ou em tesauros médicos, como o *Medical Subject Heading* (Títulos de Assuntos Médicos) (MeSH). Extensões para acrônimos e abreviações comumente usados também são permitidas como termos com várias palavras. A notação léxica de uma unidade é uma estrutura de quadro que consiste em *slots* e preenchimento. Cada registro lexical possui um identificador exclusivo (EUI) do registro.

II.8.3.4 Processamento do MetaMap e a UMLS

O processamento de mapeamento biomédico pode ser realizado por diversas combinações de ferramentas e bases biomédicas. Com a MetaMap e a UMLS, os conceitos biomédicos podem ser mapeados e divididos em cinco etapas: pré-processamento, geração de variantes, busca de candidatos, avaliação de candidatos e construção de mapeamentos conforme a Figura II.9.

Segundo Aronson [1996], no pré-processamento, na primeira etapa são definidas as *tags* sintáticas de cada palavra do texto de entrada. Na segunda, são geradas as possíveis variações flexionadas e

ortográficas das palavras de entrada. Na etapa três são recuperadas todos o termos *Metathesaurus* da UMLS com suas respectivas informações. Na etapa seguinte é calculado o *score* de cada candidato encontrado. Por fim, a construção do mapeamento define a melhor combinação de candidatos. Mais detalhes deste mapeamento estão no Apêndice B.

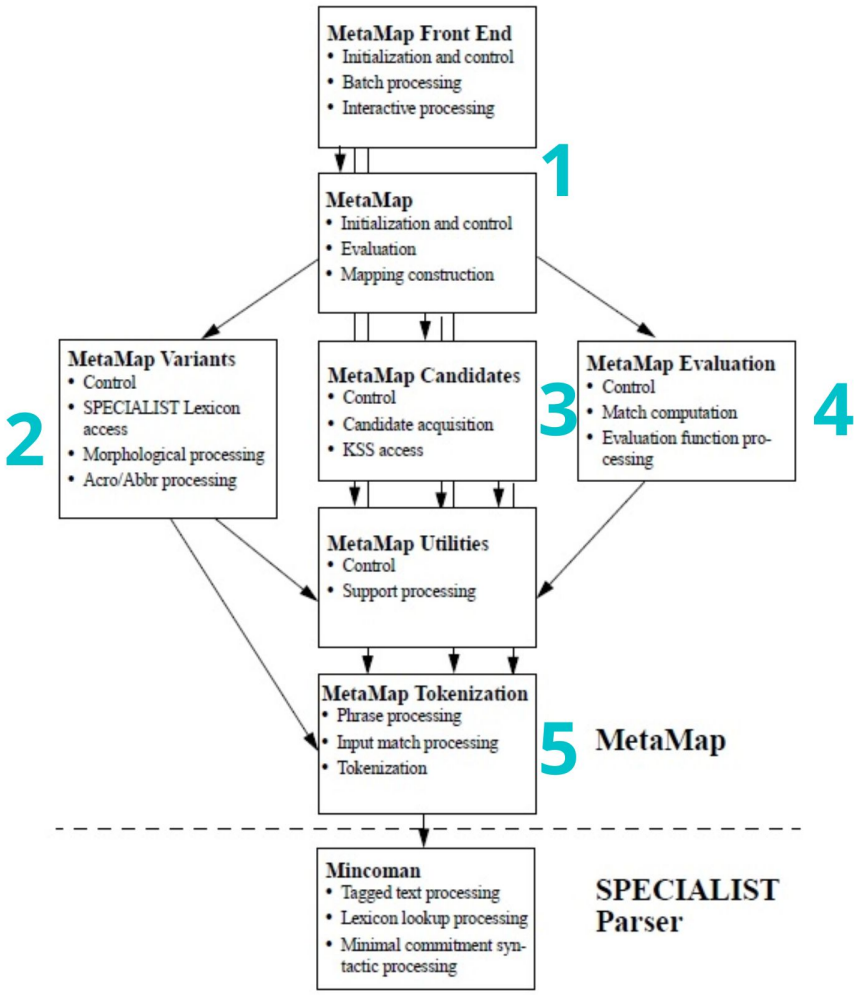


Figura II.9: Etapas de Processamento do MetaMap [Aronson, 1996].

II.8.4 Modelo de Dados em Português do Brasil

Um modelo de dados do MetaMap, é uma estrutura resultante do processo executado conforme a Subseção II.8.2. A geração de um modelo de dados está melhor detalhado no Relatório "MetaMap Instalação e Adaptação para o português do Brasil"[11]. O modelo criado para este projeto foi baseado no padrão *USABase* contendo apenas vocabulários em português do Brasil com o nome "UMLSPORTUGUES2023".

A estrutura deste modelo é armazenada e acessível por meio do banco de dados Berkeley DB da Oracle. De uma família de bibliotecas de banco de dados de chave-valor, o Berkeley DB fornece serviços de gerenciamento de dados escaláveis de alto desempenho. Os produtos Berkeley DB usam

[11] `https://github.com/LuizPerciliano/metamap-portugues`

APIs para acesso e gerenciamento de dados. Ele fornece uma coleção de tecnologias que podem ser configuradas para atender qualquer necessidade, desde dispositivo portátil a *data center*, de solução de armazenamento local a uma solução distribuída mundialmente, de *kilobytes* a *petabytes*[12]. Uma das maneiras de acessar essa estrutura e explorar resultados é utilizando uma aplicação em java, que é detalhada na Subseção II.8.5.

II.8.5 Aplicação Java para Mapeamento e Serviços

O MetaMap disponibiliza ferramentas e um programa padrão em java que pode ser adaptado para utilização do modelo de dados. Detalhes destas ferramentas e programas estão melhor detalhados no relatório referido na Subseção II.8.4. Abaixo é mostrado algumas dos principais atributos do programa MetaMap Java API[13]:

- *Candidates*: estrutura de dados que representa um possível candidato do UMLS para o mapeamento.
- Ev: instância de avaliação do candidato que contém as informações de nome de conceito, a *string Metathesaurus* localizada, o valor do *score*, o tipo semântico, a fonte de vocabulários, dentre outras.
- MetaMapAPI: *interface* que acessa o servidor do MetaMap. Esta possui métodos que realizam o processamento do MetaMap com diferentes entradas de textos e também arquivo, permite definir o modelo de dados, restringir o grupo semântico, dentre outros.
- PCM: *interface* que contém os conjuntos das classes *Phrase, Candidates* e *Mapping*.
- *Result*: um contêiner para a representação de saída da máquina do resultado do MetaMap.
- *Utterance*: representação de uma estrutura de dados que contém o texto de entrada, id próprio, tamanho total do texto e outras informações.
- *Phrase*: estrutura de dados que representa uma parte do texto de entrada, geralmente separadas por pontuações.
- *Mapping*: estrutura de dados que representa o resultado da construção do mapeamento.

Mais detalhes e execução do programa em java é detalhado na Seção III.2.1.

[12] `https://www.oracle.com/database/technologies/related/berkeleydb.html`
[13] `https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/javaapi/javadoc/index.html`

II.9 Twitter como Fonte de Dados para Estudos de Farmacovigilância

O Twitter, uma mídia social e serviço de troca de mensagens curtas e frequentes, constitui uma vasta rede de informações em texto puro, tornando-se propício para extrair dados para estudos diversos [Zhang et al., 2018]. Através dessa plataforma de *microblog*, os usuários podem pesquisar e compartilhar informações sobre uma ampla gama de tópicos, incluindo relatos de EAs, realizados por pacientes, profissionais da saúde, instituições, revistas, jornais, entre outros.

Esta rede social é utilizada por uma parcela significativa da população global. No Brasil, mais da metade da população se engaja ativamente com o Twitter, tornando-o a segunda maior plataforma em número de usuários no país. Apesar disso, a segunda língua mais prevalente no Twitter é o japonês, respondendo por 10,6% das postagens, enquanto o inglês lidera com 25,8%[14].

O Twitter têm duas APIs de coleta: a *REST API* e a *Streaming API*. Ambas permitem apenas o acesso a tuítes recentes. Recentemente, em 2023, o Twitter passou por uma mudança significativa com a aquisição por Elon Musk, CEO da SpaceX e da Tesla. Atualmente a plataforma se chama X, mas o termo Twitter é referenciado no decorrer do texto.

A mudança nessa rede social atraiu atenção global e gerou repercussões diversas na plataforma. Uma das alterações mais notáveis foi a introdução de restrições e cobranças para o uso de suas APIs. Anteriormente, o Twitter oferecia um acesso aberto e gratuito a uma quantidade limitada de dados por meio de suas APIs, os *links* antigos apontam para o novo (https://developer.x.com/en/docs/x-api).

A nova política implementada pelo Twitter após a aquisição, incide em custos, o que representa um obstáculo para pesquisadores e pequenas empresas que dependem desses dados para suas análises e serviços. Essa mudança destaca a importância da contínua evolução das redes sociais e seus impactos na pesquisa, na indústria e neste trabalho.

Os temas discutidos no Twitter abrangem desde política e guerras até vacinas e saúde, cada país, grupo ou ideologia formando suas próprias "bolhas" de notícias. Diversos estudos na área de farmacovigilância utilizam o Twitter para diferentes fins, incluindo análise de comentários, detecção de EAs desconhecidos, e fornecimento de dados em tempo real para a farmacovigilância. Segundo Duval and da Silva [2019], a identificação de Proto-AEs (postagens com semelhanças a EAs) no Twitter é quase três vezes maior do que os relatórios da FDA.

A análise realizada pela equipe da Kepios[15] revelou que em julho de 2023, a quantidade global de usuários de mídia social alcançou aproximadamente 4.88 bilhões, equivalente a 60,6% do total populacional global. Este número observou um incremento significativo ao longo dos 12 meses antecedentes, adicionando 173 milhões de novos usuários às mídias sociais desde o ano precedente.

[14] Engajamento do Twitter no Mundo

[15] `https://datareportal.com/social-media-users`

Esta expansão representa um crescimento anualizado de 3,7%, traduzido em uma média de 5.5 novos usuários registrados a cada segundo.

Corporações, indústrias farmacêuticas, jornais, presidentes de país e diversos líderes mundiais, utilizam este microblog para enviar informes diversos. A Tabela II.2 apresenta exemplos reais de tuítes relacionados à biomedicina. Tais motivos justificam a escolha desta mídia.

Fonte	**Twitter**
Revista Oeste (10/11/2020)[1]	"Anvisa decidiu interromper os estudos clínicos da vacina chinesa da Sinovac após a ocorrência de um evento adverso grave."
Rádio BandNews FM (08/01/2021)[2]	"A empresa Janssen notificou nesta sexta-feira (8) a Agência Nacional de Vigilância Sanitária (Anvisa) sobre um evento adverso grave com um voluntário brasileiro participante do estudo da vacina contra o coronavírus da empresa."

Tabela II.2: Publicações no Twitter Acerca de EA.

[1] https://twitter.com/revistaoeste/status/1326200086423068688?t=AgaREhxT9yf6m1OH2GnJMg&s=08

[2] https://twitter.com/radiobandnewsfm/status/1347668453830877184?t=9kChch5DB-w_Fg2J-6NkRg&s=08

Capítulo III Método de Pesquisa Bibliográfica

Este capítulo explicita os passos essenciais, de acordo com a Metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), para revisão da literatura, a composição dos documentos requeridos neste estudo e análise dos trabalhos.

III.1 Análise e Processamento de Artigos Científicos

A pesquisa iniciou-se com a fase de **Entendimento do Negócio** - Item (1), onde foi realizada uma extensa revisão bibliográfica nas bases de dados indexadas Scopus e PubMed. Esta fase foi essencial para compreender as necessidades do campo da Farmacovigilância e definir os objetivos da pesquisa refinando a *string* de busca. Seguiu-se a fase de **Entendimento dos Dados** - Item (2), na qual os dados relevantes foram identificados, coletados e avaliados quanto à sua qualidade e adequação ao estudo.

Posteriormente, na fase de **Preparação dos Dados** - Item (3), os dados coletados foram limpos, transformados e formatados para análise. Esta fase envolveu o uso do programa Manutenção de Documentos de Bases Indexadas (MaDIB) para garantir a precisão e relevância dos dados no contexto da Farmacovigilância. Os desdobramentos deste processo são detalhados no Capítulo III. A fase seguinte, **Modelagem** - Item (4), foi dedicada à aplicação de técnicas estatísticas e de *Data Mining* (Mineração de Dados) (MD) para identificar padrões e relações e correlações significativas entre os trabalhos, usando também a ferramenta *Connected Papers*.

A fase de **Avaliação** - Item (5) envolveu a análise crítica dos modelos gerados, considerando os objetivos iniciais da pesquisa. Por fim, na fase de **Implementação** - Item (6), os conhecimentos adquiridos foram organizados e apresentados de forma a contribuir efetivamente para o campo da Farmacovigilância. Este processo culminou na avaliação final da pesquisa, onde os resultados foram interpretados e as conclusões foram formuladas, refletindo a aplicação prática e teórica do CRISP-DM no contexto deste estudo.

III.1.1 Manutenção de Documentos de Bases Indexadas (MaDIB)

O programa MaDIB foi criado para gerenciar e analisar os arquivos obtidos das bases de dados indexadas, resultantes das pesquisas realizadas com *strings* específicas. O MaDIB[1] foi desenvolvido utilizando Python na versão 3.10.4, juntamente com a ferramenta *Jupyter Notebook*[2], um ambiente de programação interativo que facilita a criação e compartilhamento de documentos que contêm código ativo, equações, visualizações e texto explicativo. O desenvolvimento do MaDIB é um processo contínuo, refletindo a natureza dinâmica da pesquisa.

O código-fonte do MaDIB está disponível publicamente, proporcionando um recurso valioso para pesquisadores conduzirem mapeamentos sistemáticos usando as bases de dados Scopus e PubMed. Além de ser uma ferramenta eficaz para a análise de dados neste estudo, sua aplicação potencial é ampla e pode ser adaptada para diversos contextos de pesquisa.

III.1.2 Fase 1 - Entendimento do Negócio

Esta fase envolveu uma análise detalhada do domínio da Farmacovigilância, identificando as necessidades específicas e os desafios do setor. A revisão bibliográfica nas bases Scopus e PubMed proporcionou uma compreensão abrangente do campo, permitindo a definição clara dos objetivos e questões de pesquisa. Este entendimento inicial foi fundamental para direcionar todas as fases subsequentes do estudo.

A identificação dos trabalhos relacionados a esta pesquisa deu-se por meio de um mapeamento sistemático da literatura que envolvem o sistema MetaMap e o tema Reação Adversa a Medicamento (RAM)s. A revisão da literatura ocorreu por intermédio da base de dados Scopus, reconhecida por indexar revistas e conferências significativas no campo da Ciência da Computação, que abrange resumos e citações de literatura revisada por pares.

A partir da base indexada Scopus, obteve-se um total de 541 documentos, utilizando a seguinte *string* de busca:

(((“natural language processing” AND (“adverse drug reaction” OR “adverse drug event” OR pharmacovigilance OR metamap))))

Para a base de dados PubMed, com a aplicação de uma *string* de busca diferenciada, devido aos critérios específicos de sua ferramenta de busca, obteve-se um total de 188 documentos. No entanto, as mesmas palavras-chaves foram utilizadas, conforme a *string* a seguir:

[1] Projeto MaDIB no GitHub
[2] `https://jupyter.org/`

“‘ “natural language processing” [Title/Abstract] AND (“adverse drug reaction” [Title/Abstract] OR “adverse drug event” [Title/Abstract] OR (“pharmacovigilance” [MeSH Terms] OR “pharmacovigilance”[Title/Abstract]) OR “metamap” [Title/Abstract])”’

III.1.3 Fase 2 e 3 - Entendimento e Preparação dos Dados

Esta fase combinou os Itens (2) e (3), respectivamente, Entendimento dos Dados e Preparação dos Dados do CRISP-DM. Nesta, foi realizada a coleta de dados relevantes para o estudo, seguida por um processo rigoroso de limpeza e transformação dos dados. Utilizando o programa MaDIB, os dados foram estruturados e preparados para garantir a eficiência necessária e precisão nas análises subsequentes. Esta fase foi essencial para assegurar a qualidade e a relevância dos dados para a Fase 3 - Modelagem e Avaliação.

Como resultado das buscas nestas bases de dados indexadas, duas bases de dados foram criadas, tratadas, normalizadas e unificadas. Dessa forma, foi gerada uma Planilha de Controle de Leitura de Documentos (PCLD), composta por 729 documentos, correspondendo ao total de documentos encontrados nas bases indexadas. O MaDIB conduziu vários processos na PCLD que são detalhados nas próximas seções.

III.1.3.1 Determinação de Critérios para Seleção e Exclusão de Documentos

Neste tópico são abordados os critérios e regras para seleção e exclusão de documentos oriundos das bases pesquisadas. Cada base indexada gerou um arquivo no formato “.csv”. Estes arquivos são lidos e processados pelo programa MaDIB.

Critérios de Seleção de Documentos (CSD) Os Critérios de Seleção de Documentos (CSD) abarcaram:

- CSD01 - Pesquisas dentro da área de farmacovigilância que discutem Evento Adverso (EA)s e RAMs.
- CSD02 - Estudos que abordam o Processamento de Linguagem Natural (PLN) em textos, com um enfoque particular nos extraídos de plataformas de mídias sociais.
- CSD03 - Publicações que empregam ferramentas de mapeamento de textos biomédicos para a descoberta de conceitos de farmacovigilância, dando destaque para a utilização da ferramenta MetaMap.

Critérios de Exclusão de Documentos (CED) Os Critérios de Exclusão de Documentos (CED) foram:

- CED01 - Exclusão de pesquisas duplicadas na mesma base (quatro documentos na Scopus e um na PubMed).
- CED02 - Exclusão de trabalhos com menos de quatro páginas (14 documentos na Scopus).
- CED03 - Exclusão de trabalhos categorizados como: "Conference Review" (três), "Editorial" (um), "Letter" (zero) e "Note" (dois).
- CED04 - Exclusão de pesquisas idênticas presentes em ambas as bases (160 documentos).
- CED05 - Exclusão de trabalhos de acordo com o *status* atribuído. Os trabalhos foram categorizados segundo o conteúdo apresentado e a relação com o tema desta pesquisa. Após a revisão de resumos, introduções e conclusões, restaram 534 documentos, sendo 517 da Scopus e 17 da PubMed. Devido ao alto volume de documentos, apenas aqueles com *status* de excelente e bom foram selecionados para leitura completa. Os critérios para determinação dos *status* foram:

    1. Ruim: trabalhos cuja motivação e problema apresentados na introdução não se alinham à proposta desta pesquisa; resumos e conclusões não condizentes com a introdução do trabalho;
    2. Razoável: relacionado ao trabalho, mas sem total alinhamento com o tema, podendo ser reclassificado após avaliação mais aprofundada e leitura completa do trabalho;
    3. Bom: documentos que discutem o Processamento de Linguagem Natural (PLN), termos farmacológicos em mídias sociais e/ou *Electronic Medical Record* (Registro Médico Eletrônico) (EMR);
    4. Excelente: documentos que, além das características do *status* "Bom", tratam de idiomas distintos do inglês, utilizando a base de dados *Unified Medical Language System* (Sistema Unificado de Linguagem Médica) (UMLS) e o sistema MetaMap.

Com a aplicação do CED, a PCLD ficou com 131 documentos para serem lidos completamente.

III.1.4 Fase 3 - Modelagem e Avaliação

A avaliação meticulosa dos modelos gerados assegurou que eles estavam alinhados com os objetivos da pesquisa em Farmacovigilância. A utilização do CRISP-DM permitiu identificar padrões significativos e *insights* valiosos a partir dos dados.

Os métodos de categorização na literatura científica geralmente utilizam escalas de qualidade baseadas em critérios como relevância para a pergunta de pesquisa, qualidade metodológica, clareza e consistência das conclusões. No entanto, não há consenso sobre esses critérios, como apontam os

estudos de Guyatt et al. [2008]; Higgins and Green [2011], que frequentemente revisam e atualizam essas abordagens no campo da pesquisa médica. Neste trabalho, as categorias foram definidas com base na qualidade e relevância dos trabalhos analisados, conforme descrito na Seção III.1.3.1, enquanto os detalhes sobre os processos de seleção e exclusão são apresentados na Seção Fase 4 - Implementação.

III.1.5 Fase 4 - Implementação

Finalmente, esta fase marcou a conclusão do processo de pesquisa. Nesta, os resultados e conhecimentos adquiridos foram consolidados e preparados para serem compartilhados com a comunidade científica e aplicados na prática. A implementação efetiva dos achados desta pesquisa tem o potencial de contribuir significativamente para o campo da Farmacovigilância, oferecendo novas perspectivas e soluções para os desafios identificados na Fase 1 - Entendimento do Negócio.

Os dados obtidos das bases de dados indexadas passaram por uma série de fases, conforme ilustrado na Figura III.1 e detalhado nos parágrafos subsequentes.

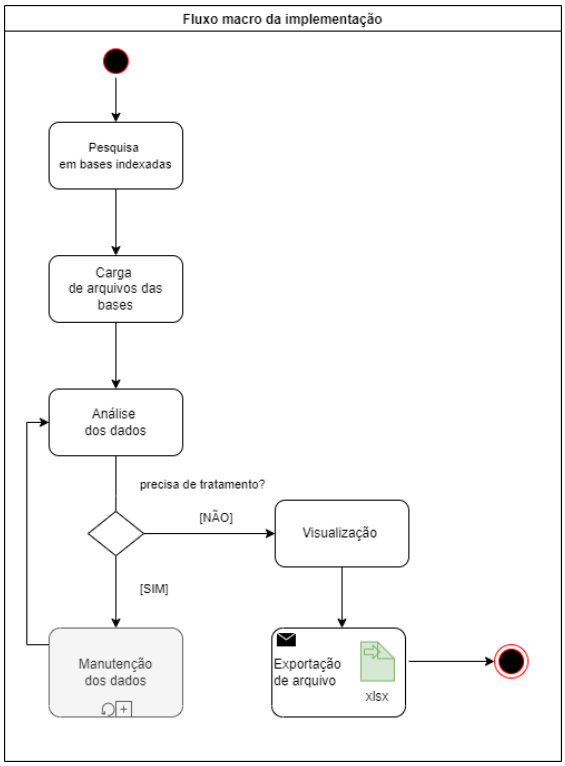


Figura III.1: Diagrama de Atividades da Implementação do programa MaDIB.

III.1.5.1 Consultas em Bases de Dados Indexadas

Os procedimentos específicos de pesquisa foram discutidos no início deste capítulo. Foram conduzidas duas consultas distintas, uma na base de dados Scopus e outra na PubMed. Embora seja possível aplicar filtros adicionais aos resultados dessas bases de dados, optamos por manter os

resultados conforme obtidos pelas *strings* de pesquisa, sem a aplicação de filtros adicionais.

III.1.5.2 Carga de Arquivos de Dados das Bases

As pesquisas realizadas nas bases de dados Scopus e PubMed resultaram em dois arquivos distintos, ambos salvos no formato ".csv". Durante o processo de exportação desses arquivos, optou-se por incluir todas as colunas possíveis de cada base de dados. Estes arquivos foram posteriormente carregados em dois *dataframes* separados.

III.1.5.3 Análise de Dados

A análise do *dataframe* consiste em verificações no conteúdo dos *dataframes* referentes ao conteúdo de pesquisa das bases indexadas, onde foram verificados:

1. o nome, quantidade de conteúdo não nulo e tipo de dado de cada coluna;
2. a quantidade de conteúdo único para cada coluna;
3. o resumo estatístico de campos numéricos.

O primeiro item apresenta o nome, tipo de dado e a quantidade de conteúdos não nulos de uma coluna. O segundo item apresenta a quantidade de conteúdo único para cada coluna, ou seja, é possível identificar se existem, por exemplo, títulos de documentos duplicados. O terceiro item é possível verificar possíveis discrepâncias de valores, por exemplo, ano mínimo e máximo dos documentos pesquisados, quantidades de citações, quantidade de páginas em um documento.

Após esta análise inicial, uma cópia do *dataframe* é realizada, uma vez que na fase de Manutenção de Dados, alguns dados e colunas são ajustados. Se a análise for realizada após a Manutenção de Dados, então, a próxima fase é a Visualização.

III.1.5.4 Manutenção de Dados

Esta parte é composta de várias fases, com objetivo de normalizar os dados, de modo a facilitar a posterior comparação de dados de ambos *dataframes*, tais como:

- **Renomeação de colunas** para facilitar o entendimento do significado e origem da coluna, por exemplo, a coluna "authors" foi renomeada para "autores_scopus" no *dataframe* referente a base indexada Scopus e para "autores_pubmed" no *dataframe* referente a base indexada PubMed.
- **Criação de colunas**, tais como, "fonte" no *dataframe* da base PubMed, coluna que existe por padrão na base Scopus e que identifica a origem da base indexada. É como cópia de colunas

já existentes, as colunas como "doi" em ambos *dataframes* e "titulo_documento_pubmed" no *dataframe* do PubMed, "titulo_documento_scopus", "inicio_pagina_" e "fim_pagina_" e "status_pagina" no *dataframe* do Scopus para posteriores análises e exclusões.

- **Manutenção nas colunas de páginas** foi necessária para ajustar os dados inseridos somente na base indexada do Scopus, pois na PubMed não existe informação referente a paginação do documento. Desses tipos de dados ajustados, podemos citar:

  1. não informado página inicial e final do documento;
  2. informado somente a página inicial do documento;
  3. informado somente a página final do documento;
  4. informado página com caractere especial, como exemplo página "1038.e1".

  Depois do ajuste das colunas de páginas, é realizada a contabilização da coluna quantidade_paginas_, que é importante na fase de exclusão de documentos com menos de 4 páginas.

- **Ajuste para minúsculo** dos textos de algumas colunas tais como, "titulo_documento" e "doi" em ambos *dataframes* e "resumo" no *dataframe* da base Scopus.

- **Remoção de caracteres especiais** na coluna "titulo_documento" de ambos *dataframes*.

- **Remoção de espaços** de textos de algumas colunas, tais como, "titulo_documento" em ambos *dataframes* e "resumo" no *dataframe* do Scopus.

- **Alteração do tipo de dados** de algumas colunas em ambos *dataframes*, como, as colunas "inicio_pagina_" e "fim_pagina_" que eram do tipo *string* e convertidas para inteiro na Scopus, o que permitiu calcular a quantidade de páginas de cada documento.

- **Exclusão de documentos** nessa fase são excluídos registros com:

  1. duplicidades;
  2. menos de 4 páginas;
  3. os tipos: *Conference Review, Editorial, Letter e Note* que não abrangem o escopo desse trabalho.

  Após essa fase, é realizada novamente a Fase de Análise de Dados.

- **Concatenação de *Dataframes*** é mais uma subfase da manutenção de dados na qual é realizada a unificação dos dois *dataframes*, para serem analisados e tratados. Em seguida são adicionadas colunas para controle da PCLD, tais como:

- status_documento: coluna para identificar se o arquivo é utilizado na pesquisa
- leitura_completa: coluna para identificar se a leitura do documento foi concluída
- observacao: coluna que serve para colocar alguma observação importante sobre o documento

Depois é realizada a tradução dos títulos e resumos dos artigos por meio da biblioteca Google Translator, a fim de se obter os termos em português para composição da nuvem de palavras na língua portuguesa.

III.1.5.5 Visualização

Esta fase visa a geração de gráficos e nuvens de palavras obtidas através do *dataframe* processado nas fases anteriores, a fim de melhorar o entendimento e análise dos dados, sendo:

- **quantidade de documentos por níveis de acessos**: os níveis de acesso especificados na base Scopus exibem a classificação por categorias de como os documentos podem ser obtidos. Esses podem ter os seguintes *status*: *Open Access Gold*, *Hybrid-Gold*, *Green* ou *Bronze*[3]. O *status Open Access Gold*, são documentos em periódicos de acesso aberto, disponíveis na plataforma da editora. O *Hybrid-Gold*, são documentos em periódicos em que os autores escolhem publicação em acesso aberto ou não, disponíveis na plataforma da editora. O *status Green*, são documentos que também podem estar disponíveis em outros repositórios de leitura gratuita na plataforma da editora. O *Bronze*, a editora escolhe fornecer acesso gratuito temporário ou permanente. Este *status Bronze* é dado a um documento caso haja outra licença específica da editora que não seja uma licença *Creative Commons*. Os *Open Access Gold* e *Hybrid-Gold* são licenças *Creative Commons*;

- **quantidade de tipo de documentos da base Scopus**: os documentos podem ser de diversas categorias, tais como: artigos de revistas/*journals* ou de conferências, capítulos de livros e editoriais;

- **veículos de publicação**: os documentos podem ser de diferentes tipos de veículos de publicação, tais como: *journal, conference proceedings, book, book series*, dentre outros. Para identificação do tipo de publicação que esses documentos pertencem, foi realizado o *download* por *Source type* da base Scopus, pois ao fazer *download* na base PubMed ele não traz esse tipo de informação;

- **quantidade de documentos por ano de publicação**: os documentos podem ser publicados em diferentes anos;

[3]https://blog.scopus.com/posts/scopus-filters-for-open-access-type-and-green-oa-full-text-access-option

- **nuvens de palavras**: as nuvens de palavras podem ser dos títulos e resumos dos documentos, bem como da junção das bases indexadas. Porém, a base PubMed não tem coluna de resumo, por esse motivo não é gerada nuvem de palavras de resumo para essa base.

III.1.5.6 Exportação de Arquivo

Da base concatenada dos dados do Scopus e PubMed é gerado um arquivo ".xlsx" - a PCLD que é utilizada para controle e análise da leitura dos documentos. Esta, à medida das leituras dos documentos e com os apontamentos, análises e observações, geram estatísticas e gráficos como as divisões dos documentos por *status*.

III.1.6 Implementação da Pesquisa

A implementação da pesquisa iniciou-se com a revisão da literatura, seguindo os critérios estabelecidos nas Seções III.1.3.1 e III.1.5. O foco foi identificar e analisar estudos relevantes na área de farmacovigilância, com destaque para pesquisas que discutem Sinal de Farmacovigilância (SF), tais como EAs e RAMs, bem como aplicações de PLN em textos biomédicos usando o MetaMap. Essa etapa permitiu mapear os principais termos e conceitos utilizados pela comunidade científica mundial, fornecendo uma base sólida para as implementações subsequentes e situando o estudo no estado da arte.

A segunda etapa envolveu o processamento de textos não estruturados provenientes do Twitter, com o objetivo inicial de realizar um mapeamento de conceitos biomédicos. Inicialmente, os textos foram normalizados e preparados para análise. Em seguida, utilizaram-se a Aplicação Java para Mapeamento e Serviços e a *API* do MetaMap para mapear os termos extraídos para a base de dados UMLS, adaptada ao português do Brasil. Esta implementação possibilitou a identificação de conceitos biomédicos relevantes e a detecção de possíveis SF. Os detalhes técnicos referentes à infraestrutura utilizada e aos resultados obtidos são apresentados no Capítulo V.

