UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Mineração de Texto aplicada às análises de intervenção de Políticas Públicas de Saúde: o caso da epidemia de sífilis no Brasil

Marcella Andrade da Rocha

Orientador: Prof. Dr. Ricardo Alexsandro de Medeiros Valentim

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Computação da UFRN (área de concentração: Engenharia de Computação) como parte dos requisitos para obtenção do título de Doutor em Ciências.

Natal, RN, julho de 2022

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI

Catalogação da publicação na fonte. UFRN - Biblioteca Central Zila Mamede

Rocha, Marcella Andrade da.
Mineração de Texto aplicada às análises de intervenção de Políticas Públicas de Saúde: o caso da epidemia de sífilis no Brasil/ Marcella Andrade da Rocha. - 2022
84 f.: il.

Tese (doutorado) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Natal, RN, 2022.
Orientador: Prof. Dr. Ricardo Alexsandro de Medeiros Valentim.

1. Sífilis - Tese. 2. Mineração de textos - Tese. 3. Apoiadores - Tese. 4. Projeto "Sífilis Não!- Tese. 5. N-gramas - Tese. I. Valentim, Ricardo Alexsandro de Medeiros. II. Título.

RN/UF/BCZM CDU 004.4:616.972
Elaborado por Ana Cristina Cavalcanti Tinoco - CRB-15/262

Às pessoas que eu amo.

Agradecimentos

Ao meu orientador, professor Ricardo Alexsandro de Medeiros Valentim, sou grata por toda a orientação e ajuda proporcionada desde que ingressei no Mestrado na UFRN em 2017, uma grande pessoa a quem devo todo meu reconhecimento e consideração.

À minha filha, Sophia Andrade Castelo Branco Araújo, principalmente por toda a paciência que teve comigo, mesmo sendo uma criança, ela sempre entendeu que eu precisava de tempo para me dedicar ao trabalho e me ajudou muito nisso.

Aos meus pais, Marineusa Damasceno Andrade e Ronaldo da Rocha e meu irmão, Madson Rangel Andrade Rocha, mesmo de longe sempre estão ao meu lado me apoiando e me motivando a continuar me esforçando.

À minha amiga, Sthefani Wanzeller da Silva e a meu amigo, Marquiony Marques dos Santos, além de serem grandes e verdadeiros amigos para a vida, me ajudaram em tudo que eu precisava, mesmo estando distantes e nos vendo poucas vezes ao ano.

Ao meu futuro marido, Matheus Cavalcante de Oliveira, mesmo entrando depois em minha vida, já na fase final do meu doutorado, me apoiou em toda a correria, surtos e crises de choro que tive. Se suportou essa minha fase sei que vai suportar várias outras na nossa vida juntos.

Aos demais colegas do Laboratório de Inovação Tecnológica em Saúde, LAIS/UFRN, pelos ensinamentos, trabalho em equipe e amizade.

À toda minha família pela motivação durante esta jornada.

Resumo

A sífilis é uma doença infectocontagiosa crônica que possui cura e é conhecida há séculos, causada pela bactéria *Treponema Pallidum*. Mesmo com fácil tratamento e diagnóstico, a sífilis continua como um sério problema de saúde pública em grande parte do mundo. Apenas no Brasil, os dados do Boletim Epidemiológico de 2017 revelaram uma elevação no número de casos de sífilis em gestantes, adquirida e congênita. Considerando a magnitude do problema a ser enfrentado, surge em 2018 no Brasil a "Pesquisa aplicada para integração inteligente orientada ao fortalecimento das redes de atenção para resposta rápida à sífilis", o projeto "Sífilis Não!", que tem como objetivo reduzir os casos de sífilis adquirida e em gestantes e eliminar a congênita no país. Esse projeto possui várias estratégias para combate à sífilis e, entre elas, a criação de um grupo de apoiadores de pesquisa e intervenção que atuaram em municípios prioritários e produziram milhares de relatos de textos e os adicionaram em uma plataforma. O objetivo do trabalho é o desenvolvimento de métodos computacionais utilizando mineração de textos que ajudam a compreender o impacto da sífilis no território utilizando as produções textuais da "plataforma LUES" dos apoiadores do projeto "Sífilis Não!". Foi utilizada a base de dados extraída da "Plataforma LUES" com 4.874 documentos em arquivo de texto e 3.071 documentos em planilhas eletrônicas entre os anos de 2018 e 2020. Seguiu-se o pré-processamento desses textos, com escolha para análise dos textos referentes aos relatórios dos apoiadores. Por fim, para essa análise, foi realizada extração dos N-gramas (N=2,3,4) utilizando a combinação da métrica TF-IDF com o algoritmo BoW para verificar a importância e a frequência dos termos, para o agrupamento dos textos que depois foram analisados utilizando técnicas de análise de conteúdo e interpretação dos termos. Assim, foram testadas associações dos dados extraídos dos relatórios com indicadores de sífilis e o impacto da epidemia no território. A mineração de textos, ao ser utilizada em conjunto ao tradicional método de análise de conteúdo, é capaz de atender objetos de pesquisa de saúde pública. O método computacional extraiu ações de intervenção dos apoiadores, como também subsidiou inferências sobre como as estratégias do projeto "Sífilis Não!" incidiram na redução dos casos de sífilis congênita no território.

Palavras-chave: Sífilis, Mineração de textos, Apoiadores, Projeto "Sífilis Não!", N-gramas, Plataforma LUES.

Abstract

Syphilis is a chronic, curable infectious disease that has been known for centuries, caused by the *Treponema Pallidum* bacterium. Even with easy treatment and diagnosis, syphilis remains a serious public health problem in much of the world. Only in Brazil, data from the 2017 Epidemiological Report revealed an increase in the number of cases of syphilis in pregnant women, acquired and congenital. Considering the magnitude of the problem to be faced, in 2018 the "Applied research for intelligent integration aimed at strengthening care networks for a rapid response to syphilis" appears in Brazil, the "Syphilis No!" project, which aims to reduce cases of acquired syphilis and in pregnant women and eliminate congenital syphilis in the country. This project has several strategies to combat syphilis and, among them, the creation of a group of field researchers who worked in priority municipalities and produced thousands of text reports and added them to a platform. The objective of the thesis is the development of computational methods using text mining that help to understand the impact of syphilis in the territory using the textual productions of the LUES platform of the field researchers of the "Syphilis No!" project. The database extracted from the "LUES Platform" with 4,874 documents in text file and 3,071 documents in spreadsheets between the years 2018 and 2020 was used. This was followed by the pre-processing of these texts, with the choice for analysis of the texts referring to the field researchers' reports. Finally, for this analysis, extraction of the N-grams (N=2,3,4) was performed using the combination of the TF-IDF metric with the BoW algorithm to verify the importance and frequency of the terms, for the grouping of the texts that were then analyzed using content analysis techniques and interpretation of the terms, thus, associations of the data extracted from the reports with indicators of syphilis and its epidemic impact on the territory were tested. Text mining, when used in conjunction with the traditional content analysis method, is able to meet public health research objects. The computational method extracted intervention actions from the field researchers, as well as subsidized inferences about how the strategies of the "Syphilis No!" project impacted the reduction of congenital syphilis cases in the territory.

Keywords: Syphilis, Text mining, field researchers, "Syphilis No!" Project, N-grams, LUES Platform.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Abreviaturas

AAPC	Average Annual Percent Change
AM	Aprendizado de Máquina
AVASUS	Ambiente Virtual de Aprendizagem do SUS
BNB	Bernoulli Naive Bayes
BoW	Bag of Word
CDW	Clinical Data Warehouse
CSV	Comma Separated Values
EHR	Electronic Health Records
GCP	Gaussian Process Classification
GES	Garantias Explícitas da Saúde
HIV	Human Immunodeficiency Virus
HSSF	Health Surveillance Software Framework
HTML	HyperText Markup Language
IDC	International Data Corporation
IoT	Internet of Things
IST	Infecções Sexualmente Transmissíveis
KNN	K-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
LSA	Latent Semantic Analysis
MS	Ministério da Saúde
NB	Naive Bayes

NLP	Natural Language Processing
PLN	Processamento de Linguagem Natural
RME	Registro Médico Eletrônico
SIGTE	Sistema de Gerenciamento do Tempo de Espera
SUS	Sistema Único de Saúde
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
UIMA	Unstructured Information Management Architecture

Capítulo 1

Introdução

Este capítulo contempla uma contextualização sucinta do tema pesquisado, problematização, questões de pesquisa, hipóteses testadas e objetivos. Ademais, as contribuições relevantes que essa pesquisa proporcionou e como essa tese está estruturada.

1.1 Contextualização

A sífilis é uma doença sexualmente transmissível que surgiu na Europa no final do século XV e é causada pela bactéria *Treponema Pallidum*. Ainda que o seu agente etiológico seja conhecido, essa doença tenha cura e o seu tratamento seja realizado com a injeção de penicilina (Crosby Jr 1969), (Rothschild 2005), a sífilis continua como um sério problema global de saúde pública, principalmente em países com recursos limitados e com baixo investimento na atenção primária à saúde, sendo a segunda causa de eventos adversos na gestação, que pode levar à morte neonatal. Portanto, trata-se de uma doença com tratamento relativamente simples, mas que desafia a efetividade de sistemas públicos de saúde (Kojima & Klausner 2018), (Korenromp et al. 2019), (Forrai 2011), (Sarbu et al. 2014).