III.2 Análise dos Trabalhos

Antes da criação da farmacovigilância, as pesquisas e trabalhos com alguns cuidados com medicamentos já existiam, contudo, sem muito critério. A partir do funcionamento da farmacovigilância diversos trabalhos têm sido desenvolvidos com intuito de ajudar e assistir os profissionais da área, seja para uma melhor prescrição ou uma tomada de decisão antecipada. Para identificação dos trabalhos relacionados desta pesquisa, foi realizado um mapeamento sistemático de trabalhos utilizando as *strings* de buscas com os **Critérios de Seleção de Documentos (CSD)** e **Critérios de Exclusão de Documentos (CED)** descritos no Capítulo III.

III.2.1 Resultados Obtidos do Método de Pesquisa Bibliográfica

Foi realizada uma pesquisa acerca dos trabalhos relacionados nas bases de pesquisas Scopus e PubMed. Essa pesquisa retornou 729 trabalhos, conforme *string* de busca descrita no Capítulo III. Após isso, foi realizada a Carga de Arquivos de Dados das Bases e em seguida a Análise de Dados em que foi possível:

- observar que na coluna *Title* (Título do Documento) na base indexada Scopus foram encontrados 541 registros não nulos e 539 únicos. Isso significa que existem títulos de documentos duplicados;
- elaborar um resumo estatístico de campos numéricos para verificar possíveis discrepâncias de valores, como por exemplo em números de páginas.

Em seguida, foi executada a Etapa de Manutenção de Dados, conforme passos descritos nessa subseção. Dentre as subetapas dessa etapa, a Concatenação de *Dataframes* resultou na unificação de base Scopus (517 registros) com a PubMed (187 registros), totalizando 704 registros para serem analisados e tratados. Foi realizada a exclusão de 160 documentos duplicados. Por fim, na Etapa de Visualização foram gerados os gráficos dos 534 trabalhos restantes, conforme apresentados a seguir:

- a Figura III.2 ilustra a quantidade de documentos por níveis de acessos existentes na base Scopus. Nem todos os documentos indexados tem um tipo de acesso, neste grupo, tiveram 282 documentos com tipo de acesso em branco. Não necessariamente o acesso é bloqueado ou pago, geralmente ao selecionar o documento para *download*, é redirecionado para outro repositório. Como exemplo o documento "*A Comparative Study of N-gram and Skip-gram for Clinical Concepts Extraction*" que ao tentar ler o documento, é redirecionado para o repositório *IEEE Xplore*. Detalhes sobre o tipo de cada acesso pode ser visto na Subseção III.1.5.5.

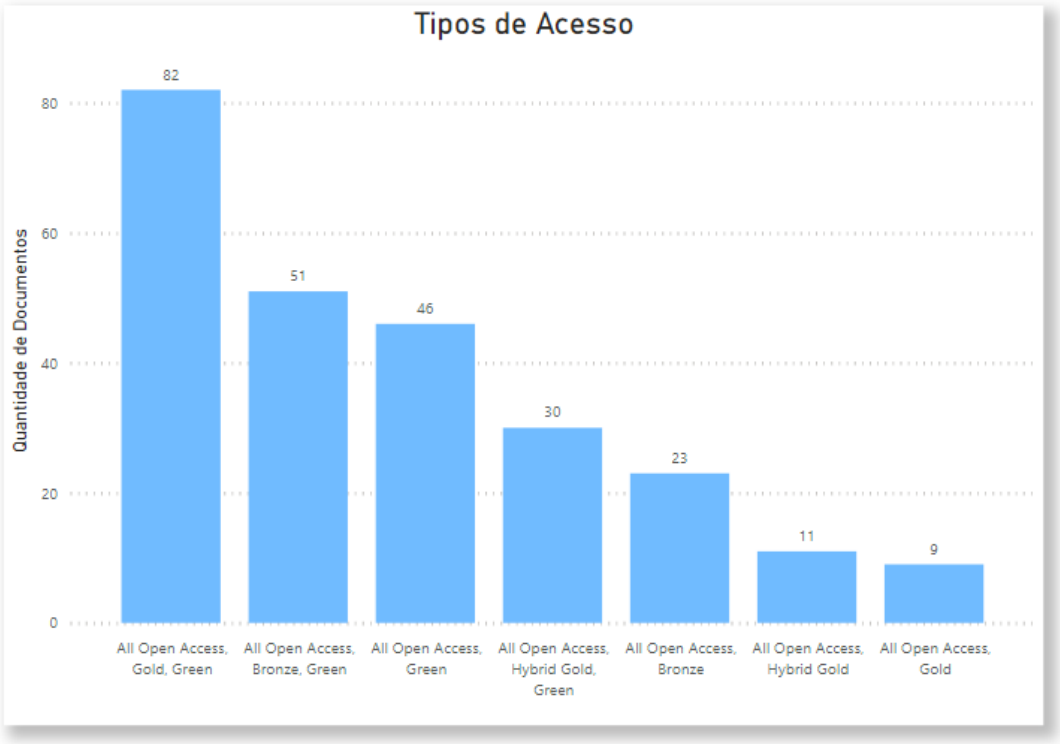


Figura III.2: Quantidade de Documentos por Níveis de Acessos.

- geralmente, o tipo de documento que mais está presente em uma pesquisa é o *Article*. Os tipos *Book, Conference Paper, Conference Review, Editorial e Note*, são encontrados em menor número, e quando eles estão presentes na *string* de busca, são eliminados devido a quantidade de páginas serem bem reduzidas e além disso, neste caso, o conteúdo não foi relevante para esta pesquisa. A base PubMed não disponibiliza em seu arquivo o tipo de documento, são os 17 documentos em branco conforme a Figura III.3.

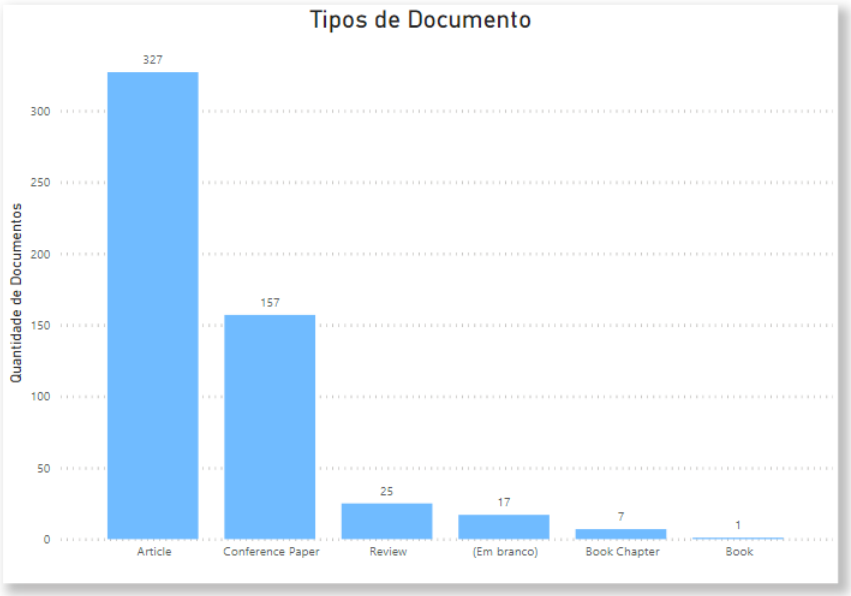


Figura III.3: Quantidade de Tipos de Documentos.

- a quantidade de documentos por veículos de publicação pode ser visualizada na Figura III.4. Na identificação do tipo de publicação, três documentos na base Scopus estavam com *Source type* como *"undefined"* e sete documentos da base PubMed não foram localizados na base Scopus. Porém, foi possível identificar o tipo de veículo ao consultar esses 10 documentos manualmente na base PubMed. Cabe salientar, que os três documentos da base Scopus que estavam como *"undefined"* ao serem consultados na base PubMed todos pertenciam ao Simpósio "AMIA Annu Symp Proc" do ano de 2012.

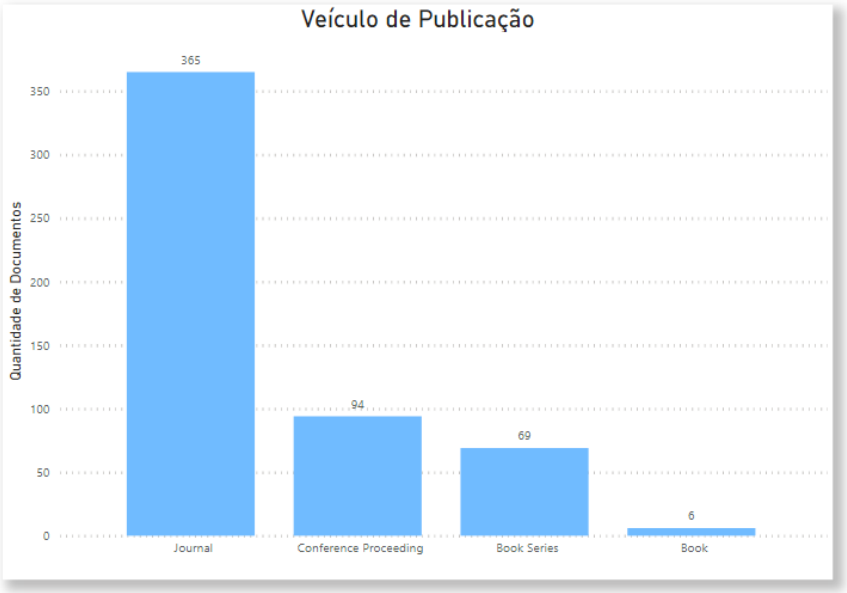


Figura III.4: Veículo de Publicação.

- apesar da ferramenta MetaMap ter sido de grande importância em seu início, as publicações foram ínfimas. Durante um período de 30 anos, apenas em 2012 atingiu uma média razoável de trabalhos publicados. Houve um aumento nos anos seguintes e estabilizou em 2017. Com a ocorrência da pandemia não houve aumento, apesar da ocorrência de maior quantidade de artigos publicados na área biomédica. Em 2016 houve uma queda brusca nas publicações assim como em 2022, apesar de ser o ano corrente desta coleta conforme a Figura III.5. De maneira similar, acontece na base PubMed de acordo com a Figura III.6. Esta possui menos documentos, pois na interseção com a Scopus, foram mantidos os documentos da Scopus.

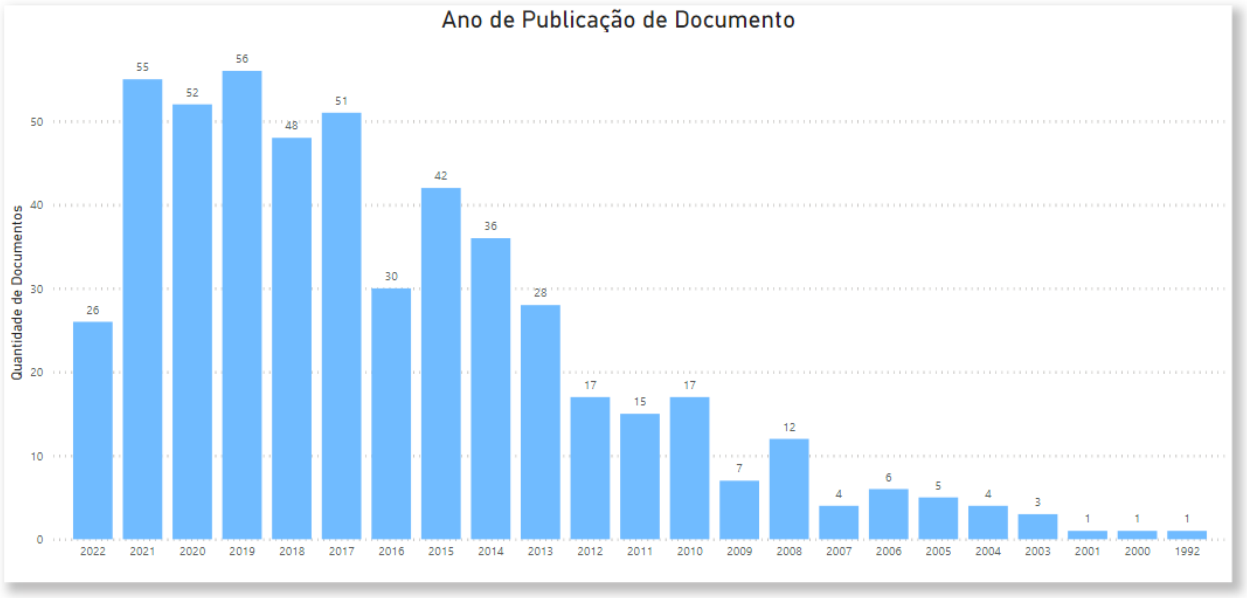


Figura III.5: Quantidade de Documentos por Ano de Publicação - Base Scopus.

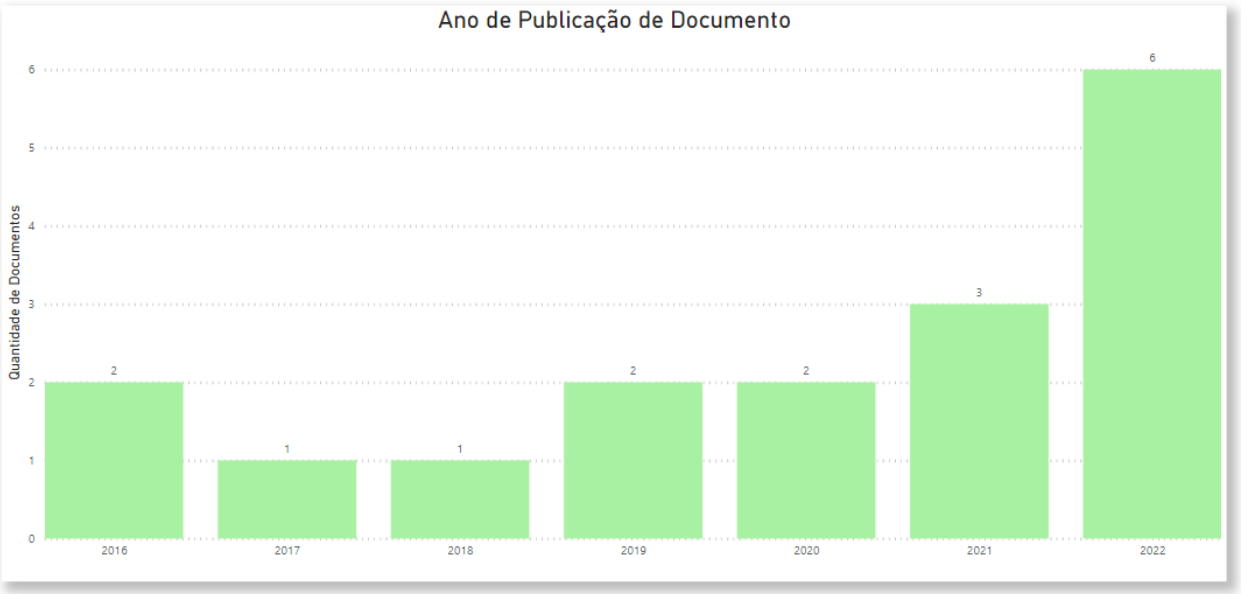


Figura III.6: Quantidade de Documentos por Ano de Publicação - Base PubMed.

- as Figuras III.7, III.8 e III.9 exibem as nuvens de palavras contidas nos títulos dos documentos de ambas as bases e resumos dos documentos da SCOPUS. São apresentadas as palavras mais significativas deste *corpus* que podem trazer diversas *insights*, algumas são descritas para cada figura. Observa-se que a palavra mais encontrada foi "*adverse drug*" nos títulos das duas bases; "*method*" e "*system*" no resumo da base Scopus. O PubMed não tem coluna de resumo, por esse motivo não foi gerado nuvem de palavras de resumo.

  A nuvem de palavras referente a Figura III.7 do *corpus* dos títulos dos documentos da base Scopus, revela uma grande variação de termos para EA, tais como: *adverse drug, drug event,*

drug reaction, helth record e *eletronic health*. Com estas variações de EA, a *string* de busca pode ser refinada para alcançar mais documentos. Nota-se que os termos referente a ferramenta MetaMap e a base UMLS, termos relevantes para este trabalho, foi quase irrelevante neste *corpus*.

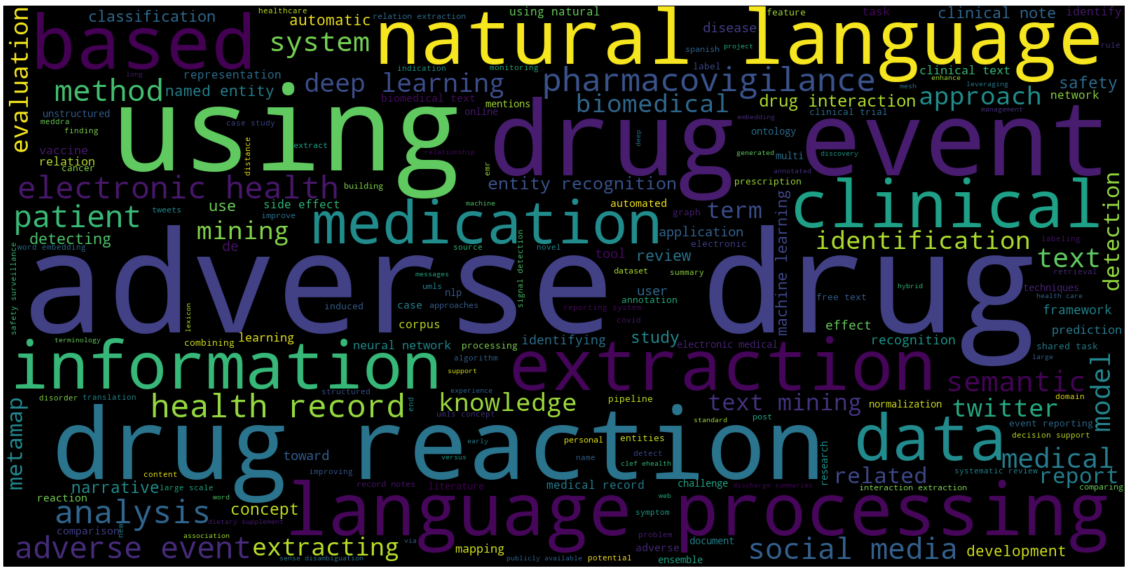


Figura III.7: Nuvem de Palavras nos Títulos dos Documentos - Base Scopus.

A nuvem de palavras referente a Figura III.8 do *corpus* dos títulos dos documentos da base PubMed, revela termos para EA similares a base Scopus, contudo, com uma relevância bem menor. Apesar desta base de indexação ser específica da área biomédica, a *string* de busca trouxe mais termos da área de tecnologia do que médicos, logo, deve-se pensar se manter a mesma *string* de busca para diversas bases faz sentido. Nota-se que termos referentes PLN foi bem mais relevante neste *corpus*.

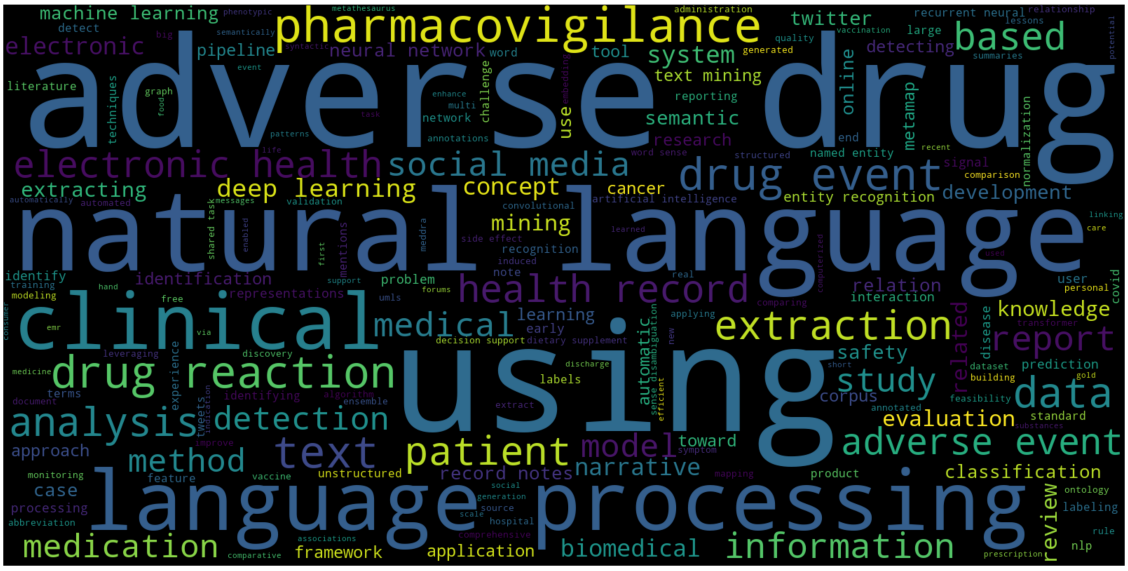


Figura III.8: Nuvem de Palavras nos Títulos dos Documentos - Base PubMed.

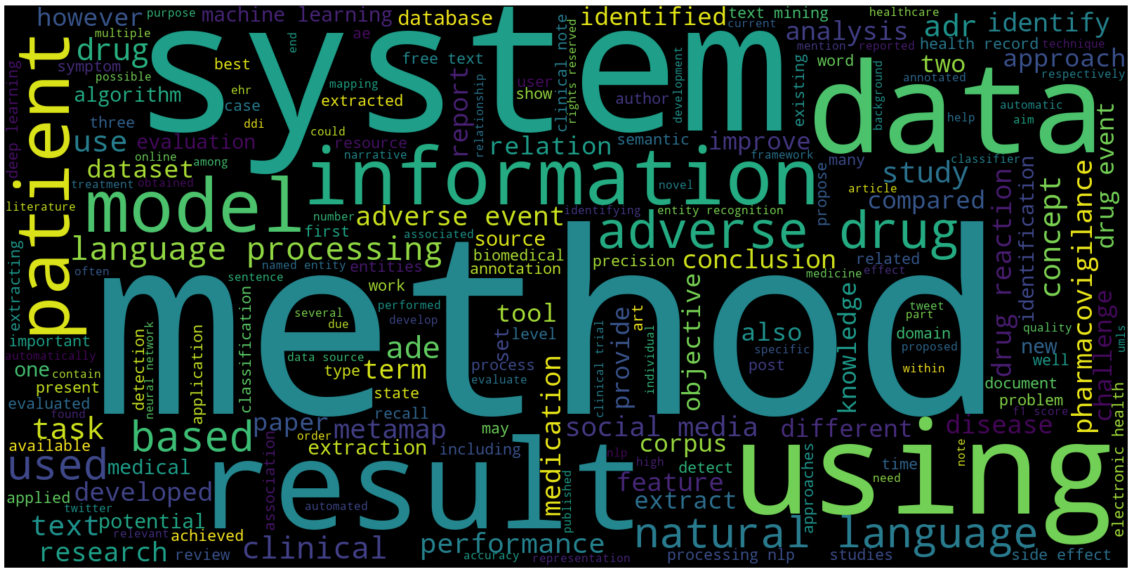


Figura III.9: Nuvem de Palavras nos Resumos dos Documentos - Base Scopus.

- A nuvem de palavras referente a Figura III.10 do *corpus* dos títulos dos documentos de ambas as bases indexadoras de documentos, trazem uma concatenação de ambas nuvens de palavras anteriores. Contudo, nota-se que os termos referentes às diversas técnicas de PLN aparecem bastante neste *corpus*. O termo MetaMap está um pouco mais visível, mas a base médica UMLS quase não aparece, geralmente os trabalhos que utilizam o sistema MetaMap, usam o UMLS. Talvez seja relevante colocar estes termos na *string* de busca para procurar trabalhos que de fato usaram o sistema e não tenha apenas menção de um destes termos. Este caso foi encontrado neste trabalho, tiveram diversos documentos que apenas mencionavam a

ferramenta, porém não havia um trabalho ou experimento com a mesma.



Figura III.10: Nuvem de Palavras nos Títulos dos Documentos em Inglês.

Após a Etapa de Visualização, ocorre a Exportação de Arquivo que teve como resultado o gráfico da Figura III.11. A barra do eixo X com o valor (Em branco), é referente aos documentos que ainda precisam ser lidos e classificados conforme os *status* pré-definidos.

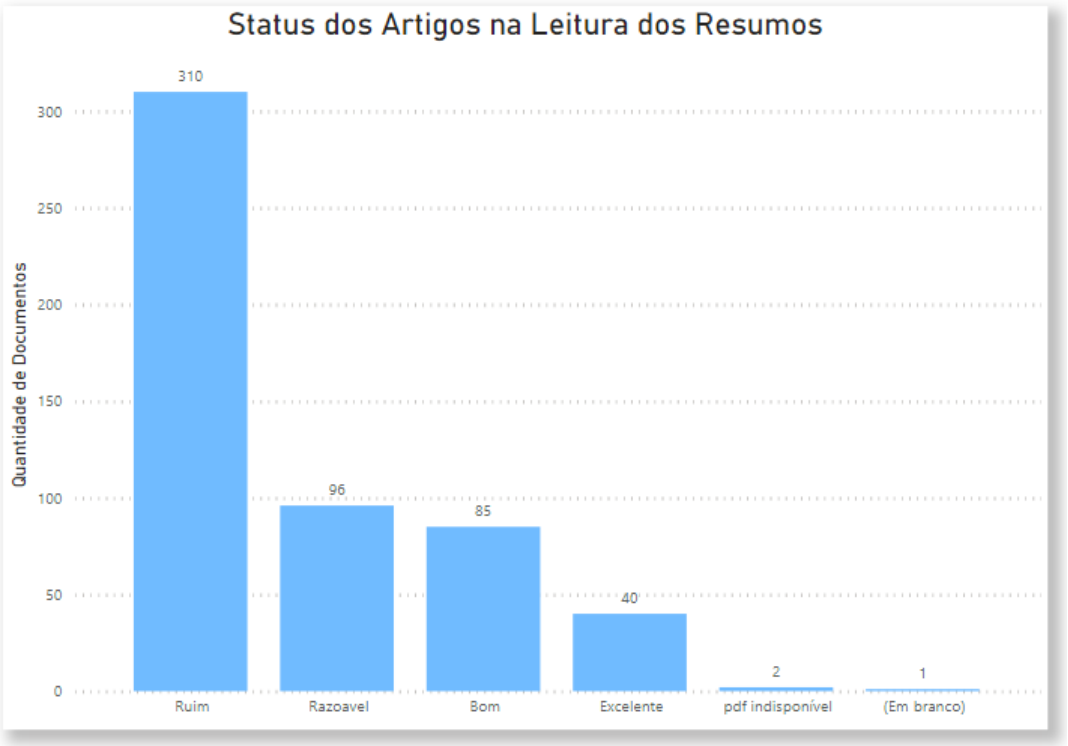


Figura III.11: Documentos por *Status*.

Os artigos cujo *status* são classificados como "Excelente", são altamente relacionados com a

proposta deste trabalho. Esta abordagem tem como um dos propósitos principais, utilizar a ferramenta MetaMap para o português do Brasil. Logo, o principal trabalho correlacionado, é o trabalho de [Aronson, 2001].

No trabalho de Aronson [Aronson, 2001] é explicado o funcionamento do MetaMap, os processos de PLN utilizados e exemplos de mapeamentos com alguns textos e os resultados obtidos pela ferramenta. Logo, em quase toda pesquisa sobre PLN em texto biomédico, este trabalho é obtido como resultado na *string* de busca.

A PCLD mostra que o trabalho de Aronson [2001] possui 1339 citações na base Scopus, o mais citado desta lista de trabalhos. Complementando a pesquisa ou validando os resultados retornados, foi colocado o trabalho de Aronson [2001] na ferramenta *Connected Papers*[4] que exibiu um gráfico com outros artigos similares. O gráfico mostra que parte dos documentos exibidos, também estão contidos na PCLD. Comparando ambos resultados das pesquisas da PCLD e do *Connected Papers*, é possível verificar que um tem documentos que o outro não tem e vice-versa. Caso o estudo precise de artigos mais recentes ou com maior número de citação, é fácil visualizar e escolher com o *Connected Papers*.

A Figura III.12 exibe o resultado quando inserido o trabalho de Aronson [2001] na ferramenta. Esta apresenta um grafo, onde o trabalho possui 2.112 citações. Nesse grafo, os nós representam os artigos e as arestas sua similaridade com os outros artigos. O tamanho do nó representa a quantidade de citações, quanto mais forte a cor de um nó, significa que este é mais recente.

[4]O *Connected Papers* é uma ferramenta visual que auxilia pesquisadores a encontrar e explorar artigos acadêmicos relacionados ao seu campo de estudo. Ela cria gráficos baseados na similaridade entre artigos, utilizando conceitos como co-citação e acoplamento bibliográfico para identificar trabalhos relevantes. Disponível em: `https://www.connectedpapers.com`.

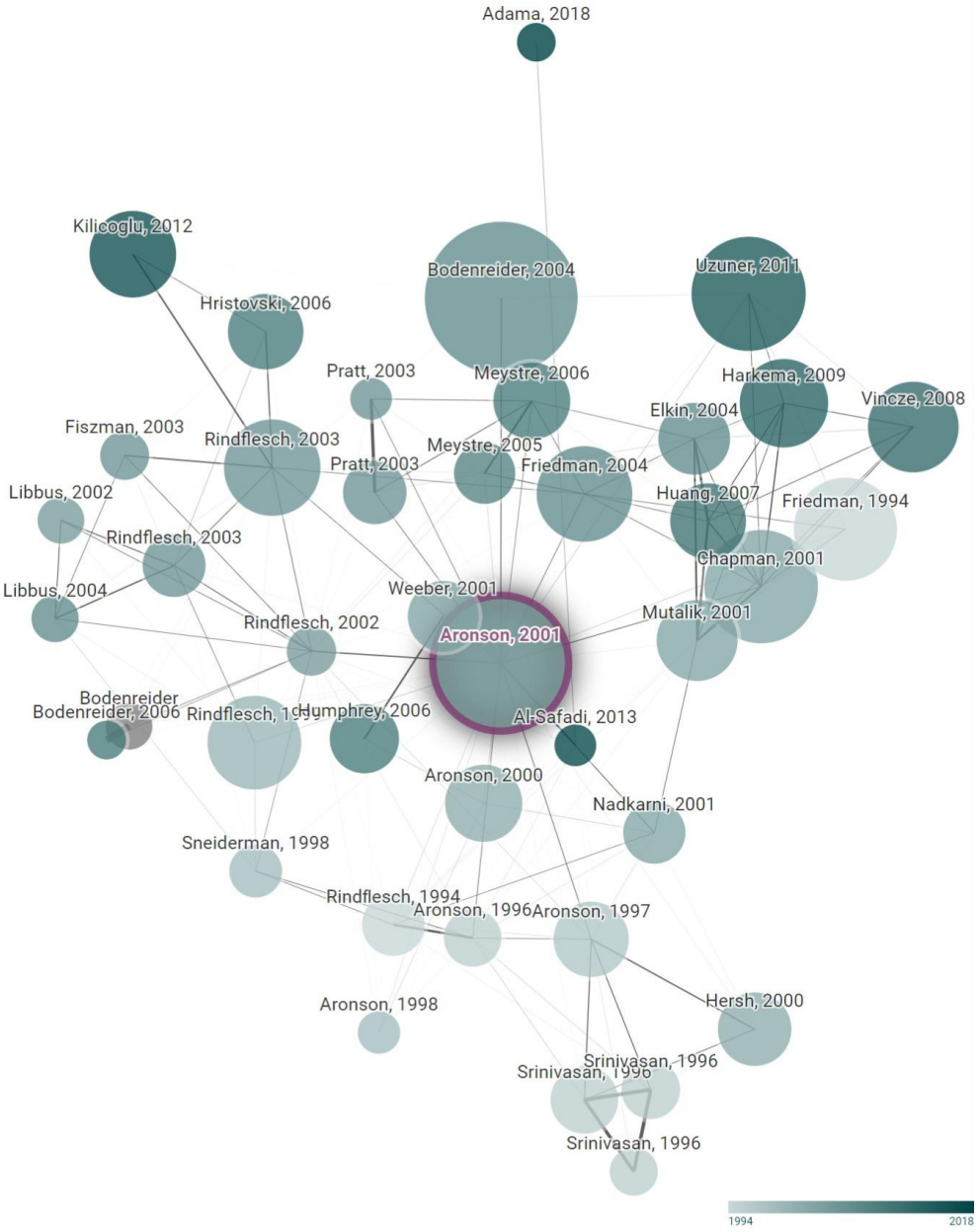


Figura III.12: Correlação do Documento de [Aronson, 2001][5].

Foram poucos artigos encontrados que utilizaram uma ferramenta de PLN da área biomédica para trabalho diretamente em sua língua mãe distinta do inglês. Os artigos geralmente falam da dificuldade do processamento e assertividade na utilização da ferramenta em línguas diferentes do inglês [Carrero et al., 2008]. Em português do Brasil, não foi encontrado nenhum trabalho utilizando especificamente o MetaMap ou outra ferramenta de processo linguístico da área biomédica, exceto, recentemente o trabalho de [Perciliano et al., 2023], oriundo desta pesquisa.

No mundo hoje, há cerca de 20 idiomas distintos mais falados dentre mais de 7.000 existentes [Ethnologue, 2022]. O conjunto de artigos deste trabalho, trata por volta de cinco idiomas. Cada pesquisa precisa utilizar ou criar uma ferramenta para processar os termos médicos de sua língua

[5]https://www.connectedpapers.com/main/1406f6b5ed4034b72ed2dccc3fcfa4c5c0810924/graph?utm_source=share_popup&utm_

específica, ou em alguns casos, traduzem para o inglês e utilizam ferramentas padrão da indústria que trabalha com inglês.

Existem diversas ferramentas multilíngues de PLN para termos médicos, grande parte dos artigos consultados neste trabalho usaram o MagiCoder, *Clinical Language Annotation, Modeling, and Processing Toolkit* (Kit de Ferramentas de Anotação, Modelagem e Processamento de Linguagem Clínica) (CLAMP) e *clinical Text Analysis and Knowledge Extraction System* (Sistema de Extração de Conhecimento de Análise de Texto Clínico) (cTAKES). A última é gratuita e usa a base UMLS, contudo, vários trabalhos de outras línguas diferentes do inglês, usaram um processo anterior de tradução. O MagiCoder tem sua maior precisão para o italiano, 91%, apesar de conter o vocabulário em português, até o momento não foi encontrado um trabalho que utilize o vocábulo em português.

III.2.2 Panorama das Abordagens e Desafios em Farmacovigilância

Esta seção visa fornecer uma visão das abordagens contemporâneas e desafios enfrentados pelos pesquisadores e profissionais da área separando os trabalhos em categorias. As categorias foram escolhidas com base na relevância para o tema da dissertação e na necessidade de abordar diferentes aspectos metodológicos e aplicados da pesquisa em farmacovigilância. A partir do documento PCLD, os trabalhos mais alinhados com esta pesquisa, foram distribuídos nas categorias das três subsecções abaixo.

III.2.2.1 Desenvolvimento de Ferramentas e Recursos

Esta subseção descreverá trabalhos que propuseram novas ferramentas, algoritmos, técnicas ou recursos especificamente para apoiar aplicações biomédicas. Isso inclui, mas não se limita a: sistemas de diagnóstico, ferramentas de análise de dados e plataformas de monitoramento de EAs. O objetivo é entender como esses desenvolvimentos podem auxiliar profissionais e pesquisadores na coleta e análise de dados, bem como na tomada de decisões clínicas.

O trabalho de Aronson [2001] aborda o MetaMap, um programa desenvolvido na Biblioteca Nacional de Medicina (*National Library of Medicine* (Biblioteca Nacional de Medicina) (NLM)) dos Estados Unidos, que mapeia textos biomédicos para o *Metathesaurus* UMLS. O MetaMap utiliza uma abordagem intensiva de conhecimento baseada em PLN e técnicas de linguística computacional. O objetivo principal é fornecer acesso eficaz ao conhecimento biomédico armazenado no *Metathesaurus* da UMLS, o que é crucial para o sucesso de várias aplicações, incluindo sistemas de apoio à decisão, gerenciamento de registros de pacientes, *Information Retrieval* (Recuperação de Informações) (IR) e mineração de dados.

O método utilizado neste programa envolve várias etapas, incluindo o *parsing* do texto em frases

substantivas simples, geração de variantes de palavras, recuperação e avaliação de candidatos do *Metathesaurus* e, finalmente, construção de mapeamentos completos. O programa é altamente configurável, permitindo que os usuários controlem tanto a saída quanto o comportamento interno, como a geração de variantes de palavras e a ordem das palavras.

Os resultados mostram que o MetaMap é eficaz na descoberta de conceitos do *Metathesaurus* em textos. Ao longo do tempo vem sendo atualizado resolvendo ambiguidades, melhoraria na precisão incluindo o desenvolvimento de um regime de tokenização extensível e hierárquico e métodos para desambiguação de sentido de palavra. Este trabalho é o mais importante desta pesquisa e de outros trabalhos que o utiliza.

O trabalho de Meystre and Haug [2005] tem como principal objetivo melhorar a qualidade e eficácia da "Lista de Problemas Médicos Eletrônica". Para atingir esse objetivo, os autores propuseram um sistema automatizado denominado "*Automated Problem List*" (APL), que visa extrair problemas médicos de documentos clínicos eletrônicos. O método utilizado envolve o uso de técnicas de PLN.

O sistema APL emprega uma ferramenta chamada MetaMap Transfer (MMTx), para mapear termos médicos em texto livre para conceitos padronizados. Além disso, o sistema utiliza um algoritmo de detecção de negação chamado NegEx para identificar a ausência de determinados problemas médicos nos registros. A ideia é que um médico ao preencher um problema no prontuário de um paciente tenha sugerido pelo sistema um problema existente em uma lista, com possibilidade de criar ou editar itens para a lista, evitando assim a escrita em texto livre.

O desempenho do sistema foi avaliado utilizando um conjunto de 160 documentos clínicos eletrônicos. O sistema alcançou um *recall* de 0,9 e uma precisão de 0,69 quando testado com um conjunto de dados personalizado. Esses resultados indicam que o sistema é eficaz na identificação de problemas médicos, superando o desempenho quando comparado ao conjunto de dados padrão completo do UMLS. As listas foram validadas por cinco médicos com quase unanimidade com os resultados e os prontuários preenchidos com mais agilidade e assertividade.

O trabalho de Osborne et al. [2006] aborda o mapeamento de dados de texto não estruturado em conceitos biomédicos comuns, como drogas, doenças e anatomia, utilizando o UMLS e o MetaMap. O método empregado envolve várias etapas. Inicialmente, o módulo de tokenização organiza o documento de entrada em seções compostas por conjuntos de sentenças e *tokens*.

Em seguida, o *Part of Speech Tagger* identifica a classe gramatical de cada *token*. Estes *tokens* são então submetidos ao LexicalLookup, o módulo que determina se algum dos elementos marcados pertence a um léxico específico. Finalmente, um analisador de frases nominais identifica frases nominais desses elementos, para as quais variantes são calculadas. Estas variantes são então usadas para identificar *strings* correspondentes no *Metathesaurus* do UMLS.

Os resultados obtidos com o uso do MetaMap são amplos e podem ser aplicados em diversas áreas, como mineração de dados e inferência de relações entre conceitos em publicações do *Medical Literature Analysis and Retrievel System Online* (Sistema Online de Análise e Recuperação de Literatura Médica) (MEDLINE)[6]. O artigo também fornece instruções detalhadas para usuários sem conhecimento de programação, tornando a ferramenta acessível para uma gama mais ampla de pesquisadores.

O principal objetivo do trabalho de [Kim et al., 2017] é abordar a escassez de dados anotados manualmente em vários subdomínios biomédicos. Para isso, foi proposto um método para gerar uma grande quantidade de dados rotulados por máquina que podem ser utilizados para treinar modelos de *Machine Learning* (Aprendizado de Máquina) (ML) em tarefas de "Biomedical NER".

O método proposto utiliza uma combinação de um *"chunker"* e o MetaMap para gerar dados rotulados inicialmente. O *"chunker"* é desenvolvido para Reconhecimento de Entidade Nomeada (REN) biomédicas, enquanto o MetaMap é utilizado para anotar as categorias dessas entidades. Posteriormente, uma abordagem de auto-treinamento é aplicada para refinar e melhorar a qualidade dos dados rotulados pela máquina.

Os experimentos realizados demonstram a eficácia do método proposto. O sistema "Biomedical NER" treinado com os dados gerados alcançou um desempenho notavelmente alto, superando o sistema que utiliza apenas o MetaMap. Especificamente, houve uma melhoria de 26,03% no F1-score. Este trabalho representa um avanço significativo, oferecendo uma solução robusta para o desafio de falta de dados anotados manualmente em vários subdomínios.

O trabalho de Herrero-Zazo et al. [2013] teve como objetivo construir o "DDI corpus", um *corpus* anotado manualmente para extração de interações droga-droga (DDIs) em textos biomédicos. O método consistiu em compilar o *corpus* a partir de 792 textos do DrugBank e 233 resumos MEDLINE, totalizando 18.502 entidades farmacológicas e 5.028 DDIs anotadas. Foram definidos quatro tipos de entidades e quatro tipos de relações DDI.

As diretrizes de anotação foram elaboradas de forma iterativa e a qualidade do *corpus* foi avaliada medindo a concordância inter-anotadores, atingindo um Kappa 0,80. O resultado foi a construção de um *corpus* anotado de alta qualidade, rico semanticamente, incluindo DDIs farmacocinéticas e farmacodinâmicas. O *corpus* foi utilizado no desafio SemEval 2013 DDIExtraction como padrão ouro para treinamento e avaliação de sistemas de IR. Os recursos desenvolvidos estão disponíveis gratuitamente para pesquisa acadêmica e podem facilitar o avanço de ferramentas de PLN para farmacovigilância.

[6] `https://www.nlm.nih.gov/medline/medline_home.html`

III.2.2.2 Processamento de Linguagem Natural (PLN) em Idiomas Não-Ingleses

O foco desta subseção é em trabalhos que aplicam técnicas PLN em idiomas que não são o inglês. Dada a predominância do inglês na literatura científica e em bases de dados biomédicas, é crucial explorar como o PLN pode ser adaptado ou aplicado em outros idiomas. Isso é especialmente relevante para o contexto brasileiro, onde o idioma predominante é o português e há poucos vocabulários médicos em português.

O trabalho de Yokotsuka et al. [2000] aborda a implementação e os desafios do uso do *Medical Dictionary for Regulatory Activities* (Dicionário Médico para Atividades Regulatórias) (MedDRA) no monitoramento de eventos de prescrição no Japão, também conhecido como "J-PEM". O objetivo principal do estudo é discutir as estratégias e desafios para o uso eficaz do MedDRA no J-PEM, especialmente em relação à entrada de dados, geração de sinais e listagem de eventos.

Para a entrada de dados no J-PEM, recomenda-se o uso de *Lowest Level Terms* (LLTs), apesar de alguns desafios específicos ao ambiente japonês. Em relação à geração de sinais, o artigo argumenta que os dados no J-PEM devem ser geralmente analisados no nível de *Preferred Term* (PT). No entanto, os autores reconhecem que alguns PTs podem não ser adequados para a geração de um sinal apropriado.

O estudo também aborda as condições que devem ser atendidas ao listar eventos codificados pelo MedDRA. Além disso, discute-se o uso de categorias de pesquisa especiais, conhecidas como *Special Search Categories* (SSCs), na análise. Por fim, o artigo delineia os problemas encontrados ao usar o MedDRA para gerar sinais e listar eventos no J-PEM, descrevendo as estratégias atuais para abordar essas questões. O trabalho serve como um guia útil para a implementação do MedDRA em sistemas de monitoramento de eventos de prescrição.

O objetivo do trabalho Merabti et al. [2011] é avaliar dois métodos lexicais em inglês e francês para mapear a Classificação *Anatomical Therapeutic Chemical* (Classificação Anatômica, Terapêutica e Química) (ATC) ao *Metathesaurus* UMLS. Os métodos lexicais usados foram: uma ferramenta de PLN francesa para mapear termos franceses da ATC aos termos franceses do UMLS e a ferramenta MetaMap para mapear termos ingleses da ATC ao UMLS. O MetaMap em inglês forneceu ligeiramente mais mapeamentos do que a ferramenta francesa (3.170 vs 2.992). Por outro lado, a ferramenta em francês teve uma precisão um pouco melhor do que o MetaMap (88% vs 86%).

Usando um mapeamento manual entre ATC e MeSH, a união dos mapeamentos validados entre ATC e MeSH resulta em 2.824 mapeamentos (68,7% dos códigos ATC do quinto nível). Métodos lexicais são poderosos para mapear terminologias de saúde ao *Metathesaurus* UMLS. Mapeamento manual ainda é necessário para completar o mapeamento. Três aplicações práticas foram desenvolvidas usando esse mapeamento ATC-UMLS: tradução MeSH para norueguês, Portal de Informações de Medicamentos do PSIP e uma ferramenta ATC para PubMed. Em resumo, eles

mapearam a classificação ATC ao UMLS usando ferramentas lexicais automatizadas com bons resultados e ressaltam que o mapeamento manual ainda se faz necessário.

O trabalho Classificação Comum de Atos Médicos (CCAM) do Bousquetab et al. [2012], é um sistema francês de codificação de procedimentos cirúrgicos usado em bancos de dados e pagamento por serviços. Mapear a CCAM a outras terminologias de procedimentos clínicos por meio do *Metathesaurus* UMLS é essencial para aumentar a interoperabilidade semântica.

Um método anterior mapeou descritores CCAM ao UMLS francês usando a ferramenta FMTI+. Este artigo propôs um novo método onde descritores CCAM são traduzidos do francês para o inglês e então mapeados ao UMLS usando o *Metathesaurus*. O *Metathesaurus* forneceu mapeamentos para 62% dos descritores de ação e 49% dos descritores anatômicos, melhorando significativamente os resultados do FMTI+.

A combinação de ambas abordagens, aumentaram a cobertura do mapeamento para 68% para ambos descritores e 95% para pelo menos um descritor. Em resumo, foi proposto um novo método de mapeamento automatizado da CCAM ao UMLS com bons resultados iniciais. A combinação de abordagens pode aprimorar a interoperabilidade semântica entre terminologias.

O trabalho de Hellrich et al. [2015] apresenta o JUFIT, um motor de regras configurável destinado a filtrar e gerar novos termos para o UMLS. Ferramentas existentes como o MetaMap e o *CASPER* são predominantemente em inglês e não oferecem soluções independentes. O objetivo principal do JUFIT é eliminar termos "não-naturais" do UMLS e adicionar novos termos através da reescrita desses termos não-naturais. Além disso, o JUFIT é multilíngue, oferecendo suporte aos idiomas inglês, espanhol, francês, alemão e holandês.

Em termos de metodologia, o JUFIT utiliza sistemas baseados em regras para transformações de termos. Ele possui tanto regras de exclusão, como *DeleteIfContainsAtSign* e *DeleteIfContains-Dosage*, quanto regras de reescrita, como *RewriteApostropheS* e *RewriteSyntacticInversion*. A avaliação do JUFIT foi realizada usando o *corpus* EMEA, mostrando um aumento na qualidade da anotação para os idiomas inglês, alemão e espanhol. Também foi observado um aumento moderado na cobertura da anotação para os idiomas francês e holandês. Em comparação com o CASPER, o JUFIT produz menos termos, porém, consegue mais de anotações. Além disso, ao contrário do CASPER, o JUFIT é de código aberto.

O trabalho de Casillas et al. [2016b] foca no desenvolvimento e transferência de tecnologias de PLN para o domínio clínico, especificamente para auxiliar na detecção de RAM em registros de saúde eletrônicos. O projeto Deteami foi estabelecido em colaboração com os hospitais de Galdakao-Usansolo e Basurto, fornecendo um *corpus* de dados reais e anotações de especialistas para criar um padrão-ouro.

O projeto utiliza uma combinação de técnicas baseadas em regras e ML para identificar entidades

clínicas e extrair informações relevantes. Além disso, o projeto também aborda o desafio da detecção de negação nos registros médicos, uma etapa crucial para descartar pares potenciais de droga-doença que não indicam RAM. O trabalho é significativo não apenas para a comunidade de PLN espanhola, mas também para os serviços de farmácia e farmacovigilância, contribuindo para a detecção precoce de RAM e, consequentemente, para a segurança do paciente.

O trabalho de Casillas et al. [2016a] aborda a extração de eventos de RAMs a partir de Registros Eletrônicos de Saúde (EMR) em espanhol. O objetivo principal é desenvolver um sistema híbrido que combine técnicas de aprendizado de máquina ML e regras linguísticas para identificar eficazmente esses eventos. O desafio adicional é lidar com dados altamente desbalanceados, onde apenas 6% dos pares entidade-doença resultam em um evento de RAM positivo.

O método proposto envolve duas etapas principais. A primeira é o Reconhecimento de Entidades Nomeadas (REN), que identifica drogas e doenças nos EMR. A segunda é a extração de RAM, que é realizada através de um sistema híbrido. Este sistema híbrido utiliza tanto regras linguísticas quanto algoritmos de ML para classificar os pares entidade-doença como RAM ou não.

Os resultados mostraram que o sistema baseado em ML é capaz de lidar com eventos tanto intra-sentença quanto inter-sentença em um ambiente de classificação enviesado. O sistema foi projetado para trabalhar com EMR reais em espanhol e tem o potencial de aliviar a carga de trabalho de diversos serviços hospitalares, como farmácia e documentação, além de melhorar a segurança do paciente ao detectar RAM de forma rápida e eficaz.

O trabalho contribui significativamente para a área de PLN aplicado à saúde, fornecendo uma solução robusta para o problema de extração RAM em EMR, particularmente em espanhol, uma língua menos estudada em comparação com o inglês neste contexto.

O trabalho de Vatian et al. [2018] aborda um tema de relevância crescente na área de Farmacovigilância: a detecção de RAM em textos da língua russa. O estudo se concentra em quatro etapas principais para a detecção de RAMs, que incluem a coleta de dados de mídias sociais, a classificação e filtragem de segmentos de texto que contêm menções de RAM, a extração dessas menções e, por fim, a análise para geração de sinais.

O método empregado pelos autores envolve o uso de ML e adaptação de algoritmos já existentes para o contexto da língua russa. Especificamente, o algoritmo "pyConTextNLP" foi adaptado para tratar textos em russo. A ausência de bancos de dados e recursos linguísticos especializados para o russo foi identificada como uma barreira significativa, tornando a adaptação dos algoritmos uma necessidade.

Em termos de resultados, o estudo alcançou uma precisão de 0,805 na etapa de extração de menções de RAM e uma medida F de 0,935 na etapa de análise para geração de sinais. Esses resultados indicam a eficácia do algoritmo adaptado, abrindo caminho para futuras investigações

que possam utilizar essa abordagem em outros idiomas ou contextos.

III.2.2.3 Análise de Dados Não-Estruturados em Saúde

A última subseção concentra-se em trabalhos que lidam com a análise de dados não-estruturados ou semi-estruturados no contexto da saúde. Isso inclui o uso de técnicas de *Text Mining* (Mineração de Texto) (MT), ML e outras abordagens de análise de dados para extrair informações úteis de fontes como *Electronic Health Record* (Registro Eletrônico em Saúde) (EHR), EMR, redes sociais e outras plataformas digitais.

O objetivo do trabalho de Brennan and Aronson [2003] é explorar a viabilidade de detectar termos em mensagens eletrônicas de pacientes. Foram aplicadas as ferramentas de PLN do Lister Hill Center no NLM para detectar conceitos relevantes do UMLS no texto livre de *e-mails* de pacientes. Foram testados vocabulários de enfermagem e clínicos do UMLS como fontes para mapeamento de termos.

O melhor desempenho foi com vocabulários de enfermagem complementados por terminologias clínicas selecionadas. A abordagem permitiu detectar conceitos padronizados relevantes no texto livre dos pacientes. Mas ainda são necessários pré e pós-processamento adicionais para aprimorar a precisão do mapeamento. A detecção de conceitos pode facilitar o acesso de pacientes a recursos de conhecimento eletrônicos indexados com vocabulários controlados.

O trabalho de Ruiz [2006] propôs uma abordagem para recuperação de imagens médicas. O objetivo principal do trabalho é avaliar a eficácia do sistema UBMedTIRS, que integra características de imagem e texto, mapeando texto livre para conceitos do UMLS.

O método utilizado combina técnicas de Recuperação de Imagens Baseada em Conteúdo (CBIR) com recuperação de texto e mapeamento automático de texto livre para uma ontologia médica. O sistema UBMedTIRS é composto por três ferramentas publicamente disponíveis: GIFT para recuperação de imagem, SMART para recuperação de texto e o MetaMap. O sistema foi avaliado utilizando a coleção ImageCLEFmed 2005, que contém aproximadamente 50.000 imagens médicas com descrições de texto associadas em inglês, francês e alemão.

Os resultados experimentais apresentados demonstraram que a combinação de características de texto e imagem melhora significativamente a precisão média na recuperação de imagens médicas. O sistema UBMedTIRS superou outros métodos, mostrando uma melhoria de 156% em relação ao sistema GIFT e uma melhoria de 42% em relação aos sistemas de recuperação baseados em texto. Além disso, o sistema se mostrou eficaz em coleções multilíngues em inglês, francês e alemão.

O trabalho de Zuccon et al. [2013] abordou a identificação de distúrbios em EMR, um tópico de grande relevância na área de PLN aplicado à saúde. O estudo propõe uma abordagem híbrida que combina MetaMap e Campos Aleatórios Condicionais (CRF) para mapear distúrbios para conceitos

SNOMED CT no *Metathesaurus* UMLS. O objetivo principal é melhorar o *recall* na identificação desses distúrbios, um desafio que muitos sistemas atuais enfrentam.

A metodologia adotada envolve duas etapas principais. A primeira utiliza o MetaMap para identificar conceitos médicos nos EMR. A segunda aplica CRF para refinar essas identificações, levando em consideração o contexto em que os termos aparecem. Essa abordagem híbrida é comparada com um método que utiliza apenas o MetaMap, servindo como um "baseline" para avaliação.

Os resultados empíricos mostram que a combinação de MetaMap e CRF oferece um *recall* significativamente mais alto, em comparação com o uso apenas do MetaMap. Especificamente, a abordagem híbrida alcançou um aumento de aproximadamente 20% no *recall*. Isso sugere que a inclusão de CRF pode ser uma estratégia eficaz para melhorar a identificação de distúrbios em EMR.

III.2.2.4 Análise Comparativa com Trabalhos Correlatos

Diversos trabalhos de revisão sistemática abordaram a aplicação de técnicas de PLN e ferramentas como o MetaMap na área de farmacovigilância. A seguir, são analisados alguns estudos relevantes, destacando seus pontos fortes e fracos, e comparando-os com a presente pesquisa.

O trabalho de de revisão sistemática de Feng et al. [2019] apresentado neste estudo abordou o uso de EHRs para a identificação de RAMs no ambiente ambulatorial, destacando métodos variados, como revisão manual de prontuários, monitores computadorizados e alertas eletrônicos. Um ponto forte do trabalho é sua análise detalhada dos métodos disponíveis, evidenciando a transição de abordagens manuais para o uso de ferramentas eletrônicas mais avançadas, o que reforça o potencial dos EHRs para melhorar a segurança do paciente. Além disso, a revisão foca nos métodos de detecção em vez de apenas caracterizar as RAMs, proporcionando uma base para o desenvolvimento de soluções futuras. No entanto, o estudo apresenta limitações importantes, como a exclusão de literatura publicada em outras fontes além do *PubMed* e *Google Scholar*, o que pode gerar viés de publicação. Além disso, a ausência de padronização na documentação dos EHRs e a dependência de abordagens retrospectivas limitam a capacidade de prevenir RAMs antes que danos ocorram. Apesar dessas limitações, o trabalho fornece *insights* valiosos e destaca a necessidade de estudos futuros que explorem ferramentas em tempo real e padronização de documentação para maximizar os benefícios dos EHRs na detecção de RAMs. Em contraste, o presente trabalho se diferencia na possibilidade de incorporar métodos híbridos de detecção de RAMs, que combinam aprendizado supervisionado e não supervisionado.

O estudo conduzido por Pilipiec et al. [2022] abordou o uso de PLN para a identificação de RAMs por meio da análise de conteúdo textual gerado por usuários na Internet. Um ponto forte deste estudo é a inovação ao propor o uso de PLN como uma abordagem suplementar para a

farmacovigilância, destacando seu potencial para reduzir custos e tempo em comparação com os métodos tradicionais. Além disso, a revisão compila evidências consistentes de 16 publicações, apontando resultados positivos em 14 delas, o que destaca a eficácia do PLN na identificação de RAMs.

Por outro lado, o estudo apresenta algumas limitações significativas. Primeiramente, todos os estudos incluídos foram avaliados como tendo confiabilidade e validade médias, o que pode comprometer a robustez das conclusões. Adicionalmente, a revisão não aborda de forma detalhada as limitações éticas e técnicas associadas ao uso de dados gerados por usuários, como questões de privacidade e anonimato. Outro ponto é a ausência de uma análise prática ou arquitetural mais aprofundada sobre como implementar as ferramentas de PLN, limitando a aplicabilidade imediata das descobertas. Apesar disso, o estudo é um marco importante na exploração de métodos automatizados para farmacovigilância e oferece uma base sólida para pesquisas futuras que busquem otimizar e validar essas abordagens em contextos reais.

O trabalho de revisão sistemática apresentado por Young et al. [2019] abordou o uso do PLN na classificação de relatórios de incidentes e EAs na área da saúde, com foco na capacidade de automatizar a análise de textos livres. Um dos principais méritos do estudo está na abrangência de sua revisão, que incluiu uma ampla variedade de bases de dados e métodos de extração, garantindo uma visão geral significativa sobre o estado atual da pesquisa. Além disso, o trabalho destaca a eficácia de técnicas de PLN em tarefas específicas, como a classificação de tipos de incidentes e a detecção de erros de medicação, demonstrando resultados favoráveis em desempenho. Outro ponto positivo é a ênfase na aplicação prática do PLN, especialmente no contexto de grandes volumes de dados, o que reforça a relevância do trabalho para a área da saúde. Apesar da riqueza dos dados apresentados, o estudo enfrenta desafios metodológicos que limitam a generalização de seus resultados. A análise não aprofunda suficientemente a comparação entre diferentes abordagens de PLN, nem aborda questões relacionadas à variabilidade dos dados de entrada. Além disso, questões éticas, como a privacidade e a confidencialidade dos dados analisados, não são discutidas de forma detalhada, o que poderia enriquecer a análise. O estudo fornece uma base valiosa para o desenvolvimento de soluções futuras que utilizem PLN na análise de EA, mas sugere a necessidade de pesquisas adicionais que explorem a padronização dos métodos e abordem lacunas éticas e metodológicas.

O estudo de Vo et al. [2022] apresenta uma revisão abrangente sobre modelos de Inteligência Artificial (IA) para predição de interações medicamentosas (DDIs), destacando avanços significativos na área e o papel crescente da IA explicável (XAI). Um dos pontos fortes do trabalho é a análise detalhada de técnicas de manipulação de dados e seleção de características, evidenciando o progresso tecnológico e a aplicação de metodologias modernas, como redes neurais profundas, na

extração de DDIs. Além disso, o texto aborda de forma clara as limitações dos métodos tradicionais de aprendizado de máquina, como a dificuldade em lidar com dados não rotulados e a falta de um "padrão-ouro" para amostras negativas, trazendo *insights* sobre as vantagens de abordagens baseadas em aprendizado profundo.

Por outro lado, o trabalho ressalta a falta de explicabilidade como uma das principais barreiras para a adoção clínica de modelos de IA, destacando que os sistemas mais sofisticados frequentemente sacrificam a transparência para alcançar maior precisão preditiva. Essa crítica é bem fundamentada e direciona para uma necessidade urgente de desenvolvimento de modelos mais interpretáveis, o que é particularmente relevante em áreas críticas como saúde. Outra limitação importante apontada é a carência de dados estruturados e anotações padronizadas, o que dificulta a validação e replicação de estudos.

O texto fornece um panorama sólido dos desafios e oportunidades no campo, enfatizando que a integração de mecanismos de XAI é essencial para aumentar a confiança e aceitação dos modelos preditivos em contextos médicos. No entanto, uma limitação da revisão é a ausência de uma discussão mais aprofundada sobre como diferentes fontes de dados textuais, como informações de patentes ou dados semiestruturados, poderiam ser utilizadas de forma mais eficaz para enriquecer os modelos. Ainda assim, a revisão contribui significativamente para a literatura ao estabelecer uma taxonomia detalhada de métodos existentes e sugerir caminhos promissores para pesquisas futuras.

Por fim, o artigo de Jeiziner et al. [2021] destaca a importância da farmacogenética (PGx) na personalização de terapias medicinais e no combate a RAM, abordando a utilização de bulas suíças como fonte de informações para os profissionais de saúde na Suíça. Um ponto forte é o uso de PLN para identificar informações relevantes de PGx, o que demonstra uma abordagem inovadora e potencialmente replicável em outros contextos. Além disso, o estudo sublinha a necessidade de padronização e harmonização das bulas para facilitar a aplicação clínica, proporcionando *insights* valiosos para a melhoria futura das bulas de medicamentos. Contudo, algumas limitações merecem atenção. A exclusão de informações farmacodinâmicas e o foco restrito à farmacocinética podem limitar a abrangência dos resultados, especialmente em áreas como oncologia, onde a genotipagem é essencial. Além disso, as dificuldades na padronização semântica, embora compreensíveis, representam um desafio na replicabilidade da metodologia. Apesar dessas limitações, o trabalho contribui significativamente ao enfatizar a necessidade de bulas mais informativas e organizadas, sugerindo um caminho para avanços na personalização de cuidados farmacêuticos.