A partir do boletim epidemiológico da sífilis no Brasil (Ministério da Saúde 2021), nota-se o aumento na taxa de incidência de sífilis congênita nos últimos dez anos até 2018 e uma redução a partir do ano 2019. Um total de 260.596 casos foram notificados no período de 1998 a 2021. Em 2020, a taxa de incidência no Brasil foi de 7,7 casos de sífilis congênita a cada 100 mil nascidos vivos. Embora o aumento de casos da sífilis congênita seja visto com cautela, em decorrência da notificação compulsória ter ocorrido a partir do ano de 2010 até 2018, as estimativas da epidemia continuam crescendo exponencialmente. Do total de casos notificados no período de notificações, 9.819 (44,5%) deles ocorreram na região sudeste (Ministério da Saúde 2021). As taxas de sífilis congênita superaram a mundial, com 880 casos para cada 100.000 nascidos vivos em 2017, com a Variação Percentual Anual Média (Average Annual Percent Change - AAPC) de 60,38% (Marques dos Santos et al. 2020).

Como consequência do crescimento da sífilis congênita no Brasil, em 2018, o Ministério da Saúde (MS) instituiu junto aos órgãos de governança do Sistema Único de Saúde (SUS), um projeto nacional para resposta integrada à sífilis nas redes de atenção, cuja cooperação técnica foi estabelecida pela Universidade Federal do Rio Grande do

Norte, e ficou conhecido como projeto "Sífilis Não!". Esse projeto é delineado pela implementação de ações universais e de ações específicas. Estas últimas foram aplicadas aos 100 municípios que, na ocasião, apresentavam os piores indicadores de sífilis congênita no Brasil, e que foram considerados como os municípios prioritários (de Andrade et al. 2020).

O projeto "Sífilis Não!" teve como intuito atuar como ferramenta de indução de política pública em saúde para contribuir com a redução da sífilis congênita e sífilis adquirida em gestantes e, nesse contexto, foi definida a estratégia de criação de um grupo de apoiadores que atuaram por meio de ações diretas nos municípios prioritários com intervenções relacionadas a quatro eixos temáticos no projeto: Vigilância em saúde; Rede de atenção à saúde; Gestão e governança e Educação/Comunicação. Esses apoiadores trabalharam junto aos gestores municipais para traçar as melhores estratégias de combate à sífilis. Todo o trabalho foi acompanhado por supervisores a partir de um sistema de gestão de pesquisa e intervenção chamado de "plataforma LUES" (LAIS/UFRN 2022). A partir dessa plataforma, os apoiadores produziram e inseriram milhares de relatos de textos de suas experiências no território (BRASIL 2018).

A produção textual referente a atuação dos apoiadores no território foi toda inserida na plataforma digital "LUES" e seu conteúdo é de fundamental importância para subsidiar análises sobre o impacto destas ações e suas relações como instrumento de indução da política de saúde para resposta à sífilis. Diante do exposto, este trabalho realizou o desenvolvimento e análise do conteúdo textual desenvolvido pelos apoiadores, por intermédio de um método de mineração de dados, que utilizou algoritmos para extração de informações nos textos produzidos, como ferramenta para auxiliar na compreensão do papel do apoiador no território. Em seguida, é reiterada a problematização, questões de pesquisa, hipóteses e objetivos da pesquisa.

1.2 Problematização

A utilização de uma grande quantidade de dados tem crescido em todas as áreas da ciência. Até 2020 foi estimado a produção de 40 trilhões de gigabytes no mundo, ou seja, 2,2 milhões de terabytes de dados todos os dias e, atualmente, no ano de 2022, é produzido cerca de 2,5 milhões de terabytes por dia (Gartner 2022), (IDC 2022). No cenário da saúde a quantidade da dados gerados é progressivamente maior a cada dia, um estudo realizado pela International Data Corporation (IDC) projetou no setor da saúde um total de 25 mil petabytes de dados, em 2020.

O panorama da saúde em comparação com as outras áreas da ciência, possui muito mais ferramentas capazes de gerar e absorver dados (Piai & Claps 2013). No entanto, é necessário ter os mecanismos adequados para transformar esses grandes volumes de dados em informação relevante (Piai & Claps 2013). Esse volume de dados recebido pelo setor da saúde é proveniente de bancos de dados públicos e privados, prontuários eletrônicos, dispositivos IoT, aplicativos, arquivos e redes sociais e, esses dados muitas vezes estão em forma de textos digitais (Alshaikh et al. 2014), (Gianfrancesco et al. 2018).

Um dos desafios de trabalhar com grandes volumes de dados em forma de textos é tentar interpretar, de forma objetiva, estruturas que ajudem a compreender o impacto do

conteúdo a partir dos relatos escritos nas plataformas digitais, principalmente no domínio da saúde como, por exemplo, prontuários eletrônicos de pacientes, relatórios técnicos produzidos por profissionais de saúde e fóruns dos cursos disponíveis no Ambiente Virtual de Aprendizagem do SUS (AVASUS). Analisar esses tipos de dados a partir de métodos tradicionais manuais toma muito tempo. Existem estratégias de análise de texto que são tradicionalmente usadas na área de saúde e são feitas com a utilização de software para busca de palavras chaves (Souza, Wall, Thuler, Lowen & Peres 2018).