Capítulo IV MetaMap e o Processo de Anotação de Dados na Pesquisa Biomédica

O MetaMap é uma ferramenta essencial para o processo de anotação de dados na pesquisa biomédica. Este capítulo explora como o MetaMap é utilizado para mapear conceitos médicos a partir de textos em linguagem natural, facilitando a identificação de sinais de farmacovigilância.

IV.1 Preparação dos Dados para Anotação

Nesta seção, são apresentados os procedimentos adotados para a preparação dos dados antes da anotação com o MetaMap. A preparação dos dados é uma etapa fundamental para garantir a qualidade e a eficácia no mapeamento semântico e, consequentemente, nos resultados dos experimentos. As subseções a seguir descrevem os processos de coleta, seleção e pré-processamento dos dados.

IV.1.1 Coleta e Seleção de Dados do Twitter

A coleta de dados foi realizada a partir do Twitter, abrangendo o período de 2016 a 2023 e considerando exclusivamente tuítes em português. Foram investigados dois temas relevantes para a saúde pública: (i) Medicamentos da Classe Terapêutica de Enxaqueca e (ii) Dengue - Sinais e Sintomas. O objetivo principal foi identificar possíveis SF relatados por usuários, contribuindo para a área de farmacovigilância.

A escolha do Twitter como fonte de dados baseou-se em sua capacidade de captar, em tempo real, o discurso e a opinião pública sobre temas de interesse médico e de saúde. A análise da primeira temática, Medicamentos da Classe Terapêutica de Enxaqueca, buscou compreender a percepção do público sobre medicamentos para enxaqueca, com foco na identificação de EAs relatados pelos usuários. A segunda temática, Dengue - Sinais e Sintomas, centrou-se nos sinais e sintomas da dengue, uma doença de importância global, com o intuito de fornecer subsídios para aprimorar a vigilância epidemiológica e apoiar ações de prevenção.

Inicialmente, a coleta foi realizada com o uso da extensão do KNIME para Twitter. Contudo, após a aquisição da plataforma por uma nova gestão, as mudanças na *Application Programming Interface* (Interface de Programação de Aplicativos) (API) tornaram essa abordagem inviável. Como

alternativa, utilizou-se a biblioteca *snscrape* em linguagem Python, que possibilitou a continuidade da coleta de tuítes para a conclusão dos experimentos.

Os dados coletados passaram por um processo de análise para garantir sua coerência e relevância para os objetivos da pesquisa. Foram aplicados filtros linguísticos e semânticos para selecionar somente tuítes que apresentassem informações potencialmente úteis para o mapeamento biomédico.

IV.1.2 Pré-processamento dos Textos

Após a coleta, os tuítes foram submetidos a um rigoroso pré-processamento, essencial para a preparação dos dados para o mapeamento no MetaMap. As etapas dessa fase incluíram:

- **Remoção de ruídos**: Exclusão de caracteres especiais, *URLs*, *emojis* e menções irrelevantes.
- **Normalização do texto**: Padronização para caixa baixa, garantindo uniformidade nas palavras.
- **Filtragem de expressões irrelevantes**: Identificação e exclusão de termos ou frases que não contribuíam para os objetivos do estudo.

O pré-processamento foi conduzido utilizando a ferramenta KNIME Analytics Platform, que forneceu um fluxo de trabalho visual e modular para as tarefas mencionadas. Após essa etapa, os textos estavam adequadamente preparados para serem mapeados no UMLS em Português do Brasil por meio do MetaMap.

IV.1.3 Fluxo de Atividades do Mapeamento

O processo completo, desde a coleta até o mapeamento dos tuítes, seguiu um fluxo de atividades bem definido, conforme ilustrado na Figura IV.1. O fluxo incluiu as seguintes etapas:

1. **Coletar Tuítes**: A *string* de busca foi inserida em um programa desenvolvido em Python, que filtrava os tuítes apenas na língua portuguesa.
2. **Analisar Tuítes**: Verificação da coerência dos dados coletados e ajuste da *string* de busca, se necessário.
3. **Pré-processar Tuítes**: Aplicação de PLN nos dados coletados, com exportação de um arquivo pronto para mapeamento.
4. **Mapear Tuítes**: Execução dos serviços do MetaMap, com a instância do modelo UMLS em português, para mapeamento conceitual.

5. **Manter Dados**: Realização de novos processamentos e enriquecimento dos dados coletados, visando melhorar a análise.

6. **Visualizar Dados**: Criação de gráficos para interpretação e identificação de *insights* utilizando KNIME e Power BI.

7. **Exportar Dados**: Armazenamento seguro dos dados processados para backup ou análises futuras.

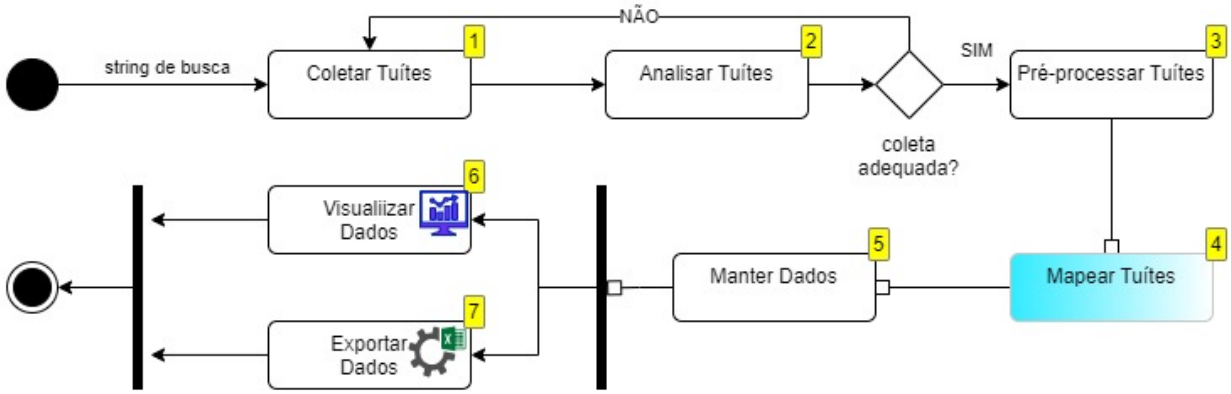


Figura IV.1: Fluxo da Coleta ao Mapeamento de Tuítes.

Os resultados obtidos a partir desse fluxo são detalhados no Capítulo V, com ênfase nas contribuições dos dados mapeados para a farmacovigilância e a saúde pública.

IV.2 Configuração e Funcionamento do MetaMap

Nesta seção, são apresentados os passos necessários para a instalação, configuração e adaptação do MetaMap ao idioma português do Brasil, bem como os aspectos técnicos relacionados à infraestrutura do experimento.

IV.2.1 Infraestrutura e Ambiente de Execução

A infraestrutura do experimento foi composta por uma máquina equipada com o sistema operacional Pop!-OS 22.04.4 LTS, processador Intel Core i5-8250U, 16GB de RAM e 1TB de armazenamento. Essa configuração foi escolhida para atender às demandas computacionais do MetaMap, especialmente no processamento de grandes volumes de dados textuais.

O armazenamento mínimo necessário para a instalação do MetaMap e seus arquivos associados foi de 100GB, considerando as bases de dados da UMLS e arquivos gerados durante os experimentos. A instalação do Java na versão 8 foi obrigatória para garantir a compatibilidade com o MetaMap.

IV.2.2 Instalação e Configuração Inicial da Ferramenta

A instalação do MetaMap e do UMLS para idiomas diferente do inglês, é extramente complexa. É altamente recomendável instalar o UMLS *Metathesaurus* e o *Metamorphosys*, pois a execução do *Metamorphosys* e a visualização da lista de vocabulários no UMLS fornecem uma noção da escala do UMLS. A seguir, os programas instalados e suas respectivas funções:

1. **Lexical Tools (*Lexical Tools* (Gerador de Variante Lexical) (LVG)) 2023 Full Version**: Utilizado para normalização e análise lexical[1].
2. **MetaMap 2020 Linux Version**: Ferramenta principal para o mapeamento semântico[2].
3. **Linux Datafile Builder Suite 2022 (*MetaMap Data File Builder* (DFB))**: Responsável pela criação de arquivos de dados personalizados[3].
4. **2023AA Full UMLS Release Files**: Base de dados completa para mapeamento[4].
5. MetaMap Java API: Utilizado para a integração e execução automatizada do MetaMap[5].

Para realizar o *download* desses componentes, foi necessário criar um cadastro no portal oficial da UMLS (`https://uts.nlm.nih.gov/uts/signup-login`). As configurações adicionais incluíram ajustes de alocação de memória e parâmetros de execução para otimizar o desempenho da ferramenta em grandes volumes de dados.

IV.2.3 Adaptação do MetaMap para o Português do Brasil

Para utilizar o MetaMap com textos em português, foi necessário criar um modelo UMLS específico para o português do Brasil. O MetaMap Data File Builder (DFB) é um módulo que permite a criação de modelos de dados UMLS personalizados. O processo de criação do modelo UMLS em português do Brasil envolveu as seguintes etapas:

- **Seleção dos Vocabulários**: A partir dos 214 vocabulários biomédicos disponíveis no UMLS, foram selecionados aqueles relevantes para o contexto da pesquisa e disponíveis em português do Brasil.
- **Geração do Subconjunto**: Utilizando a ferramenta *MetamorphoSys*, foi gerado um subconjunto do *Metathesaurus* UMLS contendo apenas os vocabulários selecionados no passo anterior.

[1] `https://lhncbc.nlm.nih.gov/LSG/Projects/lvg/current/docs/userDoc/install/releaseNotes.html`
[2] `http://metamap.nlm.nih.gov/MainDownload.shtml`
[3] `https://metamap.nlm.nih.gov/DataFileBuilder.shtml`
[4] `https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html`
[5] `https://data.lhncbc.nlm.nih.gov/umls-restricted/ii/tools/MetaMap/download/public_mm_l`

- **Construção dos Arquivos de Dados**: Com o auxílio do MMTx Data File Builder DFB, foram criados os arquivos de dados necessários para o funcionamento do MetaMap com o modelo UMLS em português do Brasil [Meystre and Haug, 2005].

A Figura IV.2 apresenta a arquitetura do ambiente utilizado para a criação e utilização do modelo UMLS em português do Brasil no MetaMap. Essa infraestrutura foi projetada para possibilitar a integração do MetaMap com dados e ferramentas biomédicas no contexto do idioma português. A seguir, descrevem-se os componentes e o fluxo de informações:

- **MetaMap**:
  - O DFB deve ser usado para idiomas diferentes do inglês.
  - Ferramentas auxiliares como 'skrmedpostctl', 'wdsserverctl' e a *API Java do MetaMap* garantem a comunicação e o processamento interno.
- **UMLS (Unified Medical Language System)**:
  - **Metathesaurus**: Inclui arquivos específicos do idioma português, como ICPPOR 1999AA, LNC-PT-BR 2022AB, MDRPOR 2022AB e MSHPOR 2022AA, e WHOPOR 1999AA.
  - **Rede Semântica**: Agrupa informações sobre conceitos, tipos semânticos e relações semânticas entre os termos.
  - **Ferramentas Lexicais (SPECIAL)**: Utiliza recursos como 'lvg static', 'metamorphosys' e outras ferramentas de apoio ao processamento linguístico.
- **Fonte de Vocabulário UMLS**:
  - Baseia-se em dados fornecidos pela NLM (National Library of Medicine) e na USABase (padrão em inglês), com suporte adicional de outras bases biomédicas.
- **Modelo UMLS Português Brasil**:
  - Este modelo é armazenado em um banco de dados Berkeley DB e contém os vocabulários e conceitos adaptados ao idioma português, permitindo o mapeamento preciso de textos biomédicos nesse idioma.
- **Servidor MetaMap**:
  - Atua como intermediário, conectando o modelo UMLS em português com as aplicações que utilizam os serviços do MetaMap.
- **Aplicação de Mapeamento JAVA**:

  - Desenvolvida em Java, essa aplicação lê os arquivos de entrada no formato de termos ou textos e os envia ao MetaMap para o mapeamento. Após o processamento, os arquivos mapeados retornam para análise e utilização.

- **Fluxo de Dados**:

  - O fluxo começa com a aplicação de mapeamento lendo os arquivos de entrada e enviando-os ao servidor MetaMap. O MetaMap processa os dados utilizando o modelo UMLS em português e retorna os resultados mapeados para os arquivos de saída.

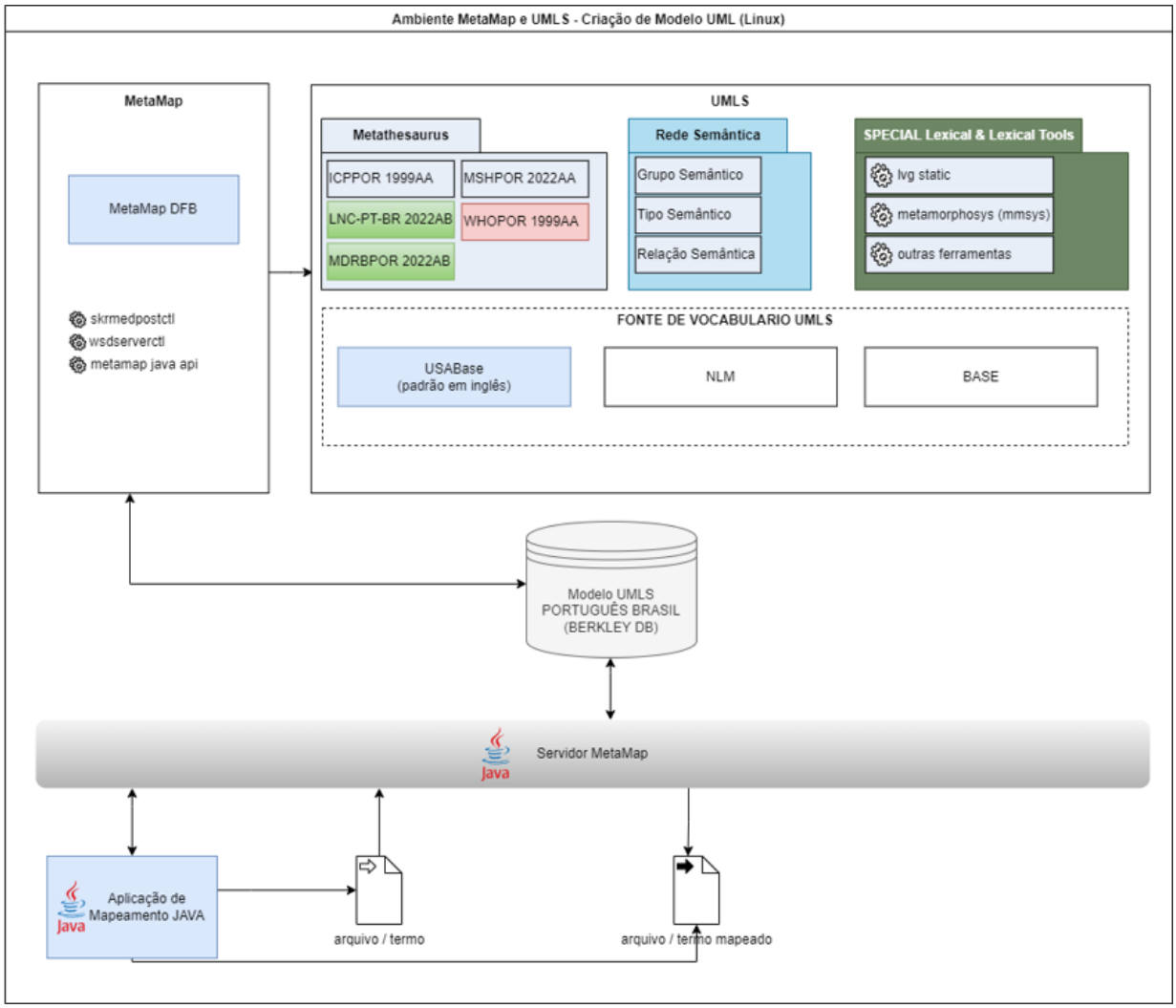


Figura IV.2: Ambiente MetaMap e UMLS - Criação de Modelo UMLS em Português.

Esta arquitetura foi fundamental para suportar a análise de textos biomédicos em português e possibilitar o mapeamento semântico de conceitos médicos, viabilizando a aplicação em estudos de farmacovigilância e pesquisa biomédica.

IV.2.4 Configuração no Ambiente de Integração

A integração do MetaMap com outras ferramentas, como o KNIME Analytics Platform, permitiu realizar tarefas complementares, como pré-processamento e visualização de dados. A versão 4.7.2 do KNIME foi utilizada para organizar fluxos de trabalho e realizar análises integradas.

Além disso, a integração com a MetaMap Java API possibilitou a automação do processo de mapeamento semântico, reduzindo o esforço manual e permitindo o processamento de grandes volumes de dados de maneira eficiente.

IV.2.5 Limitações e Requisitos Específicos

Embora o MetaMap seja amplamente reconhecido por sua eficiência no mapeamento de conceitos biomédicos, a ferramenta apresenta algumas limitações e requisitos técnicos específicos que devem ser considerados. Entre as principais restrições está a necessidade de sistemas operacionais Linux ou macOS para as versões mais recentes, sendo incompatível com sistemas Windows modernos. Além disso, a obrigatoriedade do uso do Java 8, uma versão descontinuada em muitos ambientes de produção, também representa um desafio técnico.

Versões anteriores do MetaMap (anteriores a 2014) podem ser instaladas em sistemas operacionais Windows XP ou 7. No entanto, essas versões carecem de suporte técnico e atualizações, o que as torna menos adequadas para projetos que demandam confiabilidade e recursos atualizados.

No contexto da adaptação do MetaMap ao português do Brasil, foram enfrentados desafios adicionais:

- **Disponibilidade de Vocabulários**: A quantidade limitada de vocabulários biomédicos em português do Brasil restringe a abrangência das análises e a precisão do mapeamento semântico.
- **Variação Linguística:** O português brasileiro é uma língua rica em variações regionais e contextuais, exigindo constantes atualizações nos dicionários de sinônimos e termos biomédicos para acompanhar as mudanças e garantir a relevância dos resultados.
- **Ausência de Geração de Variantes para o Português**: O MetaMap não realiza a geração de variantes léxicas e ortográficas para o português, um recurso que inclui a criação de flexões (como singular e plural), derivações (como substantivos e adjetivos) e correção de erros ortográficos. Essa limitação impacta diretamente a capacidade da ferramenta de lidar com a complexidade e as particularidades do idioma.

Apesar dessas limitações, o uso do MetaMap no contexto desta pesquisa foi bem-sucedido graças a ajustes específicos, como a personalização de dicionários e a criação de um modelo UMLS em

português do Brasil. Essas adaptações permitiram superar grande parte dos desafios, garantindo que os resultados obtidos fossem relevantes e precisos para os objetivos propostos.

IV.3 Processo de Anotação do MetaMap

O processo de anotação é um dos elementos centrais deste trabalho, consistindo na identificação de conceitos biomédicos presentes em textos por meio do MetaMap. Nesta seção, detalham-se as etapas envolvidas, desde a preparação dos dados até a atribuição de *scores*, destacando as técnicas e metodologias aplicadas [Aronson, 2001].

IV.3.1 Etapas do Processo de Anotação

O processo de anotação no MetaMap é composto por várias etapas interconectadas, conforme descrito a seguir.

IV.3.1.1 Pré-processamento

Nesta fase inicial, os textos são preparados para análise. Isso inclui:

- **Tokenização**: Divisão do texto em unidades menores, como palavras ou frases.
- **Normalização de Caracteres**: Correção de caracteres especiais e uniformização da caixa (minúsculas).
- **Remoção de Ruídos**: Filtragem de elementos desnecessários, como *emojis* e *stopwords*.

Essas etapas garantem que o texto esteja em um formato adequado para o mapeamento.

IV.3.1.2 Geração de Variantes

O MetaMap realiza a geração de variantes léxicas e ortográficas para ampliar a cobertura dos conceitos. Esse processo inclui:

- **Flexões**: Identificação de formas no singular e plural.
- **Derivações**: Reconhecimento de formas relacionadas, como adjetivos e substantivos.
- **Correção Ortográfica**: Detecção e ajuste de erros ortográficos no texto de entrada.

Essa funcionalidade é essencial para garantir maior abrangência e precisão na busca de conceitos biomédicos.

IV.3.1.3 Busca de Candidatos

Com as variantes geradas, o MetaMap busca por candidatos no *Metathesaurus* UMLS que correspondam aos termos do texto. Essa etapa envolve a recuperação de *strings* que apresentam alta similaridade com as variantes do texto de entrada, criando uma lista inicial de possíveis mapeamentos.

IV.3.1.4 Avaliação de Candidatos

Os candidatos identificados são avaliados com base em um sistema de pontuação (*scoring*) que considera os seguintes critérios:

- **Centralidade**: Relação do termo com o contexto geral do texto.
- **Variação**: Capacidade de representar diferentes formas do termo.
- **Cobertura**: Quantidade de palavras do texto de entrada que são cobertas pelo candidato.
- **Coesão**: Consistência do termo dentro do mapeamento gerado.

Essa avaliação permite a seleção dos conceitos mais adequados para o texto em análise.

IV.3.1.5 Construção de Mapeamentos

Nesta etapa final, o MetaMap combina os candidatos com os melhores *scores* para formar o mapeamento final [Aronson, 2001]. O objetivo é gerar uma representação semântica coerente do texto, que seja útil para a análise biomédica. Mais detalhes deste mapeamento estão no Apêndice B.

IV.4 Resultados da Anotação

Nesta seção, são apresentados os resultados obtidos durante o processo de anotação utilizando o MetaMap. Os dados coletados foram analisados, mapeados e organizados para fornecer *insights* sobre conceitos biomédicos extraídos de tuítes, abordando etapas desde a coleta até o processamento final.

IV.4.1 Armazenamento e Organização dos Resultados

Os dados coletados foram salvos inicialmente em arquivos no formato Excel (“.xlsx”) e, posteriormente, convertidos para arquivos CSV durante o processamento no KNIME. Esses arquivos representam tanto os registros processados quanto os resultados do mapeamento biomédico.

Para a etapa de mapeamento, foi desenvolvido um programa em Java que lê os arquivos gerados pelo KNIME e realiza o processo de anotação por meio do MetaMap. Esse programa foi ajustado para otimizar a leitura dos arquivos CSV, dividindo os resultados em três arquivos distintos:

- **Registros Mapeados**: Contendo todos os tuítes analisados e os conceitos associados.
- **Tuítes Únicos Mapeados**: Incluindo apenas os tuítes que apresentaram termos biomédicos válidos.
- **Detalhes dos Mapeamentos**: Fornecendo informações detalhadas sobre cada conceito identificado.

Essa separação possibilitou uma análise mais estruturada e facilitou o processamento posterior no KNIME.

IV.4.2 Detalhes do Mapeamento Biomédico

Os arquivos gerados pelo MetaMap incluem informações detalhadas, como conceitos mapeados, tipos semânticos, pontuações e fontes. Esses dados permitem analisar a relação entre os termos identificados nos tuítes e os conceitos biomédicos presentes no UMLS.

Cada registro mapeado contém um conjunto de campos descritos na Tabela IV.1, como identificador único do conceito, nome preferido, tipos semânticos e fontes do vocabulário utilizado no mapeamento. Esses campos são essenciais para uma análise aprofundada dos resultados, destacando a precisão e abrangência do mapeamento realizado.

Nome do Campo	Descrição do Campo
ID	Identificador do tuíte coletado
Twitter	Tuíte coletado
Tipo Conceito	Tipo do conceito, por padrão vem “Mapeado”
Map Score	Pontuação geral de um mapeamento, considerando todas as correspondências com conceitos do UMLS
Score	Avaliação da qualidade da correspondência entre um termo do texto
Id do Conceito	Identificador único para cada conceito
Nome do Conceito	Nome do conceito mapeado para determinada palavra
Nome Preferido	Nome preferido pelo MetaMap para determinada palavra
Matched Words	Palavra que foi mapeada dentro do tuíte
Quantidade de Matched Words	Número de palavras que definem um conceito único
Tipos Semânticos	Tipo semântico mapeado para o conceito
Fontes	Vocabulário utilizado no mapeamento

Tabela IV.1: Descrição dos Campos Usados no Mapeamento do MetaMap.

IV.4.3 Processamento Posterior no KNIME

Após o mapeamento realizado pelo MetaMap, os arquivos CSV foram novamente processados no KNIME, passando por etapas adicionais de análise e filtragem. Durante este processamento, foram aplicadas técnicas de PLN e mineração de dados para:

- Extrair informações relevantes dos conceitos mapeados.
- Realizar a agregação e filtragem de dados para eliminar redundâncias e focar nos conceitos mais relevantes.
- Criar visualizações gráficas no Power BI para facilitar a interpretação e a descoberta de *insights*.

Esse fluxo de atividades garantiu uma análise mais refinada dos dados coletados, consolidando os resultados para atender aos objetivos da pesquisa.

IV.4.4 Considerações sobre os Resultados

Os resultados obtidos com o mapeamento e o processamento adicional indicam a eficácia do MetaMap no contexto do português brasileiro, mesmo diante das limitações discutidas. As análises geradas a partir dos dados processados fornecem *insights* relevantes para a farmacovigilância, auxiliando na identificação de EAs e RAMs.

A separação dos arquivos e a estruturação dos resultados facilitaram a validação dos dados e a integração com outras etapas do projeto. Esses avanços destacam o potencial do MetaMap como ferramenta para análises biomédicas em mídias sociais.

Capítulo V Experimentos

Os experimentos realizados neste trabalho tiveram como objetivo principal validar o uso do MetaMap adaptado ao português do Brasil para o mapeamento de conceitos biomédicos em tuítes. A partir dos dados previamente coletados e pré-processados, buscou-se explorar a eficácia da ferramenta no contexto da farmacovigilância, identificando SFs. Para isso, foram conduzidos experimentos com diferentes conjuntos de dados relacionados a (i) Medicamentos da Classe Terapêutica de Enxaqueca e (ii) Dengue - Sinais e Sintomas de saúde pública, analisando métricas como cobertura de mapeamento, precisão e relevância dos conceitos identificados. Este capítulo descreve os passos experimentais realizados, os métodos de avaliação empregados e os resultados obtidos, destacando as contribuições práticas para a segurança de medicamentos e o monitoramento epidemiológico.

V.1 Medicamentos da Classe Terapêutica de Enxaqueca

A investigação do tema medicamentos da classe terapêutica de enxaqueca baseou-se no conjunto de dados "DADOS_ABERTOS_MEDICAMENTOS"[1]. Esse conjunto de dados foi escolhido devido à sua temática relacionada à detecção de RAM Não Identificadas de Medicamentos em Mídias Sociais. Alguns ajustes nos nomes de medicamentos foram realizados para criação de uma *string* de busca mais efetiva.

Nessa base, foram identificadas três classes terapêuticas de enxaqueca, totalizando 66 medicamentos. No entanto, algumas coletas específicas resultaram na exclusão de alguns medicamentos por motivos diversos:

1. ISOMETETEPTENO + DIPIRONA + CAFEINA - não vem nenhum resultado na busca
2. DOLOR - esse termo é um nome para dor em diversas línguas, logo, descaracterizou a coleta
3. IMIGRAN - termo que remete a imigrante, dificultando a coleta
4. PARCEL - termo que remete a parcela e parcelamento, na análise da coleta não foi identificado nenhum medicamento
5. ENJOY - termo muito comum nas redes sociais para agradecimentos

[1] `https://github.com/LucasPLTC/TCC_CEFET_RJ_ADR_EXTRACTION_PROCESS`

6. NARAMIG - este medicamento também é um perfil no Twitter de uma farmacêutica, logo retornava muitos tuítes sem relevância para este trabalho, logo, para manter este nome, foi necessário criar uma etapa no processo para verificar se era nome de uma pessoa ou um medicamento. Este mesmo caso, de forma similar aconteceu com outros "nomes de medicamentos"

Os termos pesquisados foram criados em partes, para que a coleta não tivesse uma grande quantidade de registros e dificultasse os processos posteriores pós-coleta. Sendo assim as coletas foram realizadas em partes no programa em Python da seguinte forma:

1. "CEFALIV"
2. "ENXAK"
3. "ACIDO ACETILSALICILICO" OR ACOTERMON OR BESODIN OR CEFALINA OR CEFALIUM OR CEFALOL OR "CLORIDRATO DE NARATRIPTANA" OR DESERILA OR DORIDINA OR "HIDROBROMETO DE ELETRIPTANO"
4. MIGRAL OR MIGRANE OR MIGREX OR MIGROFF OR NARANETY OR NARATANO OR NARATRIN OR NARCEF OR ORMIGREIN OR PASURTA
5. "SUMAX"
6. POSDRINK OR SEDOL OR "SUCCINATO DE SUMATRIPTANO" OR SUMATRIPTANO OR SUTRIPTAN OR SYDOLIL OR "TANACETO EC" OR "TANACETUM ORIENT" OR "TARTARATO DE ERGOTAMINA" OR TENLIV OR TONOPAN OR ZOMIG
7. "NARAMIG"
8. "MAXALT"
9. "MESILATO DE DIIDROERGOTAMINA + DIPIRONA SODICA + CAFEINA"

A Figura V.1 apresenta um painel contendo dados da coleta e processamento de tuítes relacionados a medicamentos, distribuídos entre os anos de 2016 e 2023. À esquerda, são exibidos os principais indicadores do estudo, incluindo:

- **46.386** tuítes coletados no período.
- **35.893** tuítes processados, após etapas de limpeza e filtragem.
- **23.08** tuítes mapeados como únicos, com termos biomédicos relevantes.

- **49.586** registros mapeados, representando a totalidade de conceitos extraídos com a ferramenta MetaMap.

No centro, há um gráfico de barras agrupadas que compara a quantidade de tuítes coletados (em azul claro) com os registros mapeados (em azul escuro) para cada ano.

Análises e *Insights*:

- **Tendência de Coleta e Processamento:** Há um aumento significativo no número de tuítes coletados e processados a partir de 2019. Em 2016, foram coletados 3.715 tuítes, enquanto em 2020 esse número saltou para 11.404. Esse crescimento pode ser atribuído ao aumento da popularidade e uso do Twitter como plataforma de discussão sobre saúde e medicamentos. O pico de tuítes coletados ocorreu em 2020, com 11.404 tuítes, provavelmente influenciado pela pandemia de COVID-19, que gerou um aumento substancial nas discussões sobre saúde e medicamentos nas redes sociais. A partir de 2021, observa-se uma leve redução no volume de tuítes coletados e mapeados, embora ainda mantenha números expressivos.

- **Picos Relevantes:** O ano de **2020** apresentou o maior volume de tuítes coletados (11.404) e registros mapeados (9.038), indicando maior interação e discussões sobre medicamentos, possivelmente devido à alta busca por informações sobre tratamentos e vacinas durante a pandemia. O segundo maior pico ocorre em **2021**, com 10.119 tuítes coletados e 7.519 registros mapeados, sugerindo um impacto residual do aumento de discussões gerado no ano anterior.