Os softwares atualmente disponíveis são insuficientes para acompanhamento das ações estratégicas de intervenção, principalmente por produzirem informações limitadas que não levam em consideração a formação textual (Amaral-Rosa 2019). Esses softwares limitam a capacidade do investigador de conduzir com segurança os resultados produzidos, além de existir diversas dificuldades na interpretação do processamento de textos desses softwares (Amaral-Rosa 2019). Com o crescimento acelerado da Inteligência Artificial para diversos campos de atuação da saúde, tem surgido a utilização de algoritmos de aprendizagem para análise e modelagem de Big Data (Kumar et al. 2021), (Kar & Dwivedi 2020). O método de mineração de texto pode oferecer consistência na interpretação de estratégias de monitoramento textual, além de oferecer um tratamento fiel no fornecimento de palavras nas formas de n-gramas (Li et al. 2019).

A incorporação da mineração de textos na aplicação do método de análise de conteúdo, em pesquisas que envolvem grandes volumes de dados, pode subsidiar estudos avaliativos (Lee et al. 2019), e complementar indicadores que medem processo de trabalho e desempenho, auxiliando estratégias de saúde pública para novas intervenções frente ao desafio da sífilis congênita no país.

1.3 Questões de Pesquisa

A elaboração dessa tese possui, como principal, a seguinte questão norteadora:

- Como o método de mineração de textos contribui na análise das ações dos apoiadores de pesquisa e intervenção no enfrentamento à sífilis nos municípios prioritários do projeto "Sífilis Não!" no Brasil?

E possui também as seguintes questões secundárias:

- Há conexão entre os objetivos do projeto "Sífilis Não!" e os conteúdos apresentados pelos apoiadores a partir dos métodos computacionais que aplicam algoritmos de mineração de textos?
- Esse método permite identificar relatos de ações de intervenção relacionadas ao enfrentamento à sífilis congênita?
- Esse método permite identificar relatos de ações de intervenção relacionadas ao rastreamento da transmissão vertical?
- Esse método permite identificar relatos de ações de intervenção relacionadas à gestão municipal no enfrentamento à sífilis?

1.4 Hipóteses

Partindo do cenário que o trabalho do apoiador de pesquisa e intervenção demonstra os nexos entre os objetivos do projeto "Sífilis Não!" e à gestão da sífilis no território e que sua atuação incidiu nos indicadores de diminuição de sífilis congênita nos municípios prioritários. A seguinte hipótese teórica nula foi levantada juntamente com sua hipótese alternativa:

H_0: Métodos computacionais que aplicam algoritmos de mineração de textos permitem estabelecer os nexos entre o projeto "Sífilis Não!" e a indução de política pública nos territórios.

H_1: Métodos computacionais que aplicam algoritmos de mineração de textos não permitem estabelecer os nexos entre o projeto "Sífilis Não!" e a indução de política pública nos territórios.

1.5 Objetivo

Desenvolver métodos computacionais que aplicam algoritmos de mineração de textos para analisar as conexões entre os objetivos do projeto "Sífilis Não!" e a indução da política pública de Sífilis no território.

1.6 Objetivos Específicos

O projeto irá envolver os seguintes objetivos pertinentes:

- Analisar as produções não estruturadas dos apoiadores do projeto "Sífilis Não!", por ano e região, utilizando mineração de texto;
- Identificar estruturas textuais da plataforma LUES a partir da mineração de textos;
- Associar termos da sífilis que estão relacionados aos dados extraídos.

1.7 Estrutura da tese

A organização deste trabalho consiste em 6 capítulos. O primeiro possui a introdução envolvendo desde contextualização e problematização até os objetivos, hipóteses e questões de pesquisa. O capítulo 2 conta a trajetória percorrida para chegar ao resultado da tese. O capítulo 3 aborda um breve estado da arte com os trabalhos mais relevantes entre 2010 e 2021. O capítulo 4 trata dos materiais e métodos descrevendo os conjuntos de procedimentos que são utilizados e como foram utilizados no desenvolvimento da pesquisa. O capítulo 5 traz os resultados adquiridos com o método de mineração de textos em todas as aplicações. E por fim, o capítulo 6 apresenta as considerações finais e aponta os direcionamentos para pesquisas futuras.

Capítulo 2

Fundamentação teórica

Este capítulo trata-se do referencial teórico e possui a descrição das técnicas e metodologias existentes que são essenciais ao entendimento desta tese.

2.1 Mineração de textos

A mineração de texto é o processo de extrair o conhecimento que está implícito nos dados textuais (Feldman et al. 2007). O agrupamento de texto, a classificação e a associação são tarefas típicas da mineração de textos e serão descritos com mais detalhes nas próximas subseções. A mineração de texto é um tipo especial de mineração de dados, e outro tipo, como mineração de big data, também será abordada nas subseções seguintes. Portanto, esta seção pretende explorar a visão geral da mineração de texto, antes de mencionar as tarefas de mineração de texto e um dos tipos de mineração de dados (Feldman et al. 2007).

O texto é definido como dados não estruturados que consistem em *strings* que são chamadas de palavras (Salton 1989). Mesmo que a coleção de *strings* pertença ao texto na visão ampla, ela precisa dos significados das *strings* individuais e a combinação delas por meio das regras (gramática) para fazer o texto. O escopo do texto é restrito ao artigo que consiste em parágrafos e está escrito em linguagem natural. Atribui-se que um parágrafo é referido a um grupo organizado de frases e um texto é um conjunto ordenado de parágrafos. O que consiste em palavras escritas em uma linguagem artificial, como código-fonte ou equações matemáticas, são excluídos do escopo do texto.

A mineração de texto é considerada como o tipo especial de mineração de dados, conforme foi mencionado anteriormente, e é necessário explorar a mineração de dados conceitualmente para entendê-la. A mineração de dados refere-se ao processo de obtenção do conhecimento implícito de qualquer tipo de dado na visão geral. No entanto, na mineração de dados tradicional, o tipo de dado que realiza a função de origem são os dados relacionais .

A classificação, a regressão, o agrupamento e a associação são as principais tarefas da mineração de dados. A classificação refere-se ao processo de classificar os dados em suas próprias categorias e a regressão é feita para estimar um valor de saída ou valores de saída para cada dado. Agrupamento (ou *clustering*) é considerado como o processo de segmentação de um grupo de vários dados em vários subgrupos de dados semelhantes,

que são chamados de grupos ou *clusters* . A associação é considerada a tarefa de extrair combinações de dados na forma de *se-então* (Tan et al. 2016), por exemplo, existe um pequeno mercado e ele possui uma base de dados de vendas e é necessário encontrar regras de associação entre os produtos vendidos. Presumindo que o conjunto de produtos encontrados em uma certa regra de associação, é pão, ovo, leite e que a relação encontrada é pão, ovo $\rightarrow$ leite, isto é, os produtos pão e ovo compõe o "se" da relação "se-então" e o leite compõe o "então" (Džeroski 2009).

Na Tabela 2.1 (Salton 1989), são apresentadas as diferenças entre a mineração e a recuperação. A saída da mineração de dados é o conhecimento que é necessário diretamente para a tomada de decisões, enquanto a da recuperação são alguns dados que são relevantes para a consulta dada. Por exemplo, no domínio dos preços das ações, a previsão dos preços futuros das ações é uma tarefa típica de mineração de dados, enquanto a obtenção de alguns preços de ações passados e atuais é a recuperação de informações. Pode-se observar que a certeza nunca existe na mineração de dados, em comparação com a recuperação (Salton 1989). O maior avanço na computação para obter conhecimento a partir dos dados brutos é chamado de síntese e é necessário para realizar as tarefas de mineração de dados (Feldman et al. 2007).

Tabela 2.1: Mineração x Recuperação

	Mineração	**Recuperação**
Exemplo	Valores previstos	Valores atuais ou passados
Saída	Conhecimento	Dados relevantes
Síntese	Requerida	Opcional
Certeza	Probabilidade	Explícita

2.2 Tarefas típicas de mineração de textos

Nesta seção será explorada com mais detalhes as tarefas individuais de mineração de textos já citadas introdutoriamente: a classificação, o agrupamento e a associação.

2.2.1 Classificação

A classificação é definida como o processo de atribuir uma ou algumas categorias entre as predefinidas para cada dado, de acordo com o mostrado na Figura 2.1. A tarefa preliminar na classificação é definir uma lista de categorias predeterminadas para o sistema de classificação e os dados são distribuídos para cada categoria como dados de amostra. Para essa tarefa são considerados dois tipos de abordagens: a baseada em regras, onde as regras são definidas manualmente e cada dado é classificado por essas regras; e a baseada em aprendizado de máquina, onde a capacidade de classificação é construída pelas amostras e cada dado é classificado por ela. A abordagem baseada em regras geralmente é excluída por causa do seu limite, pouca flexibilidade e os requisitos

de conhecimento prévio. Esta subseção pretende descrever a classificação em sua visão funcional, o processo de aplicação de algoritmos de aprendizado de máquina e as técnicas de avaliação.