- **Eficiência do Processamento:** A proporção entre tuítes processados e registros mapeados é consistente ao longo dos anos, demonstrando a eficácia das ferramentas utilizadas, como o MetaMap, para identificar termos biomédicos relevantes.

- **Declínio em 2023:** O número de tuítes coletados (5.026) e registros mapeados (4.144) é inferior aos anos anteriores, o que é explicado devido o final da coleta ser até maio, sendo até o momento superior ao ano de 2018 onde começou levemente uma alta de registros.

A Figura V.1 apresenta informações quantitativas do processo de coleta, processamento e análise de dados de mídias sociais. Os picos observados em 2020 e 2021 destacam a relevância de eventos globais no comportamento digital da população, enquanto a análise temporal dos dados oferece *insights* sobre padrões de uso de medicamentos e percepção pública ao longo dos anos. Esses resultados reforçam a aplicabilidade de redes sociais como fonte alternativa de dados para farmacovigilância.

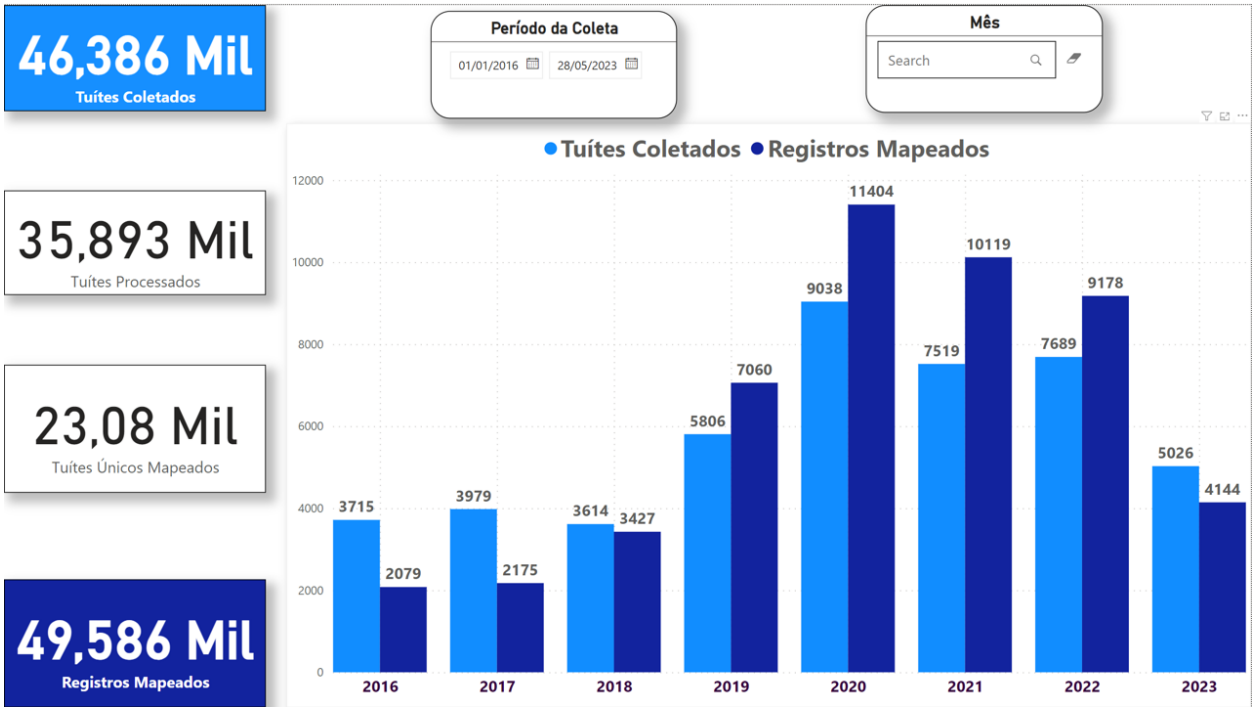


Figura V.1: Coleta e Processamento de Tuítes para Detecção de SF

A Figura V.2 exibe um tuíte coletado da plataforma Twitter. O tuíte destacado em vermelho representa uma resposta a outro tuíte. Após o processo de coleta, pré-processamento e mapeamento, obtém-se um resultado conforme ilustrado na Tabela V.1.

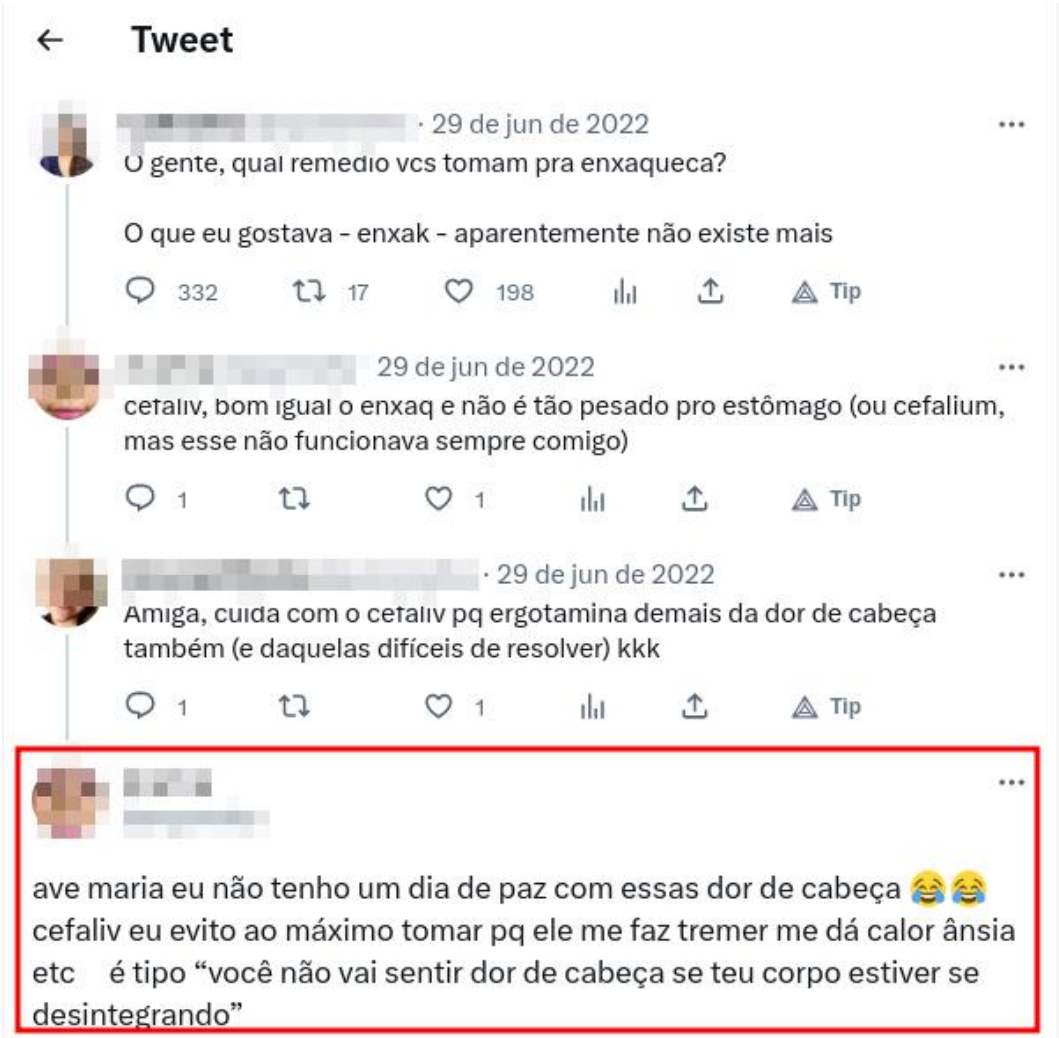


Figura V.2: Formato Original do Tuíte na Plataforma Twitter.

A Tabela V.1 é um recorte do tuíte do "ID Mapeado" de número 10.423, evidenciado na Figura V.2, que após ser pré-processado, apresentou o seguinte conteúdo:

AVE MARIA EU NÃO TENHO UM DIA DE PAZ COM ESSAS DOR DE CABEÇA CEFALIV EU EVITO AO MÁXIMO TOMAR PQ ELE ME FAZ TREMER ME DÁ CALOR ÂNSIA ETC É TIPO "VOCÊ NÃO VAI SENTIR DOR DE CABEÇA SE TEU CORPO ESTIVER SE DESINTEGRANDO"

Observa-se que na linha um da Tabela V.1, o "Nome Preferido" ainda não foi traduzido neste modelo da UMLS, mas o "Nome do Conceito" foi corretamente mapeado para "Ânsia". Na linha dois, o "Nome do Conceito" mapeado foi "Dor de Cabeça", um termo composto por três palavras, sendo que cada uma dessas palavras também teve seus mapeamentos individuais. O conjunto "Dor de Cabeça" é um termo bem conhecido e comum, sendo mapeado por quatro vocabulários diferentes.

Id do Conceito	Termo	Nome do Conceito	Nome Preferido	Score	Map Score	Tipo Semântico	Grupo Semântico	Vocabulário	Quantidade de Termos
C0870371	ânsia	Ânsia	Craving	-616	-540	Comportamento Individual	Atividades e Comportamentos	MeSH Português	1
C0018681	dor,de,cabeça	Dor de Cabeça	CEFALALGIA	-630	-540	Sinal ou Sintoma	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR	3
C0030193	dor	DOR	Dor	-616	-534	Sinal ou Sintoma	Distúrbios	MDRBPO, MSHPOR, WHOPOR	1
C0018837	calor	Calor	Calor	-616	-540	Fenômeno ou Processo Natural	Fenômenos	MeSH Português	1
C0018670	cabeça	Cabeça	Cabeça	-616	-534	Localização ou região do corpo	Anatomia	MeSH Português	1

Tabela V.1: Tuíte Mapeado.

Os resultados revelam as seguintes informações gerais: foram coletados um total de 46.386 registros de tuítes, dos quais, após o pré-processamento, restaram 35.399 registros. Dentre os tuítes pré-processados, o MetaMap conseguiu mapear 32.508 tuítes, sendo que cada tuíte pode conter mais de uma palavra mapeada, totalizando 49.586 palavras mapeadas no conjunto de dados.

Quantidade de Mapeamentos por Vocabulário

Como era esperado, os vocabulários mais específicos de cada idioma mapeiam mais palavras, como se percebe na Figura V.3 o vocabulário *Medical Subject Heading* (Títulos de Assuntos Médicos) (MeSH) Português que mapeou 33.797 termos, uma quantidade muito superior aos demais vocabulários. Em contra-partida o vocabulário WHOART Portuguese mapeou apenas quatro registros. É importante mencionar que o WHOART Portuguese é amplamente utilizado em pesquisas científicas, vigilância farmacológica e notificação de EAs em sistemas de saúde. No entanto, uma das limitações é que, infelizmente, o vocabulário não está sendo atualizado no *Metathesaurus*, o que pode afetar a disponibilidade de novos termos e atualizações para a terminologia. Mesmo assim, a terminologia continua sendo uma ferramenta valiosa para o registro e classificação de RAMs na língua portuguesa e é o vocabulário mais alinhado com essa pesquisa.

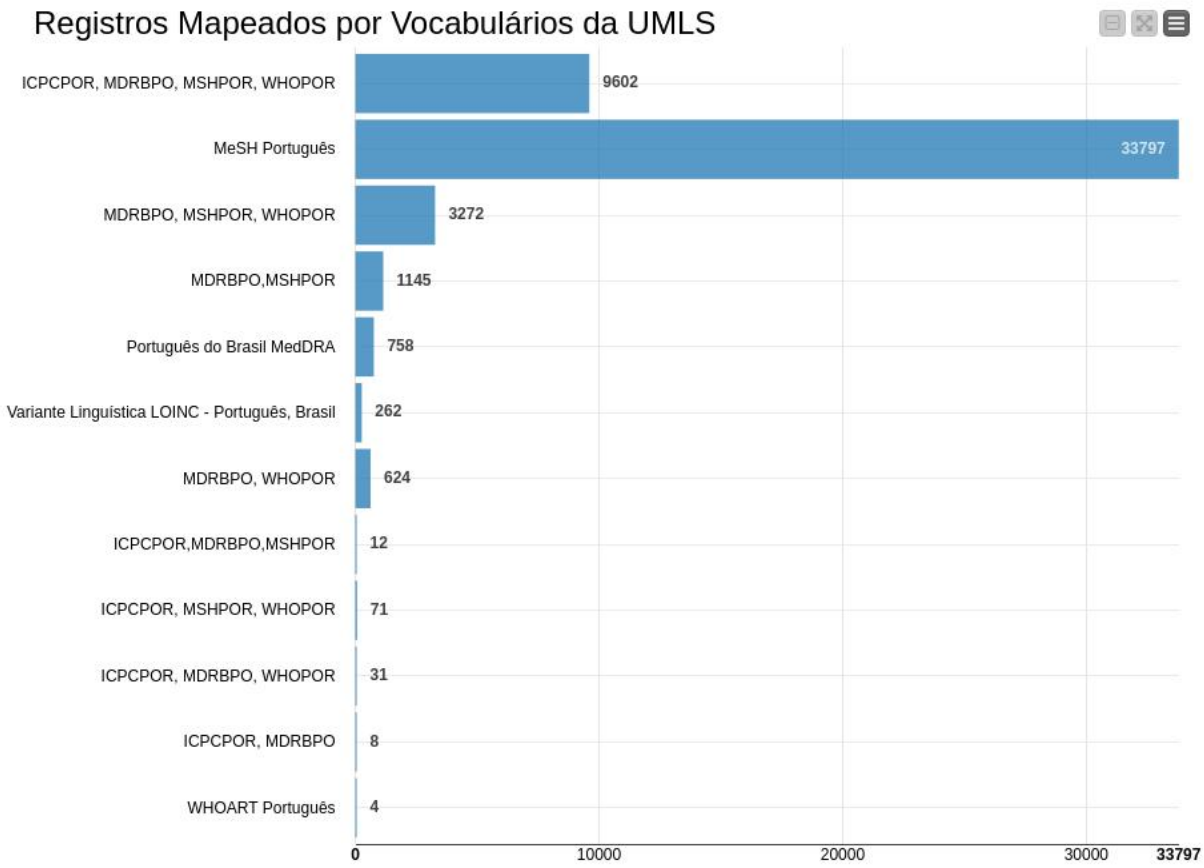


Figura V.3: Registros Mapeados por Vocabulário UMLS.

Mapeados por Grupo Semântico

Conforme esperado também, os dois maiores grupos com mais termos mapeados são "Produtos Químicos e Drogas" e "Distúrbios", devido à pesquisa de medicamentos relacionados à enxaqueca, conforme apresenta a Figura V.4. Esses medicamentos são produtos farmacológicos que visam amenizar algum distúrbio, contudo ao invés de amenizar pode ocorrer uma RAM, nada acontecer ou ainda ocorrer outro evento inesperado, um EA. Além disso, é compreensível que o grupo "Anatomia" apresente inúmeros mapeamentos, uma vez que a cabeça é a parte do corpo mais afetada nesses casos.

Os três grupos, "Produtos Químicos e Drogas", "Distúrbios" e "Anatomia", possuem respectivamente 18.050, 14.764 e 2.939 mapeamentos, representando aproximadamente 72% do total de termos mapeados.

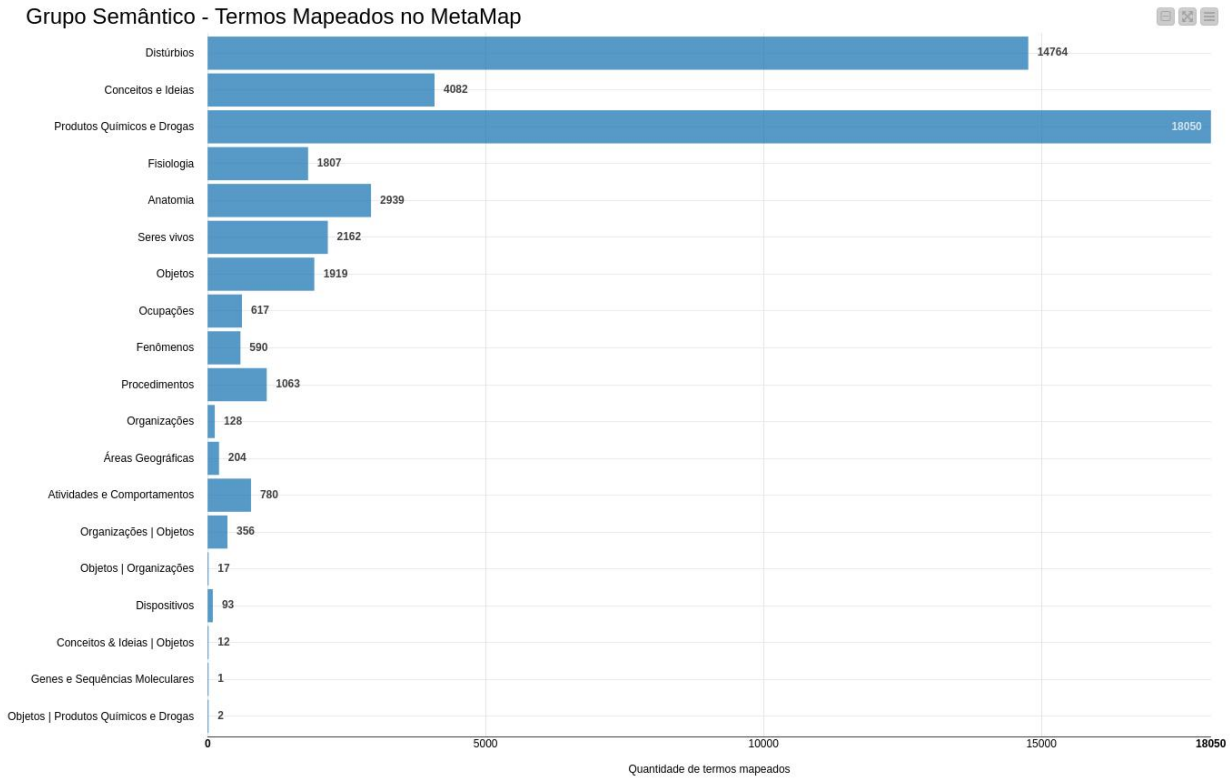


Figura V.4: Tuítes Coletados X Mapeados por Grupo Semântico.

Palavras Mapeadas por Tipo Semântico

Os tipos semânticos "Químico Orgânico" e "Substância Farmacológica" (orch e phsu) foram os que mais mapearam termos, sendo importante ressaltar que um único termo pode ter mais de um tipo de mapeamento. Um exemplo notável é o termo "ácido acetilsalicílico", amplamente conhecido no Brasil como "AAS" e "aspirina". Além desses, outros tipos semânticos relevantes incluem "Doença ou Síndrome", "Sinal ou Sintoma" e "Substância Farmacológica". Esses tipos são apenas alguns dos 133 Tipos Semânticos ilustrados na Tabela V.2. Eles desempenham um papel essencial na identificação e classificação dos termos mapeados, fornecendo informações valiosas sobre os conceitos presentes nos tuítes relacionados à enxaqueca.

Nome Tipo Semântico (PT-BR)	Count(Id Mapeado)
orch,phsu	10387
Doença ou Síndrome	7744
Sinal ou Sintoma	5340
Substância Farmacológica	4710
Produto intelectual	1889
Ideia ou Conceito	1361
Localização ou região do corpo	1330
Comida	1133
Parte do corpo, órgão ou componente do órgão	1032
Processo mental	890

Tabela V.2: Dez Tipos Semânticos com Mais Termos Mapeados.

Termos Mais Mapeados

Os termos mais mapeados estão fortemente relacionados ao tema desta coleta sobre enxaqueca, como ilustrado na Tabela V.3. Observa-se que a linguagem com maior frequência é o português, o idioma selecionado para a coleta. No entanto, é importante notar que outros idiomas também aparecem nos resultados devido ao mecanismo interno do Twitter, que tenta identificar automaticamente o idioma do texto do tuíte. Com base no conteúdo, o Twitter atribui um idioma ao tuíte, resultando na inclusão de outros idiomas além do português, como evidenciado no recorte da Tabela V.4. Essa diversidade linguística pode fornecer *insights* adicionais ao estudo, uma vez que permite uma análise mais abrangente das experiências dos usuários em diferentes contextos linguísticos [Suh et al., 2010].

Matched Words	Count(Id Mapeado)
enxaqueca	6618
remédio	3056
ácido,acetilsalicílico	3017
dor	2283
dor,de,cabeça	1979
dipirona	1523
cabeça	1211
vida	888
remédios	840
aspirina	687

Tabela V.3: Os Termos Mais Mapeados.

Linguagem	2016	2017	2018	2019	2020	2021	2022	2023	Total de Tweets
TOTAL	3715	3979	3614	5806	9038	7519	7689	5026	46386
pt	3618	3885	3529	5642	8790	7307	7490	4900	45161
ca	5	28	23	75	131	116	80	51	509
es	45	34	37	43	67	53	62	28	369
tl	11	8	6	17	25	18	21	15	121
it	21	10	8	13	13	16	22	8	111
fr	4	6	6	2	1	1	1	4	25
en	6	1	3	1	4	1	3	2	21
ht	1	?	?	8	2	3	1	4	19
in	?	2	2	2	?	2	1	3	12

Tabela V.4: Tuítes Mapeados por Ano e Linguagem.

Nuvem de Palavras

As Figuras V.5 e V.6 exibem as nuvens de palavras com as 250 palavras mais proeminentes da coleta pré-processada e dos mapeamentos, respectivamente.

Na Figura V.5, podemos observar que termos como “enxaqueca”, “Cefaliv”, “remédio” e “dor” são altamente representativos nesta coleta, indicando que estão frequentemente mencionados nos tuítes sobre enxaqueca coletados.

Já na Figura V.6, vemos as palavras mapeadas com maior destaque no contexto do mapeamento realizado pelo MetaMap. É possível notar termos como “efeito”, “dor”, “receita”, “sono” e “enxaqueca”, cruciais para compreender os conceitos e tópicos discutidos nas mensagens relacionadas à enxaqueca.

Essas nuvens de palavras proporcionam uma visão geral das palavras-chave mais relevantes e podem auxiliar na identificação de tendências e temas relevantes relacionados à enxaqueca a partir da análise dos dados coletados e mapeados.

Durante o processo de mapeamento com o MetaMap, algumas palavras foram mapeadas de forma não adequada para o contexto desta pesquisa. Por exemplo, termos como “fé”, que se refere a crenças, foram mapeados como elementos químicos, neste caso, Fe++ que representa o Ferro. Da mesma forma, “mg”, frequentemente utilizado para representar “miligrama” ou “Minas Gerais”, foi mapeado como Mg++ que corresponde ao Magnésio, assim como “na”, uma preposição, foi mapeado como Na+ que representa o Sódio. Termos comuns como “p” ou “p/”, usado informalmente para “para”, foram mapeados como P+ (Fósforo), e “se”, pronome pessoal do caso oblíquo, foi mapeado como Se++ que representa o Selênio. Essas incongruências são resultados da limitação do modelo UMLS gerado em português, que possui uma quantidade reduzida de termos e vocabulários específicos para o Brasil.

No modelo UMLS gerado em português, percebe-se ausência de muitos termos, haja vista a pequena quantidade de vocabulários específicos do Brasil, somente dois de um total de 214. A palavra comprimido nesse contexto é bem relevante assim como suas variações, contudo, na coleta haviam diversas variações como: comprimidos, comprimidinho, comprimidinhos, comprimidão e nenhum destes foram mapeados, apenas “comprimido”. Em contrapartida, as palavras remédio, remédios e drogas foram mapeadas como esperado.

Algumas palavras escritas no dia-dia brasileiro tiveram mapeamento não esperado, como “ela”, pronome, foi mapeado para a “Doença de Lou Gehrig”, assim como “tinha”, um verbo que foi mapeado para “Dermatofitose” e o outro verbo “tento” mapeado para a planta “Alcaçuz-da-América”.

Realizada uma análise aleatória em 200 tuítes, verificou-se que destes, muitos termos não foram mapeados por falta de sinônimos neste modelo em português. Ressalta-se que sinônimos e variações de palavras para o português e outros idiomas diferentes do inglês não são gerados por limitação da ferramenta MetaMap DFB.

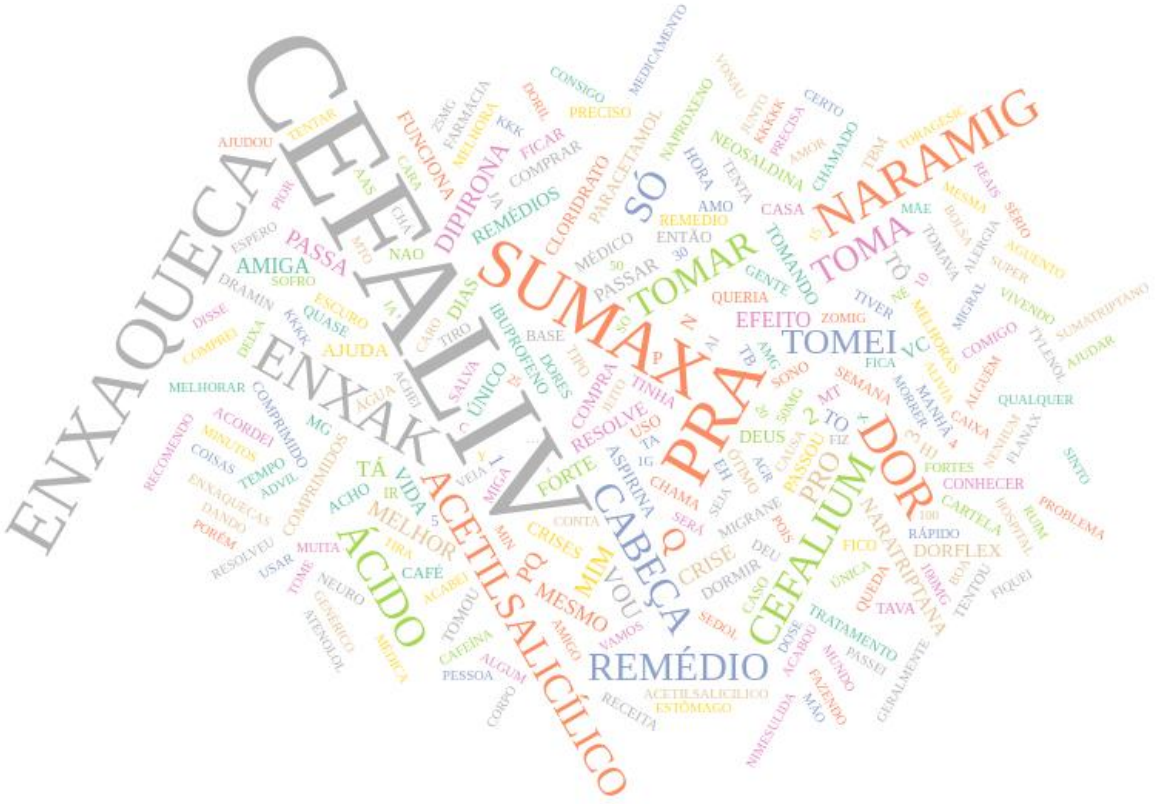


Figura V.5: Nuvem de Palavras Coletadas.



Figura V.6: Nuvem de Palavras Mapeadas Coletadas.

V.2 Dengue - Sinais e Sintomas

A dengue, uma doença tropical infecciosa causada pelo vírus da dengue, tem se apresentado como um desafio crescente para os sistemas de saúde ao redor do mundo. Esta seção explora os sinais e sintomas da dengue, ampliando o entendimento de seus efeitos sobre o organismo humano. As informações apresentadas, assim como a escolha dos termos para coleta de dados no Twitter

foram obtidas a partir do site do Ministério da Saúde[2], onde são detalhados os sintomas principais e sinais de alerta relacionados à doença. Todos os elementos listados foram avaliados e, quando necessário, adaptados ou complementados, para assegurar uma coleta mais eficaz no contexto deste estudo.

Os termos pesquisados foram:

'Dengue AND (Febre OR Dor OR "Mal estar" OR "Falta de apetite" OR "Manchas vermelhas" OR Vômito OR derrame OR Hipotensão OR lipotímia OR Letargia OR irritabilidade OR Hepatomegalia OR Sangramento OR hematócrito OR articulações OR articulação OR corpo OR olho OR cabeça OR palpação OR ascite OR sinal OR sintoma)'

A Figura V.7 apresenta um painel com os resultados da coleta, processamento e análise de tuítes relacionados à dengue no contexto da farmacovigilância. Os principais destaques do painel incluem:

- **56.829** tuítes coletados no período analisado.
- **54.943** tuítes pré-processados, após a remoção de ruídos e normalização dos dados.
- **250.264** registros mapeados, com a utilização da ferramenta MetaMap.
- **50.583** tuítes mapeados como únicos e relevantes para a análise.

Além disso, a figura exibe: um gráfico de barras, que compara os volumes anuais de tuítes coletados e pré-processados; uma nuvem de palavras, destacando os termos mais frequentes em tuítes relacionados à dengue, com destaque para palavras como "dengue", "febre", "dor", "corpo" e "articulações"; um mapa de calor, mostrando a distribuição geográfica dos tuítes coletados e uma tabela detalhada com os grupos semânticos mapeados e a quantidade de registros associados a cada grupo, com destaque para "Distúrbios" (169.873 registros) e "Sinais ou Sintomas" (36.790 registros).

Análises e *Insights*:

- **Tendência Temporal:** O gráfico de barras evidencia um crescimento constante na coleta e pré-processamento de tuítes até 2020, com picos em 2020 e 2022. Esse comportamento pode ser explicado por surtos de dengue e campanhas de saúde pública nesses períodos. Em 2021 o assunto do momento é COVID-19 e outros assuntos podem ter sido bem menos discutido nas mídias sociais.

[2] `https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/dengue`

- **Grupos Semânticos Relevantes:** A tabela de grupos semânticos mapeados mostra que a maior parte dos registros está associada a "Distúrbios", seguido por "Sinais ou Sintomas". Isso demonstra a eficácia do uso da ferramenta MetaMap na identificação de informações clínicas relevantes.

- **Distribuição Geográfica:** O mapa de calor revela uma alta concentração de tuítes em regiões tropicais e subtropicais, como América do Sul, Ásia e África, onde a dengue é mais prevalente mas não somente isso, a concentração também se limita aos tuítes em língua portuguesa.

- **Nuvem de Palavras:** A nuvem de palavras destaca os principais termos mencionados nos tuítes relacionados à dengue, indicando sintomas ("febre", "dor") e termos correlacionados a outras doenças, como "zika" e "chikungunya". Essa observação reforça a importância de compreender as discussões públicas para identificar tendências de saúde.