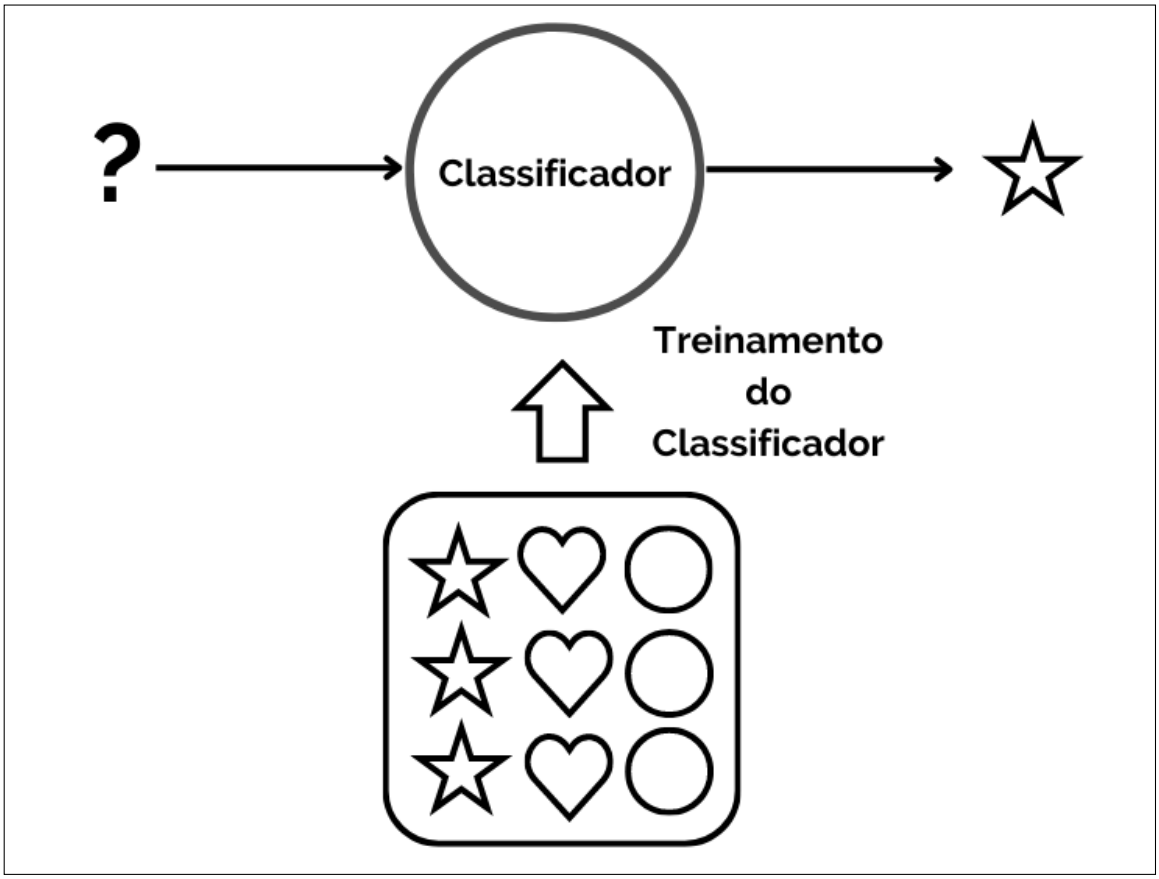

Figura 2.1: Classificação dos dados
Fonte: adaptada de (Aggarwal 2015)

A classificação é uma caixa preta cuja entrada são os dados e a saída, suas categorias. A classificação *Hard* refere-se à classificação em que apenas uma categoria é atribuída a cada dado, enquanto a classificação *Soft* refere-se à classificação em que mais de uma categoria pode ser atribuída a ele (Hart et al. 2000). Na classificação plana, as categorias são predefinidas como uma única lista enquanto na hierárquica, são feitas como uma árvore hierárquica; categorias aninhadas existem em algumas categorias (Hart et al. 2000). A classificação de visualização única é considerada como aquela em que apenas um sistema de classificação é predefinido, seja plano ou hierárquico, enquanto a classificação de visualização múltipla é feita como aquela em que mais de um sistema de classificação é predefinido ao mesmo tempo. Antes de construir o sistema de classificação automática, leva muito tempo para predefinir categorias e coletar amostras de dados, dependendo das áreas de aplicação (Jo 2006).

Considerando as etapas de classificação dos dados por um algoritmo de aprendizado de máquina, as categorias são predefinidas com uma lista ou uma árvore e os dados de

amostra são coletados e alocados para cada categoria. Ao aplicar o algoritmo de aprendizado de máquina aos dados da amostra, implanta-se a classificação que é feita de diversas maneiras: regras simbólicas,equações matemáticas e probabilidades (Mullen & Collier 2004). Ao aplicar a classificação, os dados são separados separados da amostra e são classificados. Os dados são fornecidos posteriormente como alvos de classificação e devem ser diferenciados dos dados de amostra que são rotulados manualmente e fornecidos com antecedência (Mullen & Collier 2004).

Considerando o esquema de avaliação dos resultados da classificação dos dados, um conjunto de teste que consiste em dados rotulados é dividida em dois conjuntos: conjunto de treinamento e conjunto de teste. O conjunto de treinamento é utilizado com dados de amostra para implantar a classificação usando um algoritmo de aprendizado de máquina. Classificamos os dados no conjunto de teste e observamos as diferenças entre seus rótulos verdadeiros e classificados. A precisão como a taxa de itens rotulados consistentemente em relação ao total e a medida F1 são usados como medidas de avaliação (Wiener 1993).

2.2.2 Agrupamento

O agrupamento é definido como o processo de segmentação de um grupo de vários dados em subgrupos de dados semelhantes, conforme mostrado na Figura 2.2. Na tarefa, os dados não rotulados são fornecidos inicialmente, e as medidas de similaridade entre eles devem ser definidas. Um grupo de dados é segmentado, dependendo das semelhanças entre eles, em subgrupos. Os algoritmos de aprendizado supervisionado são aplicados à classificação e à regressão, enquanto os não supervisionados são aplicados no agrupamento.

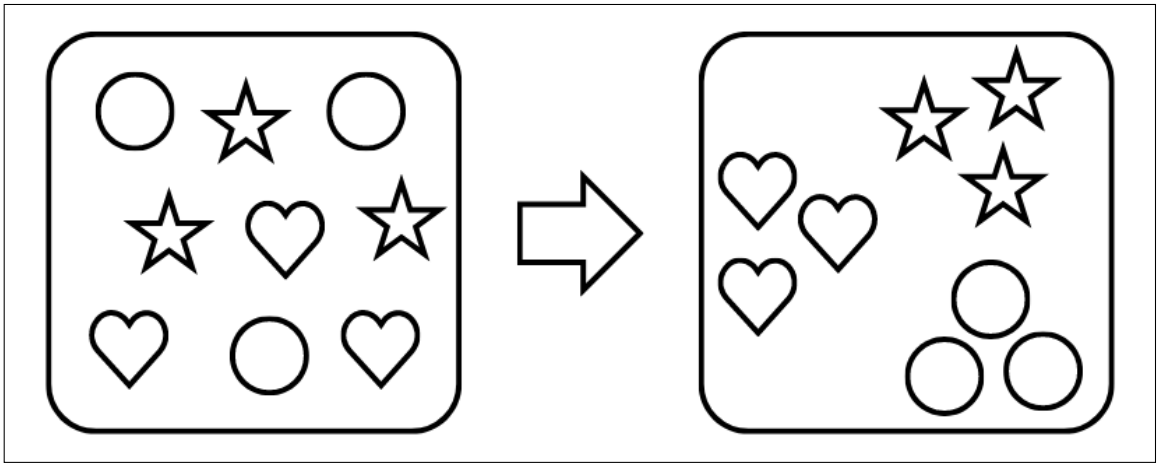

Figura 2.2: Agrupamento dos dados
Fonte: adaptada de (Duran & Odell 2013)

Os tipos de agrupamento ou clustering são descritos dependendo da sua visualização. A visualização do tipo hard clustering é o agrupamento onde cada dados é organizado em apenas um cluster, enquanto o soft é a visualização onde cada dado pode ser agru-

pado em mais de um cluster (Hart et al. 2000). O agrupamento plano é feito utilizando os agrupamentos em um único nível, enquanto o agrupamento hierárquico é realizado com os agrupamentos feitos em árvore com vários níveis (Hart et al. 2000); a tarefa em quer apenas um grupo de clusters é gerado como resultado é chamado de cluster de visualização única, enquanto uma tarefa que gera diversos grupos de clusters é chamada de cluster de visualização múltipla. O clustering tem um custo computacional alto; isto leva a complexidade quadrática ao número de dados (Mullen & Collier 2004).

Considerando o processo de agrupamento de dados pelos algoritmos de aprendizado de máquina não supervisionado, inicialmente, um grupo de dados não rotulados é fornecido como entrada e o algoritmo de aprendizado não supervisionado decide os agrupamentos. Define-se o esquema de computação das semelhanças entre os dados e configurações de parâmetros. Ao executar o algoritmo de aprendizado não supervisionado, o grupo de dados é segmentado em subgrupos de dados semelhantes. O agrupamento pode automatizar as tarefas preliminares para a classificação predefinindo as categorias com uma lista ou uma árvore de agrupamentos e organizando os dados agrupados com os dados de amostra (Jo 2006).

O objetivo em agrupar dados é maximizar as semelhanças entre os dados dentro de cada cluster e minimizar as semelhanças entre os clusters (Jo 2006), (Jo & Lee 2007). O valor que calcula a média da semelhança entre os dados de cada cluster é chamado de similaridade intra-cluster ou coesão. O valor que calcula a média das semelhanças entre os clusters é chamado de similaridade inter-cluster e faz a discriminação entre os clusters. Assim, a maximização tanto da coesão quanto da discriminação é o objetivo do agrupamento dos dados. O requisito mínimo para a implementação dos sistemas de agrupamento é que a coesão e a discriminação sejam maiores do que os resultados do agrupamento aleatório.

Considerando as diferenças entre o agrupamento e a classificação, a classificação requer as tarefas preliminares, a predefinição da categoria e a coleta de dados amostrais enquanto o agrupamento não. A classificação precisa da divisão do conjunto de dados em conjunto de treinamento e teste, enquanto o agrupamento não.