Este painel evidencia a capacidade de utilizar mídias sociais como fonte de dados para monitorar surtos de dengue e identificar percepções públicas sobre a doença. A integração de ferramentas como o MetaMap com análise semântica e geográfica oferece *insights* valiosos para intervenções de saúde pública e vigilância epidemiológica. O aumento expressivo nos registros mapeados e a diversidade dos dados analisados destacam o potencial desta metodologia para estudos futuros na área de farmacovigilância.



Figura V.7: Painel Interativo: Coleta e Processamento de Tuítes sobre Dengue

Mapeamentos por Vocabulário

O vocabulário MeSH em português desempenha um papel crucial neste modelo UMLS, mapeando, neste contexto, 80.40% mais termos que o MedDRA do Brasil. É importante destacar que, embora o MeSH não seja exclusivo para o Brasil, o vocabulário *LOINC Linguistic Variant - Portuguese, Brazil* apresenta um volume de mapeamento bastante diminuto. Esta observação é evidenciada na Figura V.8, que exibe a quantidade de tuítes mapeados em comparação com os termos identificados nesses tuítes mapeados. O volume de termos é frequentemente superior, já que normalmente mais de um termo é mapeado em cada tuíte. Juntos, o MeSH e o MedDRA mapeiam uma maior quantidade de termos, embora seja importante ressaltar que um único termo pode ser mapeado por vários vocabulários.

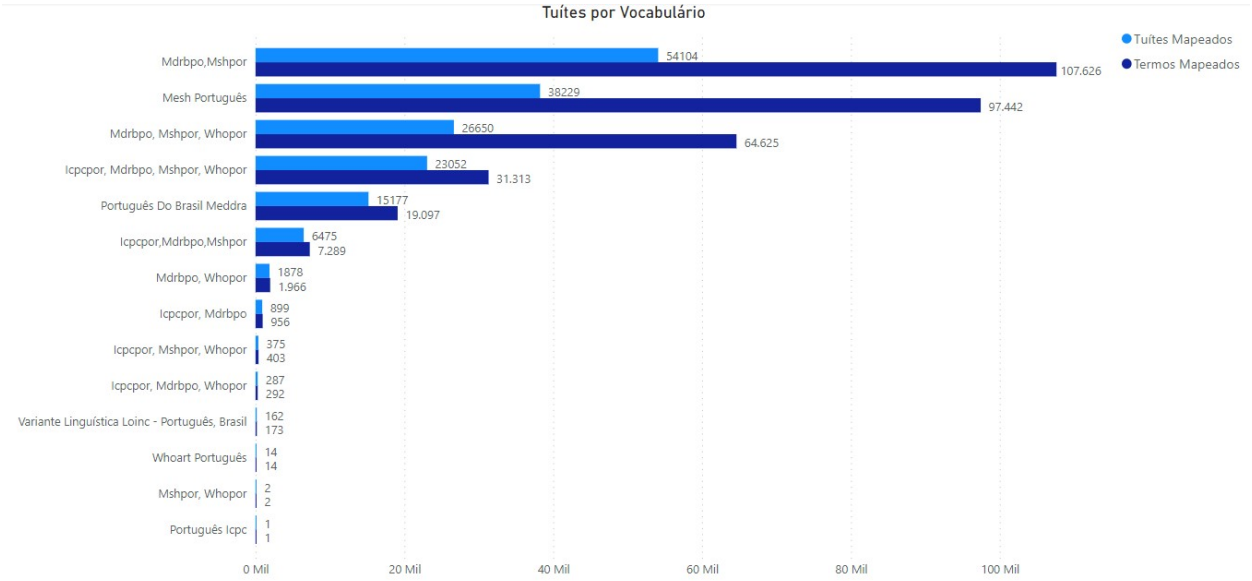


Figura V.8: Tuítes Coletados X Mapeados (Dengue).

Após a coleta de dados sobre a doença dengue e seus respectivos sinais e sintomas, dois tipos semânticos distintos emergiram como aqueles com a maior quantidade de mapeamentos: “Doença ou Síndrome” (referente à Dengue) e “Sinal ou sintoma” (relacionado aos sinais e sintomas da dengue), como apresentado na Figura V.9. Essa ilustração exibe apenas os 10 tipos mais bem classificados, e oferece uma percepção intrigante: a soma dos tipos “Disfunção Mental Ou Comportamental” e “Processo Mental” constituiria o “quinto tipo” mais frequente.

Em muitos tuítes, as pessoas expressam o profundo desconforto mental causado tanto pela doença quanto por medicamentos. Segundo a Organização Mundial da Sáude (OMS), alterações de humor ou função mental são classificadas como RAMs de gravidade moderada, e a prevalência dessas reações tem aumentado anualmente [de Moura et al., 2016].

Um dos fatores que contribuem para esse aumento é o uso de medicamentos incorretos ou sem prescrição médica [Trentin et al., 2019]. Na análise dos tuítes, é evidente que muitas pessoas estão recomendando medicamentos a outras sem um diagnóstico clínico adequado.

Este cenário não apenas aumenta o risco de EAs, mas também pode contribuir para o agravamento da angústia mental dos pacientes. A dengue por si só já é uma condição que gera desconforto e estresse, e o uso inadequado de medicamentos pode intensificar sentimentos de ansiedade e incerteza. Portanto, é essencial que as recomendações de medicamentos sejam feitas por profissionais de saúde qualificados para garantir o tratamento mais eficaz e seguro para os indivíduos afetados pela dengue [Organization et al., 2020].

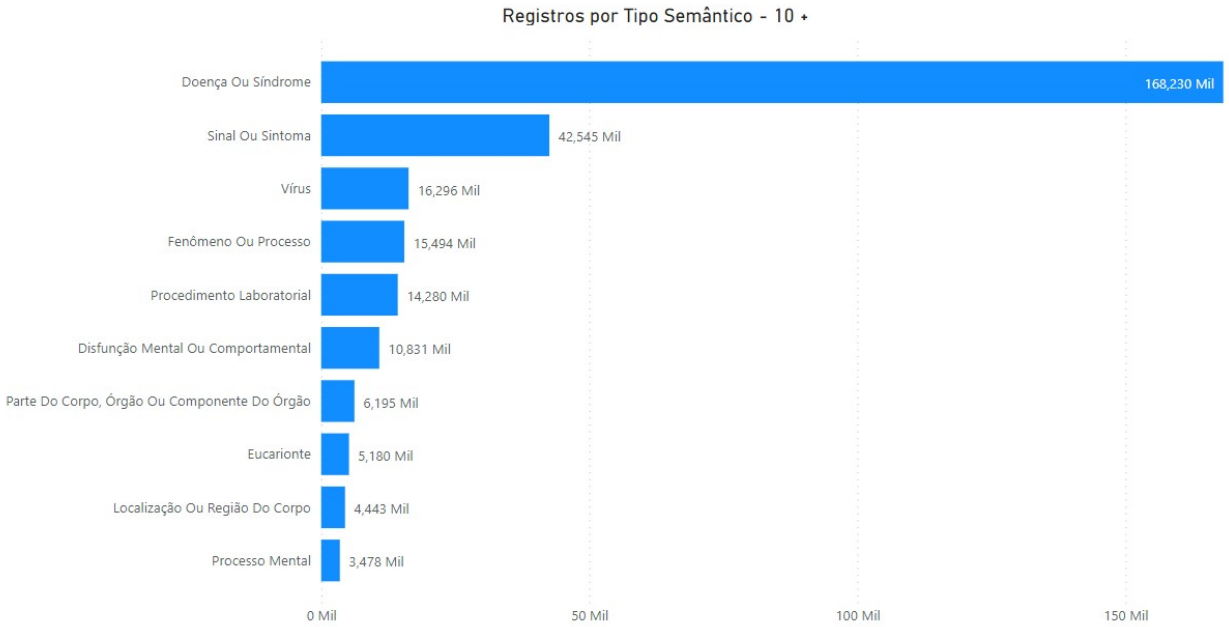


Figura V.9: Tuítes Coletados X Mapeados por Tipo Semântico (Dengue).

Análise dos Termos

A Tabela V.5 destaca os 12 termos mais frequentemente mencionados na coleta de dados. Nota-se que o termo "DOR", o terceiro do *ranking*, possui muitas variações encontradas ao longo desta lista, tais como "DOLORIDO", "DOLORIDA", "DOLORIDOS", "DOLOROSO", "DÓI", dentre outro que juntos apresentam um número significativo de ocorrências. No entanto, a maioria destes termos, quando apresentados de forma isolada, não são mapeados pelo MetaMap. Isso sugere que esses termos necessitam de um contexto ou complemento para poderem ser adequadamente mapeados. Alguns exemplos de contextos em que estes termos foram mapeados incluem "articulações doloridas", "braço dolorido", "olhos doloridos", entre outros. Para uma visão mais abrangente, a Figura V.10 ilustra uma Nuvem de Palavras com as 200 mais encontradas.

Termo	Ocorrências	Ranking
DENGUE	54550	1
FEBRE	26912	2
DOR	18640	3
AMARELA	15660	4
CORPO	14878	5
ZIKA	11909	6
GRIPE	9538	7
ANSIEDADE	9116	8
ALTA	9077	9
BAIXA	8460	10
DEPRESSÃO	8269	11
PRESSÃO	8170	12

Tabela V.5: Os Termos Mais Referenciados na Coleta.

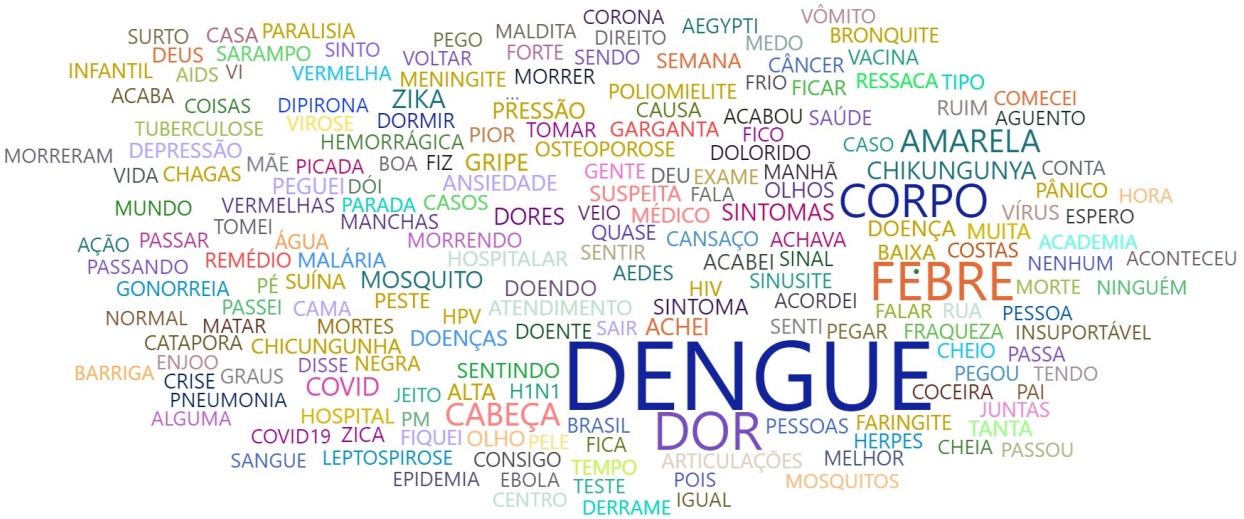


Figura V.10: Nuvem de Palavras dos Termos Mais Referenciados da Coleta de Dengue.

A Tabela V.6 evidencia os 12 termos que mais se destacaram no mapeamento dos tuítes coletados. A coluna “Termo” reflete as palavras exatas usadas nas postagens, enquanto a coluna “Nome do Conceito” representa a terminologia padrão armazenada no modelo UMLS. Importante notar que, dentro do UMLS, cada conceito possui um “Nome Preferido”. Por exemplo, o termo “ansiedade” foi identificado como “ANSIEDADE” no conceito, mas seu nome preferido é “Angústia”. Este termo foi mapeado em três vocabulários distintos.

É importante destacar que, ao contrário da Figura V.11, a tabela não foi submetida a um processo de limpeza para eliminar palavras sem relevância nesta temática. Isso fica evidente com a inclusão do termo “se” na sétima posição, que foi identificado como “Se++”, referindo-se ao

elemento químico Selênio. Esse mapeamento, aparentemente inadequado, ocorreu exclusivamente pelo vocabulário MedDRA em português. As comparações entre as tabelas e as nuvens de palavras da coleta e do mapeamento mostram claras distinções, com exceção dos termos predominantes, Dengue e Dor.

Termo	Nome do Conceito	Nome Preferido	Tipo Semântico	Grupo Semântico	Vocabulário	Quantidade de Termos
dengue	Dengue	Dengue	Doença ou Síndrome	Distúrbios	MDRBPO,MSHPOR	52721
pressão	Pressão	Pressão	Fenômeno ou Processo	Fenômenos	MeSH Português	14560
febre,amarela	Febre amarela	Febre amarela	Doença ou Síndrome	Distúrbios	MDRBPO,MSHPOR	14480
dor	DOR	Dor	Sinal ou Sintoma	Distúrbios	MDRBPO, MSHPOR, WHOPOR	13119
febre	FEBRE	Doenças Febris	Sinal ou Sintoma	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR	9498
ansiedade	ANSIEDADE	Angústia	Disfunção Mental ou Comportamental	Distúrbios	MDRBPO, MSHPOR, WHOPOR	9368
se	Se++	Se++	Procedimento laboratorial	Procedimentos	Português do Brasil MedDRA	8229
hiv	HIV	HIV	Vírus	Seres vivos	MeSH Português	7023
osteoporose	OSTEOPOROSE	Osteoporose, não especificada	Doença ou Síndrome	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR	7014
gripe,suína	Gripe suína	Gripe Porcina	Doença ou Síndrome	Distúrbios	MDRBPO,MSHPOR	6970

Tabela V.6: Os Termos Mais Mapeados.



Figura V.11: Nuvem de Palavras dos Termos Mais Mapeados.

Um tuíte, selecionado de forma aleatória para a análise, é ilustrado na Figura V.12. A imagem exibe a mensagem original exatamente conforme apresentada pelo usuário na plataforma Twitter.

O respectivo mapeamento do tuíte é retratado na Tabela V.7. A tela de mapeamento evidencia a ocorrência de certos termos mapeados mais de uma vez. Por exemplo, o termo “dengue” aparece em duplicidade no texto. Embora ele seja identificado pelo mesmo vocabulário, tipo e grupo semântico em ambas as instâncias, o “Map Score” varia. Isso ocorre porque o cálculo desta coluna leva em consideração a posição do termo dentro da frase, assim como os termos adjacentes.

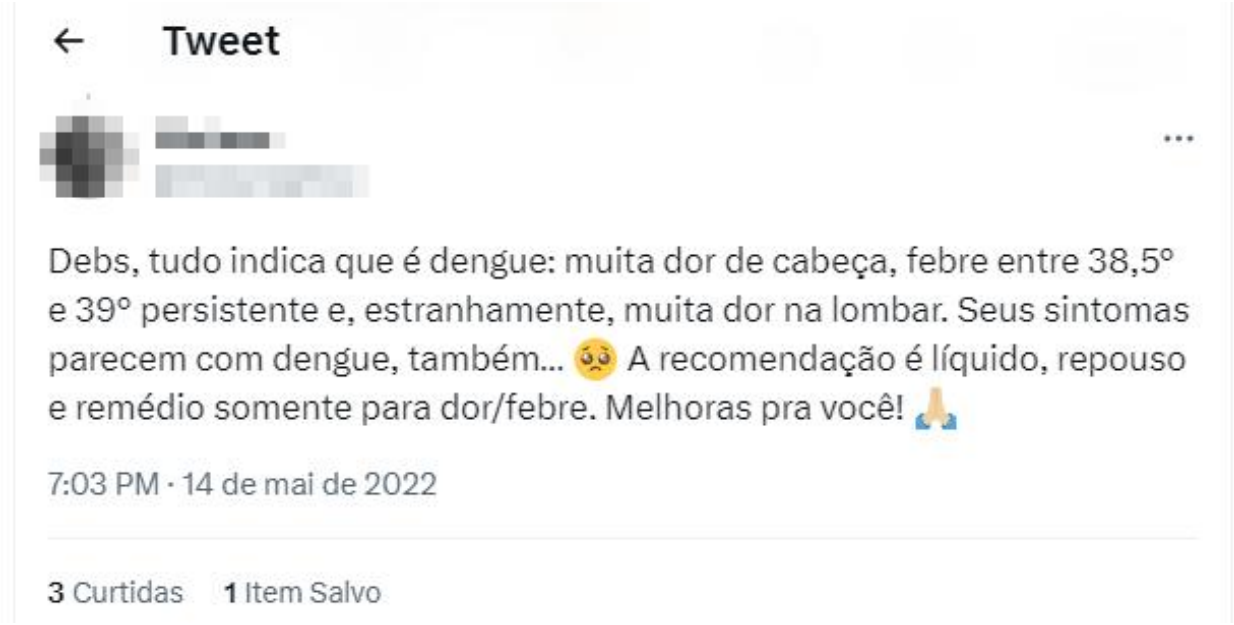


Figura V.12: Dengue: Formato Original do Tuíte na Plataforma Twitter.

Map Score	Score	Id do Conceito	Nome do Conceito	Nome Preferido	Matched Words	Quantidade de Matched Words	Nome Tipo Semântico (PT-BR)	Nome Grupo Semântico (PT-BR)	Nome do Vocabulário (PT-BR)
-804	-804	C0011311	Dengue	Dengue	dengue	1	Doença ou Síndrome	Distúrbios	MDRBPO,MSHPOR
-923	-923	C0018681	Dor de Cabeça	CEFALALGIA	dor,de,cabeça	3	Sinal ou Sintoma	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR
-827	-827	C0015967	FEBRE	Doenças Febris	febre	1	Sinal ou Sintoma	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR
-824	-875	C0024031	DOR LOMBAR	Dor lombar	dor,lombar	2	Sinal ou Sintoma	Distúrbios	MDRBPO, MSHPOR, WHOPOR
-824	-645	C0337443	Na+	Na+	na	1	Procedimento laboratorial	Procedimentos	Português do Brasil MedDRA
-764	-637	C0037088	Sintomas	Manifestações Clínicas	sintomas	1	Sinal ou Sintoma	Distúrbios	MeSH Português
-764	-804	C0011311	Dengue	Dengue	dengue	1	Doença ou Síndrome	Distúrbios	MDRBPO,MSHPOR
-735	-629	C0035253	Repouso	Descanso	repouso	1	Atividade Diária ou Recreativa	Atividades e Comportamentos	MeSH Português
-735	-629	C0013227	Remédio	Drogas	remédio	1	Substância Farmacológica	Produtos Químicos e Drogas	MeSH Português
-735	-629	C0030193	DOR	Dor	dor	1	Sinal ou Sintoma	Distúrbios	MDRBPO, MSHPOR, WHOPOR
-735	-795	C0015967	FEBRE	Doenças Febris	febre	1	Sinal ou Sintoma	Distúrbios	ICPCPOR, MDRBPO, MSHPOR, WHOPOR

Tabela V.7: Tuíte Mapeado referente a Figura V.12.

Além desta análise e demais observações, este tema apresentou os seguintes resultados: foram coletados um total de 55.326 tuítes, destes, o MetaMap mapeou 54.940 tuítes totalizando 331.199 registros mapeados.

O crescimento substancial dos registros em 2020, tanto para a Medicamentos da Classe Terapêutica de Enxaqueca quanto para a Dengue - Sinais e Sintomas, sugere que a pandemia de COVID-19 tenha influenciado o comportamento e os interesses das pessoas nas redes sociais, espe-

cialmente no Twitter. A disseminação da COVID-19 e as consequentes campanhas de informação e sensibilização podem ter elevado a consciência da população sobre questões de saúde em geral, incluindo temas como enxaqueca e dengue.

No entanto, enquanto a atenção à enxaqueca continuou crescendo em 2021, os registros sobre a dengue sofreram uma queda notável. É possível que a intensa cobertura midiática e os esforços contínuos para enfrentar a COVID-19 tenham ofuscado outras preocupações de saúde, como a dengue, em 2021.

Em conclusão, os padrões observados em ambos os experimentos refletem não apenas as dinâmicas inerentes às respectivas doenças, mas também o impacto mais amplo da COVID-19 no comportamento das mídias sociais e nas prioridades de saúde percebidas pelo público.

V.3 Modelagem e Inferência de Sinais de Farmacovigilância (SF) em Mídias Sociais

No processo de análise de tuítes para detecção de SF, uma série de técnicas e algoritmos de PLN e ML podem ser empregados. Neste trabalho foi abordado de maneira sucinta as técnicas de Modelos Baseados em Regras e REN. O REN foi aplicado pensando na possibilidade de utilizar modelo supervisionado, onde foi iniciado um experimento inicial detalhado na Subseção V.3.2.2.

O Modelo Supervisionado exige dados rotulados, que, no contexto atual não existem. Porém, é viável desenvolver um *corpus* anotado empregando técnicas de REN, e a partir dele, treinar classificadores para identificar relações entre entidades. Inicialmente, um classificador pode ser empregado para determinar a presença ou ausência de relação entre pares de entidades. Se uma relação for identificada, um segundo classificador, especializado em tipificar a natureza da relação, pode ser aplicado.

Nesta seção são apresentadas as seguintes tarefas: Carga e Novo Pré-processamento do *Corpus* de Tuítes e Aplicação de Modelos. Todas estas tarefas foram realizadas usando o *Google Colab*.

V.3.1 Carga e Novo Pré-processamento do *Corpus* de Tuítes

Nesta fase do trabalho, uma nova carga de dados e um novo pré-processamento foi realizado. Nesta carga foi aproveitado o *corpus* mapeado pelo MetaMap. O conteúdo do campo dos tuítes coletados originalmente é que foram novamente pré-processados.

Deste novo *corpus*, foram realizados os seguintes pré-processamentos: exclusão de registros duplicados; escolha de alguns tipos semânticos; textos transformados para minúsculos; remoções de *urls, hastags, emojis, stopwords*, pontuações dentre outras remoções. Uma amostra aleatória de 100 registros foi selecionada, aplicado correção ortográfica e substituição de palavrões.

Neste *corpus* contém apenas os tipos semânticos: Doença ou Síndrome; Sinal ou Sintoma; Substância Farmacológica; Parte do Corpo, Órgão ou Componente do Órgão; e Processo Mental,

estes foram escolhidos por serem os mais alinhados com a pesquisa de medicamentos de enxaqueca. Este *corpus* também contém uma coluna anotada manualmente dizendo se o registro é ou não um SF.

A etapa de correção ortográfica constitui um aspecto vital no processo de detecção de SFs. Para esta tarefa, foram testadas as bibliotecas da linguagem Python *SpellChecker*[3] e *autocorrect*[4] com a *speller* como classe principal. As bibliotecas são ferramentas simples e eficiente para realizar correção ortográfica em textos, com suporte a múltiplos idiomas, incluindo o português. Estas bibliotecas ajudam a reduzir ruídos nos dados textuais, garantindo que erros de digitação ou ortografia não prejudiquem a eficácia das técnicas analíticas subsequentes. Contudo, nenhuma das duas foram eficientes para este *corpus*. Então, foi necessário criar um dicionário de dados específico para remoção dos ruídos destes textos.

Por fim, a limpeza de palavrões foi realizada com a ajuda da biblioteca *profanity* do Python, à qual foram adicionadas palavras específicas do português brasileiro. Esta ação foi imprescindível para assegurar a adequação do conteúdo para análises acadêmicas e profissionais.

V.3.2 Aplicação de Modelos

Neste experimento inicial, foram utilizados dois modelos: o Modelo Baseado em Regras e o Reconhecimento de Entidade Nomeada (REN). Para execução de ambos, foi necessário a execução de todo o *pipeline* da Subseção V.3.1. Estes modelos são detalhados a seguir.

V.3.2.1 Modelo Baseado em Regras

Este modelo utilizou Expressões Regulares, uma estratégia antiga e comum em Modelo Baseado em Regras, embora apresente desafios significativos em termos de manutenção e flexibilidade. A complexidade de criar regras específicas para cada padrão textual é inegável, e muitas vezes, métodos alternativos podem ser mais eficientes. Por outro lado, o Modelo Baseado em Regras, exemplificado pelo Algoritmo 1, oferece uma abordagem interessante para inferir relações. No entanto, essa metodologia também apresenta limitações quanto à manutenibilidade.

O Algoritmo 1 apresentado implementa um método de detecção de possíveis RAMs em textos extraídos de mídias sociais. O processo pode ser descrito da seguinte forma:

1. Entradas do Algoritmo:

    - Conjunto T de tweets a serem analisados
    - Conjunto M de palavras-chave relacionadas a medicamentos

[3] `https://pypi.org/project/pyspellchecker`
[4] `https://pypi.org/project/autocorrect/`

- Conjunto R de palavras-chave indicativas de RAMs
- Conjunto E de termos de exclusão

2. Processo Principal:
   - Para cada texto, o algoritmo executa três verificações principais: a) Presença de termos relacionados a medicamentos b) Presença de termos indicativos de RAMs c) Ausência de termos de exclusão
3. Função VerificaPadraoRAM:
   - Normaliza o texto para minúsculo
   - Verifica a presença de medicamentos
   - Verifica a presença de termos indicativos de RAMs
   - Verifica a ausência de termos de exclusão
   - Retorna verdadeiro apenas se todas as condições são satisfeitas
4. Saída:
   - Conjunto T' contendo apenas os tweets que satisfazem todas as condições estabelecidas

V.3.2.2 Reconhecimento de Entidade Nomeada (REN)

A biblioteca *spaCy* foi utilizada neste trabalho para implementar a técnica de Reconhecimento de Entidade Nomeada (REN). Ela oferece modelos pré-treinados em português do Brasil em três tamanhos distintos: 'pt_core_news_sm', 'pt_core_news_md' e 'pt_core_news_lg'. Esses modelos variam em tamanho, cobertura de vocabulário e complexidade, sendo:

- **pt_core_news_sm**: Modelo compacto com aproximadamente 12 MB, adequado para tarefas mais simples e com restrições de memória.
- **pt_core_news_md**: Modelo intermediário, com 40 MB, que oferece melhor equilíbrio entre desempenho e eficiência computacional, sendo mais robusto para identificação de entidades e análises sintáticas.
- **pt_core_news_lg**: Modelo mais avançado, com 541 MB, treinado em um volume maior de dados de notícias e mídias, proporcionando maior cobertura e precisão na identificação de entidades.

Algorithm 1 Algoritmo de Detecção de RAMs em Textos de Mídias Sociais

Require: Conjunto de tuítes T
Require: Conjunto de palavras-chave de medicamentos M
Require: Conjunto de palavras-chave de RAMs R
Require: Conjunto de termos de exclusão E
Ensure: Subconjunto de tuítes T' com potenciais RAMs

function VerificaPadraoRAM(*tweet*)
    tweet ← Converter(*tweet* para minúsculo)
    $containsMed \leftarrow$ Falso
    $containsRAM \leftarrow$ Falso
    $containsExcl \leftarrow$ Falso
    for cada $med \in M$ **do**
        if *med* está contido em *tweet* **then**
            $containsMed \leftarrow$ Verdadeiro
            break
        end if
    end for
    for cada $ram \in R$ **do**
        if *ram* está contido em *tweet* **then**
            $containsRAM \leftarrow$ Verdadeiro
            break
        end if
    end for
    for cada $excl \in E$ **do**
        if *excl* está contido em *tweet* **then**
            $containsExcl \leftarrow$ Verdadeiro
            break
        end if
    end for
     return $(containsMed \land containsRAM \land \neg containsExcl)$
end function
$T' \leftarrow \emptyset$
for cada $tweet \in T$ **do**
    if VerificaPadraoRAM(*tweet*) **then**
        $T' \leftarrow T' \cup \{tweet\}$
    end if
end forreturn T'

Neste estudo, foi necessário realizar o *fine-tuning* do modelo para reconhecer entidades específicas, como 'MEDICAMENTO' e 'DATE', no padrão brasileiro, com o objetivo de aprimorar a detecção de Sinais de Farmacovigilância (SF). A inclusão de novas entidades, como 'MEDICAMENTO', mostrou-se essencial para melhorar a identificação de menções a medicamentos, facilitando a correlação dessas menções com possíveis SFs. Além disso, essa abordagem permite o mapeamento de relações sintáticas entre termos e a concatenação dessas informações com o tipo de conceito 'Sinal ou Sintoma' identificado pelo MetaMap, enriquecendo significativamente o *corpus* utilizado nesta pesquisa.

Foram realizados testes com os três modelos pré-treinados disponibilizados pela *spaCy*. O modelo intermediário 'pt_core_news_md' apresentou o melhor desempenho inicial na identificação das novas entidades ajustadas durante o processo de *fine-tuning*, destacando-se pela eficiência na detecção de 'MEDICAMENTO'. No entanto, optou-se por seguir o trabalho utilizando o modelo 'pt_core_news_lg', uma vez que, após o retreinamento com dados sintéticos, este apresentou maior cobertura e robustez, mesmo com custo computacional mais elevado. A base de treinamento do modelo 'pt_core_news_lg' proporcionou resultados mais consistentes devido à sua maior capacidade de generalização, o que contribuiu diretamente para a qualidade dos resultados obtidos nesta pesquisa.

V.3.3 Inferência de Sinais e Sintomas (SF)

Observou-se que, apesar de algumas tentativas de correções ortográficas automatizadas, a intervenção manual mostrou-se mais eficaz, principalmente em termos de assertividade na identificação de gírias, abreviações e termos regionalizados.

A Figura V.13 mostra alguns tuítes originais e as correções aplicadas pelas bibliotecas testadas. Percebe-se que nesta amostra todos os ajustes foram errados, alterando termos que são relevantes para esta pesquisa. Destaca-se o 'Id_Mapeado' 12414, ambas bibliotecas trocaram quase todos os nomes dos medicamentos para termos sem sentido.

Como essas bibliotecas não foram satisfatórias, foi utilizado um dicionário específico para ajustes desses erros.

A Figura V.14 traz vários *insights*. Os textos sublinhados aparentemente são SFs, conforme os Id_Mapeados 404 e 27893, destaques em vermelho são erros ou colocações não triviais que podem ser melhor trabalhados.

O resultado desse processo compreensivo de normalização dos dados foi satisfatório, estabelecendo uma base para as etapas de inferências das Subseções V.3.3.1 e V.3.3.2.

Id Mapeado	Twitter	SpellChecker	Speller
189	@XXX ai que triste n diga que tem alergia a dipirona :(((((ele resolve tanto quanto o cefaliv? Porque nossa <3 esse remédio	ai que triste n diga que tem alergia a dipirona ele resolve tanto quanto o cefaliv porque nossa 3 esse remédio	ai que triste n diga que tem alegria a dipirona ele resolve tanto quanto o cefaliv porque nossa 3 esse remédio
11184	Que Enxaqueca do cão 😩😩 Deito dói. Levanto dói . Cefaliv, summax e Dorflex e nem pra fazer cócegas, quiçá alívio 😔😣😭	que enxaqueca do cão deito dói levanto dói cefaliv **suma** e dorflex e nem pra fazer cócegas quiçá alívio	que enxaqueca do cão deito dói levanto dói cefaliv **suma** e dorflex e nem pra fazer cócegas quiçá alívio
12414	@XXX simm, tô tomando cefaliv, deocil e dramin b6 p miorar o enjoo, mas n tá adiantando mto	sim tô tomando cefaliv **desci** e **drama by** p **morar** o enjoo mas n tá adiantando **mão**	sim te tomando cefaliv **docil** e **bramia** b6 p **mirrar** o enojo mas n tá adiantando **máo**
15413	@XXX amiga quando eu tomei a segunda dose da minha eu fiquei com uma puta enxaqueca mas passou depois de algumas horas e um enxak	amiga quando eu tomei a segunda dose da minha eu fiquei com uma puta enxaqueca mas passou depois de algumas horas e um **entao**	amiga quando eu tomei a segunda dose da minha eu fiquei com uma **puxa** enxaqueca mas passou depois de algumas horas e um **enxu**
32645	@XXX1 @XXX2 @XXX3 @XXX4 O ator Bruce Lee teve choque anafilático após tomar o analgésico mais comum - ácido acetilsalicílico. Aspirina, AAS, Melhoral, muitos remédios são à base dele - mas uma pessoa em cada 40.000 tem alergia a ele.	o ator bruce ele teve choque anafilático após tomar o analgésico mais comum ácido acetilsalicílico aspirina **as melhorar** muitos remédios são à base dele mas uma pessoa em cada 40 000 tem alergia a ele	o ator **breve lhe** teve choque anafilático após tomar o analgésico mais comum ácido acetilsalicílico **aspirava as melhorar** muitos remédios são à base dele mas uma pessoa em cada 40 000 tem alegria a ele

Figura V.13: Comparativo dos Corretores Ortográficos.

Id Mapeado	Twitter	Twitter Normalizado
404	Cefaliv é o único remédio que da uma melhoradINHA na minha enxaqueca mas **TODA vez que eu tomo ele eu tenho pesadelos horríveis**	cefaliv é o único remédio que da uma melhoradinha na minha enxaqueca mas toda vez que eu tomo ele eu tenho pesadelos horríveis
8850	@XXX Quando **q** minha ataca é foda consigo nem ficar de olho aberto parece que tem alguma coisa martelando o **celebro** daí nesse caso tenho que ir no hospital. Ah não sei **si** vc toma algum remédio mais esse daqui o "cefaliv" foi o melhor remédio que já tomei ele é próprio para enxaqueca.	quando **que** minha ataca é **** consigo nem ficar de olho aberto parece que tem alguma coisa martelando o celebro daí nesse caso tenho que ir no hospital ah não sei si você toma algum remédio mais esse daqui o cefaliv foi o melhor remédio que já tomei ele é próprio para enxaqueca
20387	Dois dias e duas noites com enxaqueca e **overdose** de Dipirona com Naramig mas até o momento sem dor. **A boca tá amarga de tanto remédio que tomei** rs 🙏🙏🙏🙏🙏🙏🙏	dois dias e duas noites com enxaqueca e overdose de dipirona com naramig mas até o momento sem dor a boca tá amarga de tanto remédio que tomei rs
25023	@XXX @XXX O sumax eu já tomei qnd tive crise de enxaqueca, **tô** com umas crises de dor de cabeça tem uns três meses e tô penando. O paracetamol + codeína é top mas meus comprimidos acabaram. Obrigada mesmo assim gente 🙏❤️	o sumax eu já tomei quando tive crise de enxaqueca **estou** com umas crises de dor de cabeça tem uns três meses e estou penando o paracetamol codeína é top mas meus comprimidos acabaram obrigada mesmo assim gente
27893	**qm q** indica sumax pra enxaqueca meu médico é louco sério a dor cessou mas **agora sinto cada veia do meu corpo se esticando internamente minha pressão tá baixa galera acho q vo morrer galera esotu morendo**	**** que indica sumax para enxaqueca meu médico é louco sério a dor cessou mas agora sinto cada veia do meu corpo se esticando internamente minha pressão está baixa galera acho que vou morrer galera esotu morendo

Figura V.14: Normalização dos Textos.

V.3.3.1 Inferência em Modelos Baseados em Regras

Na amostra de 100 registros, após aplicado a regra detalhada na Subseção V.3.2.1, o código trouxe como resultado 32 registros, porém, apenas 16 destes eram de fato SF e estes estavam anotados. Apesar de errar em 50% destes casos, a regra está bem simples e é possível refinamentos a fim de atingir um melhor *score*.

V.3.3.2 Aplicando (REN)

Ao aplicar a biblioteca *spaCy* no *corpus* foi selecionado um registro de exemplo para análise, apresentado a seguir:

```
'naramig me tira a enxaqueca mas me deixa tão indisposta com umas dores
estranhas pelo corpo ****'
```

A Figura V.15 reconheceu apenas uma entidade (PER). Contudo as relações e dependências geradas conforme a Figura V.16 podem ser trabalhadas e utilizadas em um classificador como discutido anteriormente.

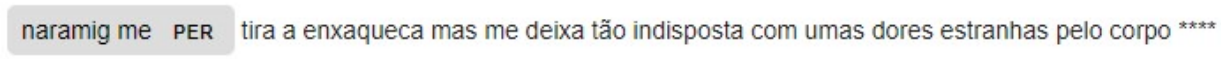


Figura V.15: Entidades Identificadas.

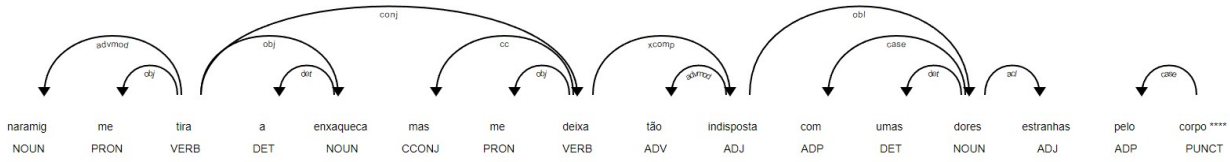


Figura V.16: Dependências e Relacionamentos Identificados.

Outro tuíte (apresentado a seguir) foi selecionado para análise. Foram identificadas apenas as entidades 'jesus' e '****' (palavra sensível), com o mesmo modelo retreinado com o *spaCy*, foram identificadas mais duas entidades, contudo, percebe-se que a entidade MISC não foi identificada nesse novo modelo conforme a Tabela V.8.

```
'e após mais de 50 horas com enxaqueca no meio da tarde após dois cefalium
jesus em cápsulas ela teve fim desesperador dessa vez foi **** uma ressaca que
desencadeou uma enxaqueca e quase me matou'
```

Termo	pt_core_news_lg	pt_core_news_lg Retreinado
jesus	PER	PER
****	MISC	-
50	-	DATE
cefalium	-	MEDICAMENTO

Tabela V.8: Resultado dos modelos *spaCy* aplicados no tuíte de Id Mapeado 29865.

Capítulo VI Conclusões e Trabalhos Futuros

As terminologias médicas do português do Brasil são bem maiores do que em inglês devido seu vasto vocabulário. As ferramentas e vocabulários biomédicos são na maioria voltadas para língua inglesa. Este trabalho fez uma revisão da literatura na área de Processamento de Linguagem Natural (PLN) no apoio à farmacovigilância e posterior experimento utilizando a ferramenta de PLN MetaMap visando identificar SFs, RAMs e EAs em tuítes.

Dois experimentos foram realizados com um Modelo de Dados em Português do Brasil, criado a partir do MetaMap Data File Builder (DFB) UMLS. O mapeamento de textos coletados no Twitter foi feito na ferramenta MetaMap, a fim de auxiliar profissionais de área médica e empresas de fármacos na obtenção e correlação de informações relacionadas as RAMs que podem ou não ter sido capturadas pelos meios oficiais de notificação.

No experimento Medicamentos da Classe Terapêutica de Enxaqueca, foi verificado que do total de tuítes enviados para serem mapeados (35.399), aproximadamente 69.44% foram mapeados. Este é um número razoável e interessante considerando quantidade de vocabulários pequena disponível em português. No experimento sobre Dengue - Sinais e Sintomas, foi verificado que dos 54.943 tuítes enviados para serem mapeados, 54.940 foram identificados. Essa quase totalidade se deu pois o termo "dengue" estava em todos os tuítes. Alguns mapeamentos não representam a realidade do vocabulário em português, contudo, o resultado foi bem satisfatório.

O desenvolvimento deste trabalho passou por diversos impasses devido à complexidade das ferramentas escolhidas:

- todas as ferramentas de mapeamento biomédico deste trabalho foram criadas para tratar especificamente termos em inglês, apesar de possuírem possibilidades de trabalhar com outros idiomas, muitas opções não funcionam;
- ajustes e atualizações foram sendo realizadas no MetaMap e DFB durante o projeto;
- a instalação das ferramentas, na maior parte em java é toda manual. Há mais passos a serem realizados quando o modelo a ser tratado difere do inglês. Além de mais passos, vários *scripts* precisam ser ajustados manualmente antes de sua execução;
- o material de referência para instalação são antigos, desatualizados e pouco claros.

No entanto, a realização deste trabalho verificou várias oportunidades de melhorias, tais como:

- utilizar métodos de ML para inferir se é há SF em um tuíte;
- criar uma imagem no Docker Hub com um modelo de dados em português do Brasil para testes por outros pesquisadores, pois a instalação e geração de um modelo para trabalhar é muito custosa e demorada;
- como os programas são de código aberto, criar um processo de automatização para a instalação em português ser mais facilitada;
- após identificado medicamentos, validar se estes estão ativos e se a comunidade fora do Brasil mantém o medicamento liberado ou se tem alguma restrição;
- após identificado uma RAM, validar esta com a base de reações da Agência Nacional de Vigilância Sanitária (ANVISA) "VigiMed_Reacoes";
- pegar um mesmo *"corpus"* com seus resultados mapeados em português do Brasil, traduzir este *"corpus"* para o inglês e mapear novamente com a base UMLS em inglês e comparar os resultados. Espera-se uma quantidade maior de termos mapeados em inglês, conforme trabalhos identificaram [Chiaramello et al., 2016];
- explorar a possibilidade de expandir o léxico português do MetaMap, adicionando termos e sinônimos, inspirado no trabalho de Campillos-Llanos [2019] que ampliou um léxico médico em espanhol;

Este trabalho apresentou contribuições significativas para a área de farmacovigilância, especialmente no contexto do português do Brasil, ao propor um *pipeline* de PLN para a detecção de Sinais de Farmacovigilância (SFs) em textos provenientes de mídias sociais. Entre as principais contribuições, destacam-se: a criação pioneira de uma base UMLS adaptada ao português brasileiro; o uso da ferramenta MetaMap para o mapeamento de conceitos biomédicos; e a aplicação de modelos de ML e abordagens baseadas em regras para a identificação de EAs e RAMs em tuítes.

A relevância deste estudo se evidencia na ampliação do conhecimento sobre a aplicação do MetaMap em português do Brasil e no desenvolvimento de um DFB mais robusto e de instalação facilitada, fruto de ajustes e sugestões aceitos pela equipe de desenvolvimento da NLM. Este trabalho também forneceu um MaDIB como sistema replicável e adaptável, com grande potencial de uso em futuras pesquisas.

Além disso, os dois artigos publicados durante a realização desta pesquisa reforçam essas contribuições. O primeiro artigo, Sousa et al. [2024], apresentou uma análise detalhada das notificações de RAMs no Brasil, utilizando dados disponibilizados pelo sistema VigiMed, destacando padrões

regionais e a predominância de notificações realizadas por farmacêuticos. Já o segundo artigo, Perciliano et al. [2023], detalhou o processo experimental de uso do MetaMap para mapear conceitos médicos extraídos de tuítes, demonstrando a viabilidade do uso de mídias sociais como fonte alternativa para a farmacovigilância.

Portanto, os resultados desta pesquisa oferecem não apenas um avanço metodológico no uso de PLN para a farmacovigilância, mas também contribuições práticas e replicáveis que poderão ser empregadas para melhorar a identificação e monitoramento de RAMs, promovendo a segurança no uso de medicamentos no Brasil e potencialmente em outros países.

Referências

Abbas, A., Lee, S., and Afzal, M. (2019). Meaningful information extraction from unstructured clinical documents. *Proc. Asia Pac. Adv. Netw*, 48:42–47. , 42, 43

Aguiar, T. O. D. d., Capuccho, H. C., Varallo, F. R., and Mastroianni, P. d. C. (2015). Notificações espontâneas de eventos adversos a medicamentos hospitalares: estudo piloto. *Revista Ciência em Extensão*, pages 46–61. 32

Allam, N., Audeh, B., Jaulent, M.-C., and Bousquet, C. (2021). Visualising patterns associated with adverse drug reactions in french forums. *Studies in health technology and informatics*. 22

ANVISA (2022). O que é farmacovigilância? 30, 31

Aronson, A. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, page 17 – 21. Cited by: 1353. 69, 70, 71, 88, 89, 140

Aronson, A. R. (1996). Metamap technical notes. 48, 49, 138, 139, 140, 141, 142, 143

Assurance, W. H. O. Q. and of Medicines Team, S. (2002). Safety of medicines: a guide to detecting and reporting adverse drug reactions: why health professionals need to take action. 17, 31

Avong, Y. K., Jatau, B., Gurumnaan, R., Danat, N., Okuma, J., Usman, I., Mordi, D., Ukpabi, B., Kayode, G. A., Dutt, S., et al. (2018). Addressing the under-reporting of adverse drug reactions in public health programs controlling hiv/aids, tuberculosis and malaria: A prospective cohort study. *PLoS One*, 13(8):e0200810. 20

Azevedo, A. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*. 35

Barsi, Maysa; Borja, A. (2005). A importância da farmacovigilância e os desafios de sua implementação no brasil. 27, 29, 30

Bhattacharya, M., Snyder, S., Malin, M., Truffa, M. M., Marinic, S., Engelmann, R., and Raheja, R. R. (2017). Using social media data in routine pharmacovigilance: a pilot study to identify safety signals and patient perspectives. *Pharmaceutical Medicine*, 31:167–174. 34

Bompelli, A., Silverman, G., Finzel, R., Vasilakes, J., Knoll, B., Pakhomov, S., and Zhang, R. (2020). Comparing nlp systems to extract entities of eligibility criteria in dietary supplements clinical trials using nlp-adapt. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings*, page 67–77, Berlin, Heidelberg. Springer-Verlag. 22

Bousquetab, C., Souvignetab, J., Merabtid, T., and Rodriguesabc, M. (2012). Method for mapping the french ccam terminology to the umls metathesaurus. *Quality of Life Through Quality of Information: Proceedings of MIE2012*, 180:164. 75

Brasil, T. S. (2022). Farmacovigilância - série educação farmacêutica em oncologia. 30, 31

Brennan, P. F. and Aronson, A. R. (2003). Towards linking patients and clinical information: detecting umls concepts in e-mail. *Journal of Biomedical Informatics*, 36(4-5):334–341. 77

Campillos-Llanos, L. (2019). First steps towards building a medical lexicon for spanish with linguistic and semantic information. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 152–164, Florence, Italy. Association for Computational Linguistics. 22, 23, 120

Cardoso, M. A. and Amorim, M. A. L. (2013). A farmacovigilância e sua importância no monitoramento das reações adversas a medicamentos. *Revista Saúde e Desenvolvimento*, 4(2):33–56. 18, 19, 20, 21, 27, 29

Carrero, F., Cortizo, J. C., and Gómez, J. M. (2008). Building a spanish mmtx by using automatic translation and biomedical ontologies. In Fyfe, C., Kim, D., Lee, S.-Y., and Yin, H., editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2008*, volume 5326 LNCS, pages 346–353, Berlin, Heidelberg. Springer Berlin Heidelberg. 23, 70

Casillas, A., de Ilarraza, D., Pérez, A., Oronoz, M., Gojenola, K., and Santiso, S. (2016a). Learning to extract adverse drug reaction events from electronic health records in spanish. *Expert Systems With Applications*. 76

Casillas, A., Díaz de Ilarraza, A., Gojenola, K., Mendarte, L., Oronoz, M., Peral, J., and Pérez, A. (2016b). Deteami research-transference project: natural language processing technologies to the aid of pharmacy and pharmacosurveillance. *Procesamiento del Lenguaje Natural*, 57:155–158. 75

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *Computer Science, Business.* , 35, 36

Chen, X., Faviez, C., Schuck, S., Lillo-Le-Louët, A., Texier, N., Dahamna, B., Huot, C., Foulquié, P., Pereira, S., Leroux, V., et al. (2018). Mining patients' narratives in social media for phar-

macovigilance: adverse effects and misuse of methylphenidate. *Frontiers in pharmacology*, 9:541. 19, 23

Chiaramello, E., Pinciroli, F., Bonalumi, A., Caroli, A., and Tognola, G. (2016). Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of metamap annotation of italian medical notes. *Journal of Biomedical Informatics*, 63:22–32. 21, 22, 120

Chowdhury, G. G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, 37:51–89. 36, 37

Christian, A. and Emmanuel, P. (2009). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, 5. 38

Costa, N. N., de Camargo Silva, A. E. B., de Lima, J. C., de Sousa Barbosa, M. R., de Freitas, J. S., and Bezerra, A. L. Q. (2016). O retrato dos eventos adversos em uma clínica médica: análise de uma década. *Cogitare Enfermagem*, 21(5). 32

Davanzo, L. and Moreira, W. (2014). O vocabulário controlado como ferramenta do processo de organização e recuperação da informação. 41

de Moura, D. C. N., Pinto, J. R., Martins, P., de Arruda Pedrosa, K., and Carneiro, M. d. G. D. (2016). Revisão integrativa da literatura sobre o aumento no consumo de psicotrópicos em transtornos mentais como a depressão. *SANARE-Revista de Políticas Públicas*, 15(2). 106

De Rosa, M., Fenza, G., Gallo, A., Gallo, M., and Loia, V. (2021). Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating twitter and pubmed. *Future Generation Computer Systems*, 114:394–402. 22

Delbecque, T. and Zweigenbaum, P. (2007). Metacode: A lightweight umls mapping tool. In *Artificial Intelligence in Medicine*, volume 4594 LNAI, pages 242–246. 22

DeLucia, R., Planeta, C. S., and Lepsch, L. (2016). Capítulo 4 farmacoterapêutica. desenvolvimento de fármacos farmacovigilância. *Farmacologia Integrada*, page 52. 32

Deléger, L., Grouin, C., and Zweigenbaum, P. (2010). Extracting medication information from french clinical texts. *Studies in Health Technology and Informatics*, 160:949–953. 22

Di Felippo, A. and Dias-da Silva, B. C. (2009). O processamento automático de línguas naturais enquanto engenharia do conhecimento linguístico. 36

Donato, M. A. M., Ribeiro, E. L., de Albuquerque Silva, Y. J., and Peixoto, C. A. (2013). Além da disfunção erétil. *Caderno de Graduação-Ciências Biológicas e da Saúde-UNIT-PERNAMBUCO*, 1(2):11–22. 20

Duval, F. V. and da Silva, F. A. B. (2019). Mining in twitter for adverse events from malaria drugs: The case of doxycycline. *Cadernos de Saude Publica*, 35. 19, 20, 21, 22, 24, 51

Dávila Rodríguez, M. J. (2023). Embase, búsqueda de información científica. 137

Ethnologue (2022). What are the top 200 most spoken languages? 70

Feng, C., Le, D., and McCoy, A. B. (2019). Using electronic health records to identify adverse drug events in ambulatory care: a systematic review. *Applied clinical informatics*, 10(01):123–128. 78

Ferreira, L., Teixeira, A., and Cunha, J. P. (2013). *Medical Information Extraction in European Portuguese*. Universidade de Aveiro. 22

Figueiredo, T. A., Fialho Neto, R. G., and Magalhães, J. L. d. (2021). The public production of medicines in brazil. *Ciência & Saúde Coletiva*, 26:3423–3434. 18

Fonseca, E. B., Chiele, G. C., Vieira, R., and Vanin, A. A. (2015). Reconhecimento de entidades nomeadas para o português usando o opennlp. *Anais do ENIAC 2015, 2015, Brasil.*, page 5. 39

Gaino, L. V., Souza, J. d., Cirineu, C. T., and Tulimosky, T. D. (2018). O conceito de saúde mental para profissionais de saúde: um estudo transversal e qualitativo. *SMAD. Revista eletrônica saúde mental álcool e drogas*, 14:108 – 116. 17

Golder, S., O'Connor, K., Wang, Y., Hernandez, G. G., et al. (2023). The role of social media for identifying adverse drug events data in pharmacovigilance: protocol for a scoping review. *JMIR Research Protocols*, 12(1):e47068. 22

Grus, J. (2016). *Data science do zero*. Alta Books. 38

Guo, D., Duan, G., Yu, Y., Li, Y., Wu, F.-X., and Li, M. (2020). A disease inference method based on symptom extraction and bidirectional long short term memory networks. *Methods*, 173:75–82. 23

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926. 57

Hanauer, D. A., Saeed, M., Zheng, K., Mei, Q., Shedden, K., Aronson, A., and Ramakrishnan, N. (2014). Applying metamap to medline for identifying novel associations in a large clinical dataset: A feasibility analysis. *Journal of the American Medical Informatics Association*, 21:925–937. 23

Hellrich, J., Schulz, S., Buechel, S., and Hahn, U. (2015). Jufit: A configurable rule engine for filtering and generating new multilingual umls terms. In *AMIA Annual Symposium Proceedings*, volume 2015, page 604. American Medical Informatics Association. 75

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., and Declerck, T. (2013). The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920. 73

Higgins, J. P. and Green, S. (2011). *Cochrane handbook for systematic reviews of interventions*, volume 4. John Wiley & Sons. 57

Iyer, S. V., Harpaz, R., LePendu, P., Bauer-Mehren, A., and Shah, N. H. (2013). Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association*, 21(2):353–362. 23

Jeiziner, C., Suter, K., Wernli, U., Barbarino, J., Gong, L., Whirl-Carrillo, M., Klein, T., Szucs, T., Hersberger, K., and Meyer Zu Schwabedissen, H. (2021). Pharmacogenetic information in swiss drug labels–a systematic analysis. *The pharmacogenomics journal*, 21(4):423–434. 80

Jensen, P., Jensen, L., and Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395–405. 23

Joshi, C., Attar, V. Z., and Kalamkar, S. P. (2022). An unsupervised topic modeling approach for adverse drug reaction extraction and identification from natural language text. In Tiwari, S., Trivedi, M. C., Kolhe, M. L., Mishra, K., and Singh, B. K., editors, *Advances in Data and Information Sciences*, pages 505–514, Singapore. Springer Singapore. 17, 18, 32

Jung, E., Jain, H., Sinha, A. P., and Gaudioso, C. (2021). Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis. *Health Informatics Journal*, 27. 23

Kang, N., Singh, B., Afzal, Z., van Mulligen, E. M., and Kors, J. A. (2013). Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 20(5):876–881. 38

Kim, J., Kwon, S., Ko, Y., and Seo, J. (2017). A method to generate a machine-labeled data for biomedical named entity recognition with various sub-domains. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*. 73

Klauck, R. et al. (2022). *Mapeamento de depósitos de patentes referentes à atividade terapêutica de compostos químicos e de estudos clínicos em andamento no Brasil*. PhD thesis, Universidade Federal de Santa Catarina. 27, 29

Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., and Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158. 22

Laan, R. H. v. d. (2002). *Tesauro e terminologia: uma inter-relação lógica*. PhD thesis, Universidade Federal do Rio Grande do Sul. 41

Lardon, J., Abdellaoui, R., Bellet, F., Asfari, H., Souvignet, J., Texier, N., Jaulent, M.-C., Beyens, M.-N., Burgun, A., and Bousquet, C. (2015). Adverse drug reaction identification and extraction in social media: A scoping review. *J Med Internet Res*, 17(7):e171. 22

Li, F., Liu, W., and Yu, H. (2018). Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. *JMIR Med Inform*, 6(4):e12159. 21

Liddy, E. D. (2001). Natural language processing. *Encyclopedia of Library and Information Science*. 37

Lima, R. F., Machado, A. V., and Bezerra, M. L. (2020). Uso seguro de medicamentos na atenção primária à saúde no brasil: revisão integrativa. *Revista Enfermagem Atual*, 92(30):art.658. 20

Luo, Y., Thompson, W. K., Herr, T. M., Zeng, Z., Berendsen, M. A., Jonnalagadda, S. R., Carson, M. B., and Starren, J. (2017). Natural language processing for ehr-based pharmacovigilance: a structured review. *Drug safety*, 40(11):1075–1089. 21, 22

Mahendran, D. and McInnes, B. T. (2021). Extracting adverse drug events from clinical notes. *CoRR*, abs/2104.10791. 34

Maigetter, K., Pollock, A. M., Kadam, A., Ward, K., and Weiss, M. G. (2015). Pharmacovigilance in india, uganda and south africa with reference to who's minimum requirements. *International journal of health policy and management*, 4(5):295. 20

McNaught, C. and Lam, P. (2010). Using wordle as a supplementary research tool. *Qualitative Report*, 15(3):630–643. 38

Melo, J. R. R., Duarte, E. C., Moraes, M. V. d., Fleck, K., Silva, A. S. d. N. e., and Arrais, P. S. D. (2021). Reações adversas a medicamentos em pacientes com covid-19 no brasil: análise das notificações espontâneas do sistema de farmacovigilância brasileiro. *Cadernos de Saúde Pública*, 37(1). 19, 20, 21

Merabti, T., Abdoune, H., Letord, C., Sakji, S., Joubert, M., and Darmoni, S. J. (2011). Mapping the atc classification to the umls metathesaurus: some pragmatic applications. In *Patient Safety Informatics*, pages 206–213. IOS Press. 74

Meyboom, R. H., Egberts, A. C., Edwards, I. R., Hekster, Y. A., de Koning, F. H., and Gribnau, F. W. (1997). Principles of signal detection in pharmacovigilance. *Drug safety*, 16:355–365. 18, 33

Meystre, S. and Haug, P. J. (2005). Evaluation of medical problem extraction from electronic clinical documents using metamap transfer (mmtx). *Studies in health technology and informatics*, 116:823–828. 22, 23, 72, 85

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2007). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform*, pages 128–144. 21

Monteiro, T. F. (2012). *Desenvolvimento de novo método ex vivo para estudo da permeabilidade de fármacos utilizando epitélio intestinal de rã-touro (Rana catesbeiana)*. PhD thesis, Universidade de São Paulo. 28

Nasciutti, P. R. and Carvalho, R. D. O. A. (2012). Seminários aplicados desenvolvimento de novos fármacos. 18, 27, 28, 29

National Library of Medicine (US) (2009a). Umls® reference manual. Internet. , 44, 45, 46

National Library of Medicine (US) (2009b). UMLS® Reference Manual. Internet. 47

Nguyen, T., Larsen, M. E., O'Dea, B., Phung, D., Venkatesh, S., and Christensen, H. (2017). Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics*, 102:130–137. 22

Nguyen, V., Yip, H. Y., and Bodenreider, O. (2021). Biomedical vocabulary alignment at scale. In *Proceedings of the Web Conference 2021*, WWW '21, page 2672–2683, New York, NY, USA. Association for Computing Machinery. , 42

Niazi, S. K. (2019). Waiver of in vivo bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on a biopharmaceutics classification system: Guidance for industry. In *Handbook of Pharmaceutical Manufacturing Formulations, Third Edition*, pages 27–35. CRC Press. 28

Oliveira, D., Do Amaral, F., and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields a tool for the named entity recognition using conditional random fields. *Linguamática*. 39

Olivier, B. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(D267-70). 40, 42, 44, 47

OMS, O. (2005). Segurança dos medicamentos: um guia para detectar e notificar reações adversas a medicamentos. *Organização Mundial da Saúde. Brasília: OPAS/OMS*, page 18. 32, 33

Opitz, J., Wein, S., and Schneider, N. (2024). Natural language processing relies on linguistics. 36

Organization, W. H. et al. (2020). Dengue and severe dengue. geneva, switzerland: Who. Internet. 107

Oronoz, M., Casillas, A., Gojenola, K., and Pérez, A. (2013). Automatic annotation of medical records in spanish with disease, drug and substance names. In Ruiz-Shulcloper, José, di Baja, S., and Gabriella, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8259, pages 536–543, Berlin, Heidelberg. Springer Berlin Heidelberg. 22

Osborne, J. D., Lin, S., Zhu, L. J., and Kibbe, W. A. (2006). Mining biomedical data using metamap transfer (mmtx) and the unified medical language system (umls). *Methods in Molecular Biology*, 408:153–161. 72

Pappa, D. and Stergioulas, L. K. (2019). Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. *International Journal of Data Science and Analytics*, 8(2):113–135. 34

Pepe, V. L. E. and Novaes, H. M. D. (2020). Sistema nacional de farmacovigilância no brasil e em portugal: semelhanças, diferenças e desafios. *Cadernos de Saúde Pública*, 36(7):e00043019. 17

Perciliano, L., Belloze, K., and Pintor, P. (2023). Twitter adverse event detection using metamap for brazilian portuguese. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 47–51. IEEE. 40, 70, 121

Piatetsky, G. (2014). Crisp-dm, ainda é a principal metodologia utilizada para análise, mineração de dados ou projetos de ciência dos dados. 36

Pilipiec, P., Liwicki, M., and Bota, A. (2022). Using machine learning for pharmacovigilance: a systematic review. *Pharmaceutics*, 14(2):266. 78

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W., and Savova, G. (2014). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154. 23

Pustejovsky, J. and Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. "O'Reilly Media, Inc.". 38

Rogers, W., Lang, F.-M., and Gay, C. (2012). Metamap data file builder. *National Library of Medicine, US*. 22, 39, 40

Rozenfeld, S. and Rangel, I. (1988). A farmacovigilância. *Cadernos de Saúde Pública*, 4:336–341. 30, 31, 32

Ruiz, M. E. (2006). Combining image features, case descriptions and umls concepts to improve retrieval of medical images. In *AMIA Annual Symposium Proceedings*, volume 2006, pages 674–678. American Medical Informatics Association. 23, 77

Santos, R. E., Souza, E. P., Correia-Neto, J. S., Magalhães, C. V., and Vilar, G. (2015). Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. *Revista de Sistemas e Computação-RSC*, 4(2). 37

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212. 33

Scripcaru, G. (2018). *Eventos Adversos a Medicamentos em Contexto de Internamento Hospitalar em Portugal Continental*. PhD thesis, Universidade NOVA de Lisboa (Portugal). 32

Segre, M. and Carvalho, Ferraz, F. C. (1997). O conceito de saúde. *Revista de Saúde Pública*, 31:538–542. 17

Shah, A., Martinez, C., and Hemingway, H. (2012). The freetext matching algorithm: A computer program to extract diagnoses and causes of death from unstructured text in electronic health records. *BMC Medical Informatics and Decision Making*, 12. 18, 32

Shang, N., Xu, H., Rindflesch, T., and Cohen, T. (2014). Identifying plausible adverse drug reactions using knowledge extracted from the literature. *Journal of Biomedical Informatics*, 52:293–310. 22

Shannon, G. J., Rayapati, N., Corns, S. M., and Wunsch, D. C. (2021). Comparative study using inverse ontology cogency and alternatives for concept recognition in the annotated national library of medicine database. *Neural Networks*, 139:86–104. 23, 39

Silva, R. d. A. et al. (2021). Uma metodologia para criação de um corpus textual adequada ao reconhecimento de entidades nomeadas em português. 38

Skeppstedt, M., Kvist, M., Nilsson, G. H., and Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics*, 49:148–158. 22

Sloane, R., Osanlou, O., Lewis, D., Bollegala, D., Maskell, S., and Pirmohamed, M. (2015). Social media and pharmacovigilance: a review of the opportunities and challenges. *British journal of clinical pharmacology*, 80(4):910–920. 22

Sousa, A. B., Perciliano, L., and Belloze, K. (2024). Explorando dados de notificações em farmacovigilância. *Anais do XVIII Brazilian e-Science Workshop (BRESCI 2024)*. 120

Stefanelli, L. C. J. (2011). Experimentação animal: considerações éticas, científicas e jurídicas. *Ensaios e Ciência: Ciências Biológicas, Agrárias e da Saúde*, 15(1):187–206. 28

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE second international conference on social computing*, pages 177–184. IEEE. 100

Trentin, K. M., Andrade, S. C., Renner, J. D. P., Krug, S. F., and Garcia, E. L. (2019). Segurança do paciente no uso de antibiótico hospitalar: uma revisão da literatura. *Revista de Epidemiologia e Controle de Infecção*, 9(4). 106

Vatian, A., Dobrenko, N., Makarenko, A., Nigmatullin, N., Vedernikov, N., Vasilev, A., Stankevich, A., Gusarova, N., and Shalyto, A. (2018). Adaptation of algorithms for medical information retrieval for working on russian-language text content. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21*, pages 106–114. Springer. 76

Vaz, R., Santos, C., and Correia, C. (2016). Promoting adverse drug reaction reporting: comparison of different approaches. *Rev. Saúde Pública, 2016a*, 50(0):1–9. 20

Vo, T. H., Nguyen, N. T. K., Kha, Q. H., and Le, N. Q. K. (2022). On the road to explainable ai in drug-drug interactions prediction: A systematic review. *Computational and Structural Biotechnology Journal*, 20:2112–2123. 79

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *Journal of the American Medical Informatics Association*, 16:328–337. 23

Yokotsuka, M., Aoyama, M., and Kubota, K. (2000). The use of a medical dictionary for regulatory activities terminology (meddra) in prescription-event monitoring in japan (j-pem). *International journal of medical informatics*, 57 2-3:139–53. 22, 74

Young, I. J. B., Luz, S., and Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International journal of medical informatics*, 132:103971. 79

Zhang, M., Zhang, M., Ge, C., Liu, Q., Wang, J., Wei, J., and Zhu, K. Q. (2018). Automatic discovery of adverse reactions through chinese social media. *Data Mining and Knowledge Discovery*, 33:848–870. 22

Zhang, X., Liu, Y., and Chen, J. (2018). Network embedding based on external word vectors. In *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*, pages 1–2. 51

Zucchetti, C. and Morrone, F. B. (2012). Perfil da pesquisa clínica no brasil. *Clinical & Biomedical Research*, 32(3). 27, 28, 29

Zuccon, G., Holloway, A., Koopman, B., and Nguyen, A. (2013). Identify disorders in health records using conditional random fields and metamap aehrc at share/clef 2013 ehealth evaluation lab task 1. In *Proceedings of the CLEF 2013 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*, pages 1–8. The CLEF Initiative (Conference and Labs of the Evaluation Forum). 77

Apêndice A Vocabulários Biomédicos

O Apêndice A apresenta os vocabulários utilizados nessa pesquisa e outros relevantes para esta temática.

- **Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10)**[1]

  A Décima Revisão CID-10 foi desenvolvida e é mantida pela OMS. O Centro Nacional de Estatísticas de Saúde dos EUA (NCHS) é o Centro Colaborador para a Família de Classificações Internacionais na América do Norte e responsável por supervisionar o uso da CID. É usada para promover comparabilidade internacional na coleta, processamento, classificação e apresentação de estatísticas de mortalidade de certidões de óbito. Iniciada 400 anos atrás, passa por diversas atualizações. Consiste em: listas contendo títulos e códigos de causa de morte; termos de inclusão e exclusão para títulos de causa de morte; um índice alfabético de doenças e natureza da lesão, causas externas de lesão, tabela de medicamentos e produtos químicos e descrição, diretrizes e regras de codificação. Trabalha também com diagnóstico padrão para epidemiologia, gestão da saúde e finalidades clínicas. Aprovada pela OMS em maio de 1990, entrou em uso pelos países membros a partir de 1994. O Brasil possui um centro colaborador para o CID-10[2] com um foco maior em doenças. É atualizada anualmente em janeiro, o *Metathesaurus* possui diversas versões e variações do CID-10 e nem sempre está atualizado.

- **Descritores em Ciências da Saúde (DeCS)**[3]

  Criado pelo Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME) para indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos e outros materiais. Usado especificamente para IR de assuntos da literatura científica nas bases de informação na Biblioteca Virtual em Saúde (BVS) como LILACS, MEDLINE e outras. Foi criado baseado no vocabulário MeSH para o espanhol, francês, inglês e português. Além de conter os termos do MeSH, contém as áreas de Saúde Pública, Homeopatia, Ciências e Saúde e Vigilância Sanitária do Brasil. O DeCS/MeSH é um vocabulário dinâmico totali-

[1] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICD10/index.html`
[2] `http://www2.datasus.gov.br/cid10/V2008/cid10.htm`
[3] `https://decs.bvs.br/`

zando 34.387 registros de descritores e 77 qualificadores, sendo 30.454 registros de descritores provenientes do MeSH e 3.310 exclusivamente do DeCS.

As categorias exclusivas e seus totais de descritores são: Ciência e Saúde (330); Homeopatia (1.907); Medicinas Tradicionais, Complementares e Integrativas (110); Saúde Pública (2761); e Vigilância em Saúde (818). Possui um processo constante de crescimento e mudança, dentre alterações, substituições e criações de novos termos ou áreas, dados de março de 2023. É mais voltado para questões da América Latina visando atender esta realidade cultural. Com a atualização do MeSH, o DeCS também é atualizado mas não pertence ao UMLS. O Brasil possui setores para atualização anual deste vocabulário.

- **ICPC Portuguese**[4]

  A Classificação Internacional de Atenção Primária (ICPC) foi criada pela Organização Mundial de Médicos de Família (Wonca). Atualmente é mantido pelo Comitê Internacional de Classificação de Wonca (WICC). É uma ferramenta epidemiológica usada para classificar dados sobre três elementos de cuidados da saúde: motivos de atendimento, diagnóstico ou problema e processo de atendimento. É baseado nos sistemas do corpo humano, problemas psicológicos e problemas sociais. Foi inserido a partir do *Metathesaurus* 1998AA e sua última atualização e tradução para o *Metathesaurus* foi em 1999. Em 2000 houve uma revisão eletronicamente.

- ***LOINC Linguistic Variant - Portuguese, Brazil* (Nomes e Códigos de Identificadores de Observação Lógica - Português, Brasil**[5]

  A terminologia foi criada e é mantida pelo *Regenstrief Institute, Inc.*, uma organização de estudo médico sem fins lucrativos vinculados à *Indiana University*. Ele pretende facilitar o intercâmbio e o agrupamento de resultados clínicos ou laboratoriais para atendimento clínico, gestão de resultados e pesquisa, fornecendo um conjunto de códigos e nomes universais. Os códigos LOINCs permitem que os usuários mesclem resultados clínicos de diversas fontes em uma base de dados para atendimento ao paciente, pesquisa clínica ou gestão. Ele fornece nomes formais e códigos padronizados para observações laboratoriais e outras observações clínicas. Os dados contêm: terminologia laboratorial, sinais vitais, ingestão / saída de hemodinâmica, eletrocardiograma, ultrassom obstétrico, eco cardíaco, urológico, imagem, procedimentos gastroendoscópicos, gerenciamento de ventilador pulmonar e instrumentos de pesquisa selecionados. A representação *Metathesaurus* compreende abreviações expandidas, bem como dados de peças e a hierarquia multiaxial. É atualizado em fevereiro e agosto de todos os anos. A cada versão do *Metathesaurus* ele é atualizado.

[4] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICPCPOR/index.html`
[5] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/LNC-PT-BR/index.html`

- ***Medical Dictionary for Regulatory Activities (Dicionário Médico para Atividades Regulatórias) MedDRA***[6]

  Foi desenvolvido pela *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (Conselho Internacional para Harmonização de Requisitos Técnicos para Produtos Farmacêuticos para Uso Humano) (ICH), pertence à Federação Internacional de Fabricantes e Associações Farmacêuticas (IFPMA). A Organização de Serviços de Manutenção e Suporte (MSSO) serve como repositório, mantenedor e distribuidor do MedDRA, assim como a fonte das informações mais atualizadas e sua aplicação na indústria biofarmacêutica e nos reguladores. Este é designado para o uso específico do compartilhamento de informações regulatórias e dados de segurança clínica para produtos médicos humanos no mundo. O MedDRA é uma terminologia médica internacional com destaque na entrada, recuperação, análise e exibição de dados. Aplica-se a todas as fases do desenvolvimento de medicamentos, menos: a toxicologia animal, e aos efeitos na saúde e mau funcionamento dos dispositivos. Um anexo inclui descrições conceituais que descrevem como um conceito médico é interpretado, usado e classificado. As informações de conceito pretende auxiliar o uso consistente e preciso do MedDRA na codificação, recuperação e análise. Sua representação pelo *Metathesaurus* inclui todos os dados dos arquivos ".asc" no diretório MedAscii da versão MedDRA. É incluído também certas expansões de abreviações em termos, assim como americanizações para grafias britânicas. É atualizado em março e setembro. Cada versão do *Metathesaurus* conta com a versão mais atualizada.

- **MeSH Portuguese**[7]

  Os títulos de assuntos médicos são geridos pela NLM. Este tesauro é usado para indexação, catalogação e IR e documentos biomédicos relacionados à saúde. Permite que sistemas de recuperação de artigos e outras publicações como o PubMed, forneçam pesquisa de dados por assunto. O MeSH contêm três tipos básicos de registros: descritores, qualificadores (subposições) e registros químicos complementares. Grande parte dos descritores são organizados em uma estrutura hierárquica de árvore numerada possibilitando a navegação de tópicos. São quatro as classes do descritor: cabeçalhos principais, tipos de publicação, verificação de *tags* e geografia. Esta última é a exceção possuindo uma estrutura hierárquica. Seus descritores são organizados em estrutura alfabética e hierárquica por categorias como "Anatomia" e "Transtornos Mentais", possuindo outras entradas para os descritores. Por exemplo, "Vitamina C" é um termo de entrada para o descritor "Ácido Ascórbico". Este tesauro é específico para IR de artigos científicos. As subposições são utilizadas com os descritores a fim de agrupar citações

[6] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MDRBPO/index.html`
[7] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHPOR/index.html`

para aspectos específicos de um assunto. Exemplo: efeitos no fígado / drogas. Descritores e qualificadores são atualizados todos os anos, contudo, os registros de conceito suplementares são atualizados continuamente.

PubMed[8] é um banco de dados gratuito que fornece acesso a mais de 35 milhões de citações bibliográficas de literatura biomédica, incluindo resumos e *links* para artigos de revistas em ciências da vida e saúde. Foi desenvolvido e é mantido pelo *National Center for Biotechnology Information (NCBI)*, um componente dos Institutos Nacionais de Saúde (NIH) dos Estados Unidos. O PubMed abrange áreas como medicina, enfermagem, odontologia, medicina veterinária, ciências pré-clínicas, bem como tópicos relacionados, como química, biologia, genética e muito mais. Ele permite que pesquisadores, profissionais de saúde e outros usuários encontrem facilmente literatura científica relevante para sua pesquisa, prática clínica e tomada de decisões.

- ***Systematized Nomenclature of Medicine-Clinical Terms* (Nomenclatura Sistematizada de Medicina - Termos Clínicos) (SNOMED-CT)**[9]

  Atualmente é propriedade do *International Health Terminology Standards Development Organisation (IHTSDO)* e gerido pela Dinamarca. Sua pretensão é ser a maior terminologia clínica multilíngue do mundo, cujo foco principal é em prontuário eletrônico do paciente. O SNOMED-CT Web[10] pode ser usado diretamente nos idiomas dos países membros, até o momento o Brasil não faz parte, contudo, há o português de Portugal. Todos os termos estão interligados com suas respectivas relações associativas, formando uma rede de significados conhecido como Rede Semântica.

- ***WHOART Portuguese* (Terminologia de Reações Adversas a Medicamentos da OMS (WHO-ART))**[11]

  A Terminologia de Reações Adversas a Medicamentos da OMS foi desenvolvida e é mantida pelo Centro de Monitoramento de Uppsala, o centro de Monitoramento Internacional de Medicamentos. O WHO-ART é usado para codificar informações clínicas relacionadas a RAM. É uma terminologia hierárquica de quatro níveis, iniciando nas classes de nível de sistema corporal/órgão. Essas classes consistem em termos de agrupamento amplo, sendo os termos preferenciais mais específicos. A OMS também possui termos comumente usados, chamados de termos incluídos, que se assemelham como termos de entrada para os termos preferidos. É atualizada a cada três meses, porém, sua última atualização no *Metathesaurus* foi em 1999AA. Esse é o tesauro o mais específico para esse trabalho de dissertação, contudo,

[8] `https://pubmed.ncbi.nlm.nih.gov/`
[9] `https://www.snomed.org/?lang=pt`
[10] `https://browser.ihtsdotools.org/`
[11] `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/WHOPOR/index.html`

a sua desatualização junto ao *Metathesaurus* não ajuda com mapeamento dos conceitos.

Nem todas as áreas de conhecimento tem seu tesauro, é muito custoso manter um destes. No Brasil, na área de saúde os vocabulários mais utilizados são o DeCS, o MeSH e o EMTREE, este último só é acessível pelo portal da Capes e tem mais sinônimo que o MeSH. O EMTREE possui mais de 25 milhões de registros, oriundos de mais de 7.000 revistas científicas de mais de 70 países indexadas, dados de 2020 [Dávila Rodríguez, 2023].

Apêndice B Processo de Mapeamento do MetaMap

O Apêndice B apresenta com melhor detalhamento as etapas do processo de mapeamento do MetaMap conforme [Aronson, 1996].

1. **pré-processamento** - nesta primeira etapa ocorrem os seguintes processos: um determinado texto "Text" é considerado como entrada, o processamento do MetaMap busca as marcações (*tags*) sintáticas dos termos do texto e como resultado é gerado uma lista de palavras com suas respectivas marcações sintáticas "TagList". A Figura B.1 apresenta na segunda linha o texto de entrada passado para o mapeamento. Na terceira linha da figura, cada termo é separado dentro da "TagList" com suas respectivas funções sintáticas, os sinais e pontuações também são separados, perceba que o segundo termo "-" e o último, estão marcados respectivamente como hífen e ponto.

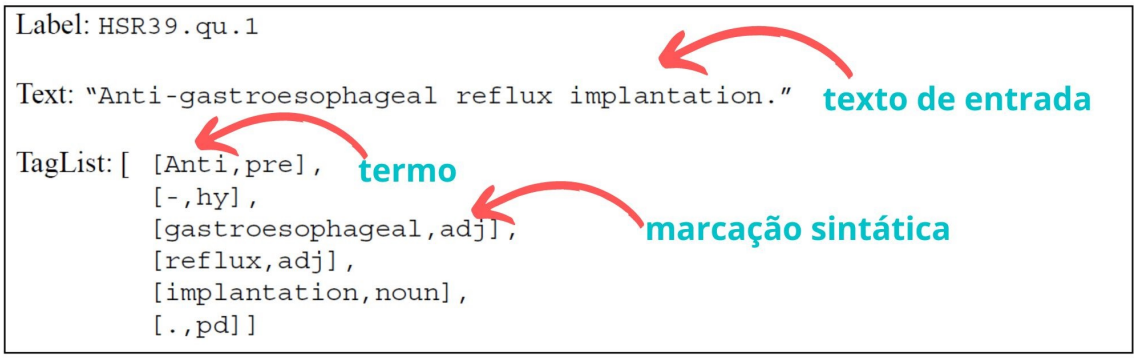


Figura B.1: Texto, Termo e Marcação Sintática (adaptado) [Aronson, 1996].

Em seguida o analisador SPECIALIST gera uma análise sintática mínima possível conforme a Figura B.2 que na sequência, é aumentada com *tokens* conforme a Figura B.3.

Por fim, o processamento preliminar consiste na coleta de informações da análise sintática aumentada com tokens, encontrando a correspondência entre a análise sintática e as palavras do texto e limpando os componentes sintáticos sem relevância, conforme a Figura B.4.

Mais especificamente, o MetaMap executa os seguintes predicados:

- adicionar tokens à frase. Neste é realizada a análise sintática onde geralmente contém itens com várias palavras, por exemplo, *gastroesophageal reflux*. A tratativa que constitui

```
[  not_in_lex([inputmatch([Anti])]),
   punc([inputmatch([-])]),
   mod([lexmatch([gastroesophageal reflux]),inputmatch([gastroesoph-
     ageal,reflux])]),
   head([lexmatch([implantation]),inputmatch([implantation])]),
   punc([inputmatch([.])])]
```

Figura B.2: Análise Sintática de Compromisso Mínimo [Aronson, 1996].

```
[  not_in_lex([inputmatch([Anti]),tokens([anti])]),
   punc([inputmatch([-]),tokens([])]),
   mod([lexmatch([gastroesophageal reflux]),inputmatch([gastroe-
     sophageal,reflux]),tokens([gastroesophageal,reflux])]),
   head([lexmatch([implantation]),inputmatch([implanta-
     tion]),tokens([implantation])]),
   punc([inputmatch([.]),tokens([])])]
```

Figura B.3: Análise Sintática Aumentada com *tokens* [Aronson, 1996].

uma "palavra" defende a tokenização de itens com várias palavras. Os *tokens* para *gastroesophageal reflux* são simplesmente *gastroesophageal and reflux*, conforme a Figura B.3.

- analisar a informação da palavra da frase. Baseado nos *tokens* calculados anteriormente e em preparação adicional para variantes de computação e localização de candidatos no *Metathesaurus*, este segundo predicado, determina a correspondência entre os componentes sintáticos e os *tokens* da frase. Isso é mostrado na Figura B.5. O primeiro elemento da sintaxe, not_in_lex(anti) corresponde ao primeiro *token*, [1,1], mapeado por extensão de palavra começando e terminando na palavra 1. O segundo elemento da sintaxe, punc(-), não corresponde a nenhum *token*, mapeado por [0,-1]. O terceiro elemento corresponde aos *tokens* 2 e 3, [2,3]. O quarto corresponde à palavra quatro, [4,4]. O último corresponde a nada, [0,-1]. A concatenação dessas correspondências produz o mapeamento da sintaxe para os *tokens*: [[1,1], [0,-1], [2,3], [4,4], [0,-1]]. Além do mapeamento, este segundo predicado realiza os seguintes processos: calcula a lista de todos os *tokens*, [anti, gastroesophageal, reflux, implantation], e uma versão em minúscula; lista os *tokens* principais, [implantation], e uma versão em minúsculas (a mesma coisa); e outra análise como acima calculada filtrando os *tokens* de componentes sintáticos não significativos, isto é, aqueles com as marcações prep, det, aux, modal, compl, punc, num, conj e pron. Para este exemplo, a versão filtrada é idêntica à versão não filtrada. As informações reais da palavra da frase para o exemplo são mostradas na Figura B.6.

2. **geração de variantes** - são definidos os geradores de variantes, estes são as sequências de

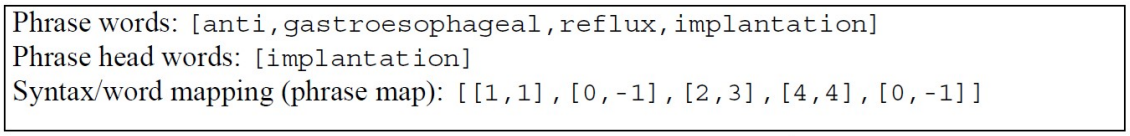


Figura B.4: Informações da Palavra da Frase [Aronson, 1996].

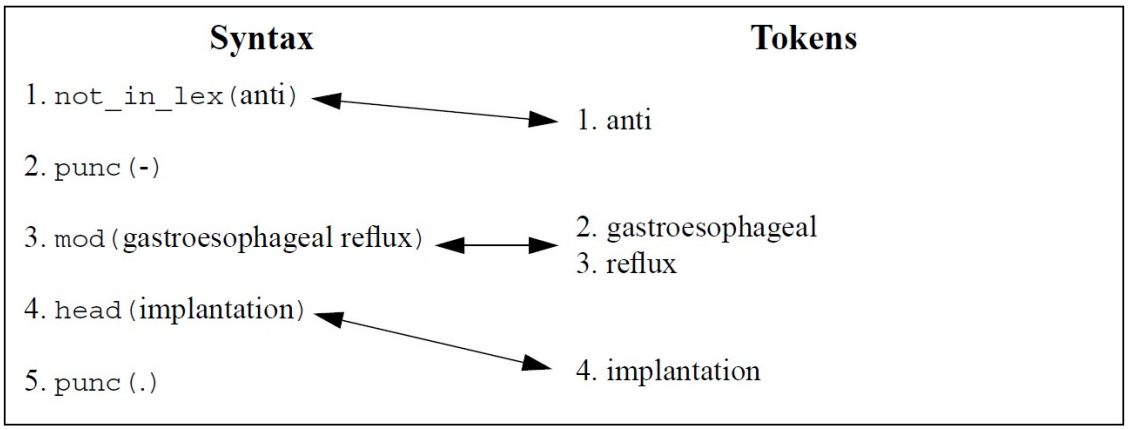


Figura B.5: Sintaxe / Correspondência *token* [Aronson, 1996].

palavras ocorridas no Léxico ESPECIALISTA da UMLS conforme a entrada do texto. Por conseguinte, para cada gerador, são produzidas todas as variantes gramaticas possíveis, tais como abreviações, acrônimos, flexões, sinônimos, dentre outras. As variantes são geradas por etapas, a Figura B.7. exibe a parte final com todas as possíveis variantes com: as partes da fala, variantes e suas classes gramaticais, distância variante e história.

3. **busca de candidatos** - na terceira etapa, são recuperadas todas as *strings Metathesaurus* da UMLS com as principais informações, tais como: ID, grupo semântico e nome do conceito. Estas iniciam com umas das variações da etapa anterior. Algumas das 239 *strings* do exemplo são mostradas na Figura B.8. Cada exemplo equivale aos *tokens* da *string*, na própria *string* e no conceito da *string*.

4. **avaliação de candidatos**, neste, são calculados o *score* de cada candidato achado. O *score* é um valor numérico que define se o candidato mapeia o texto original da melhor forma esperada.

   O *score* utiliza os seguintes parâmetros conforme Aronson [2001]:

   - *Centrality* (CE) - valor que indica se o candidato contém escrito o núcleo da frase original.
   - *Variation* (VA) - valor que indica a diferença entre as variantes do candidato e as palavras da frase original.
   - *Coverage* (CO) - valor que indica a cobertura do candidato na frase original.
   - *Cohesiveness* (CH) - valor que indica o quão suave o candidato cobre a frase original.

```
pwi(  wdl([anti,gastroesophageal,reflux,implantation],
          [anti,gastroesophageal,reflux,implantation]),
      wdl([implantation],[implantation])
      [[1,1],[0,-1],[2,3],[4,4],[0,-1]])
:
pwi(  wdl([anti,gastroesophageal,reflux,implantation],
          [anti,gastroesophageal,reflux,implantation]),
      wdl([implantation],[implantation]),
      [[1,1],[0,-1],[2,3],[4,4],[0,-1]])
```

Figura B.6: Informação da Palavra da Frase [Aronson, 1996].

```
v(anti,[],0,[],anti,4)
v(gastroesophageal reflux,[noun],0,[],gastroesophageal reflux,3)
v(gastrooesophageal reflux,[],0,"p",gastrooesophageal reflux,3)
v(gastroesophageal,[adj],0,[],gastroesophageal,3)
v(gastro-oesophageal,[],0,"p",gastro esophageal,3)
v(reflux,[verb,noun,adj],0,[],reflux,2)
v(refluxed,[],1,"i",reflux,2)
v(refluxes,[],1,"i",reflux,2)
v(refluxing,[],1,"i",reflux,2)
v(implantation,[noun],0,[],implantation,1)
v(implant,[verb],3,"d",implant,1)
v(implantable,[adj],6,"dd",implantable,1)
v(implantations,[],1,"i",implantation,1)
v(implanted,[],4,"id",implant,1)
v(implanting,[],4,"id",implant,1)
v(implants,[],4,"id",implant,1)
```

Figura B.7: Informação da Palavra da Frase [Aronson, 1996].

- *Involvement* (IV) - valor alternativo a cobertura e coesão que indica quanto a frase está envolvida na correspondência entre a frase e o candidato. Este valor existe em casos onde a ordem das palavras de entrada não importa, ocorrendo isso, esse parâmetro fica no lugar do CO e CH.

A fórmula do *score* se dá de duas formas, a primeira é o processamento normal, a outra ocorre quando a ordem das palavras de entrada não importa no processamento.

- 1000 * (CE + VA + 2 * CO + 2 * CH)/6
- 1000 * (CE + VA + 4 * IV)/6

Cada um dos candidatos ao *Metathesaurus* é avaliado relacionado ao texto da sentença. A Figura B.9 exibe os resultados. Cada resultado equivale na negação da votação para o candidato, o próprio candidato, o conceito correspondente do candidato, os *tokens* do candidato, os tipos semânticos do conceito, o mapeamento da frase para o candidato, caso o mapeamento envolva o candidato principal da frase, e se há ou não um excesso de correspondência [Aronson, 1996].

```
csc([anti,3,4,diol,1,2,oxide,benz,a,anthracene],anti-3,4-diol 1,2-
     oxide benz(a)anthracene,anti-3,4-diol 1,2-oxide
     benz(a)anthracene)
csc([anti,5,methylchrysene,1,2,dihydrodiol,3,4,epoxide],anti-5-
     methylchrysene-1,2-dihydrodiol-3,4-epoxide,anti-5-methylchry-
     sene-1,2-dihydrodiol-3,4-epoxide)
csc([anti,a,carbohydrate,test],Anti-A-carbohydrate test,Anti-A-car-
     bohydrate test)
csc([anti,a,cho,test],Anti-A-CHO test,Anti-A-carbohydrate test)
csc([anti,abortion,group],Anti-Abortion Group,Anti-Abortion Groups)
csc([gastroesophageal,reflux],GASTROESOPHAGEAL REFLUX,Gastroesoph-
     ageal Reflux)
csc([gastroesophageal,reflux],Gastroesophageal Reflux,Gastroesoph-
     ageal Reflux)
csc([gastro,esophageal,laceration,syndrome],Gastro-esophageal lac-
     eration syndrome,Mallory-Weiss Syndrome)
csc([gastro,esophageal,reflux],Gastro Esophageal Reflux,Gastroe-
     sophageal Reflux)
csc([reflux],Reflux, NOS,Reflux, NOS)
csc([reflux],reflux,Reflux, NOS)
csc([implant],Implant,Implant <1>)
csc([implant],Implant, NOS,Implantation, NOS <1>)
csc([implant],implant,Implants, Artificial)
csc([implantation],Implantation,Implantation, NOS <1>)
csc([implantation],Implantation,Ovum Implantation)
```

Figura B.8: Candidatos do *Metathesaurus* [Aronson, 1996].

5. **construção de mapeamentos** - finalizando, esta etapa escolhe a melhor combinação de candidatos que formam o mapeamento de maneira mais completa possível. O resultado pode ter diversas saídas, contudo o padrão é exibir o com maior *score*, em caso de mais ocorrências de mapeamentos com este valor, este valor aparecerá nas combinações. Os melhores mapeamentos para o exemplo, ambos com pontuação final de 840, são apresentados na Figura B.10.

```
ev(-812,Implantation,Implantation, NOS <1>,[implantation],[Thera-
     peutic or Preventive Procedure], [[[4,4],[1,1],0]],yes,no)
ev(-812,Implantation,Ovum Implantation,[implantation],[Organism
     Function], [[[4,4],[1,1],0]],yes,no)
ev(-741,Implant,Implant <1>,[implant],[Medical
     Device],[[[4,4],[1,1],3]],yes,no)
ev(-741,implant,Implants, Artificial,[implant],[Medical
     Device],[[[4,4],[1,1],3]],yes,no)
ev(-729,Implanted,Implanted,[implanted],[Functional Con-
     cept],[[[4,4],[1,1],4]],yes,no)
ev(-729,Implants,Implants <1>,[implants],[Biomedical or Dental
     Material], [[[4,4],[1,1],4]],yes,no)
ev(-729,Implants,Implants <2>,[implants],[Medical
     Device],[[[4,4],[1,1],4]],yes,no)
ev(-694,GASTRO-OESOPHAGEAL REFLUX,Gastroesophageal Reflux,[gas-
     tro,oesophageal,reflux],[Disease or Syndrome],
     [[[2,2],[1,2],0],[[3,3],[3,3],0]],no,no)
ev(-694,GASTROESOPHAGEAL REFLUX,Gastroesophageal Reflux,[gastroe-
     sophageal,reflux],[Disease or Syndrome],
     [[[2,3],[1,2],0]],no,no)
ev(-645,Reflux, NOS,Reflux, NOS,[reflux],[Finding,Sign or Symptom],
     [[[3,3],[1,1],0]],no,no)
```

Figura B.9: Avaliação de candidatos [Aronson, 1996].

```
map(-840,[ev(-694,GASTRO-OESOPHAGEAL REFLUX,Gastroesophageal
              Reflux,[gastro,oesophageal,reflux],[Disease or
              Syndrome],[[[2,2],[1,2],0],[[3,3],[3,3],0]],no,no),
        ev(-812,Implantation,Implantation, NOS <1>,[implantation],
              [Therapeutic or Preventive Procedure],
              [[[4,4],[1,1],0]],yes,no)])
map(-840,[ev(-694,GASTRO-OESOPHAGEAL REFLUX,Gastroesophageal
              Reflux,[gastro,oesophageal,reflux],[Disease or
              Syndrome],[[[2,2],[1,2],0],[[3,3],[3,3],0]],no,no),
        ev(-812,Implantation,Ovum Implantation,[implantation],
              [Organism Function],[[[4,4],[1,1],0]],yes,no)])
```

Figura B.10: Os Melhores Mapeamentos para “Anti-gastroesophageal reflux implantation”. [Aronson, 1996].