2.2.3 Associação

A associação é definida como o processo de extração das regras de associação na forma de se-então, conforme mostrado na Figura 2.3 (Pereira et al. 2008). A associação destina-se inicialmente a analisar as tendências de compra dos clientes de grandes mercados como o Wall Mart, por exemplo, pretende-se descobrir se um cliente compra cerveja, então ele também compra fraldas. Os grupos de dados chamados conjunto de dados são fornecidos como entrada desta tarefa e uma lista de regras de associação se-então é gerada como saída. A associação de dados é diferenciada do agrupamento onde a semelhança entre os dois dados é bidirecional, já na associação é unidirecional (Jo 2019).

A associação de dados, que é a tarefa inicial da mineração de dados, tem como objetivo obter tendências de compra dos clientes, por exemplo. O conjunto de dados consiste nos itens comprados pelos clientes; cada dado do conjunto é uma lista dos itens comprados pelo cliente em uma única transação. Para cada item, os conjuntos de dados que o incluem

Figura 2.3: Associação dos dados
Fonte: adaptada de (Jo 2019)

são selecionados e as regras de associação são extraídas dos conjuntos selecionados. A regra de associação é dada, simbolicamente, como a regra do se-então; se A então B (A $\rightarrow$ B). No entanto, A $\rightarrow$ B nem sempre é igual a B $\rightarrow$ A nas regras de associação (Jo 2019).

O algoritmo Apriori é determinado por Agrawal et al. (1994) para aprendizagem constante de regras de associação na mineração. Funciona identificando as variáveis individuais frequentes na base de dados e expandido-as a conjuntos de varáveis cada vez maiores, desde que esses conjuntos de varáveis apareçam com frequência suficiente. O nome Apriori vem do fato de o algoritmo utilizar o conhecimento prévio das propriedades dos conjuntos de dados (Tan et al. 2016).

Será explanado as medidas importantes, suporte e confiança, para fazer a associação de dados (Tan et al. 2016). Suporte refere-se à taxa de ocorrências de A e B em conjuntos de dados para uma relação A $\rightarrow$ B (A implica em B), indicando desta forma sua relevância (Tan et al. 2016). Confiança refere-se à taxa de ocorrência nos conjuntos de dados em que A ocorre e também há ocorrência de B, indicando desta forma a validade da relação. No processo de associação de dados, o suporte é utilizado para expandir os subconjuntos de dados e a confiança é utilizada para gerar a relação de associação para cada subconjunto de dados. Ao comparar o suporte com a confiança, o numerador é o mesmo em ambos, mas os denominadores são diferentes, como visto nas equações 2.1 e 2.2, onde σ é o número de itens, N é o número total de itens.

$$Suporte = \frac{\sigma(A \bigcup B)}{N} \tag{2.1}$$

$$Confianca = \frac{\sigma(A \bigcup B)}{\sigma(N)} \tag{2.2}$$

2.2.4 Mineração de Big Data

Nos últimos anos foi introduzido um tipo novo de mineração de dados chamado mineração de big data. Big data é caracterizado por sua variedade, velocidade, variabilidade e veracidade, assim como seu volume (Wu et al. 2013). Mídias, como smartphones, sensores e outros equipamentos onipresentes coletam dados diariamente gerando uma quantidade massiva de dados, conhecida como geração Big data. Os algoritmos tradicionais de processamento de dados possuem algumas limitações para processar big data.

O big data é caracterizado da seguinte forma:

- Variedade: Big Data consiste em dados de diversos formatos, então o pré-processamento torna-se demasiadamente complicado em comparação à mineração de dados relacional e mineração web;
- Velocidade: Em big data, os dados são atualizados com muita frequência; um grande volume de dados é adicionado e excluído dentro de um determinado tempo;
- Variabilidade: Os dados provenientes de big data têm muita inconsistência e ruído, por esta razão leva-se bastante tempo para limpá-los;
- Veracidade: Em big data é muito usual discriminar os dados por sua qualidade, sendo alguns dados considerados muito confiáveis e outros não confiáveis.

As três camadas da mineração de big data podem ser vistas na Figura 2.4. As técnicas de mineração de dados relacionais são mostradas no círculo interno da Figura 2.4, como base da mineração de big data. No círculo médio, são apresentadas as técnicas de mineração de texto, mineração web e mineração biológica. No círculo mais externo, estão as técnicas de mineração de big data, onde os dados têm estruturas mais complexas, atualizações constantes e valores incompletos que são fornecidos como fonte. As técnicas de mineração de dados relacionais tornam-se a base para o desenvolvimento de técnicas para mineração de texto, mineração web e mineração biológica, os três tipos de mineração tornam-se a base para as técnicas de mineração de big data (Jadhav 2013).

Em virtude das técnicas individuais de codificação de dados brutos em formas estruturadas e da classificação ou organização de dados utilizados nos três tipos de mineração de dados não serem viáveis para as tarefas de mineração de big data, é necessário considerar os objetivos para desenvolvê-las (Wu et al. 2013). Como a mais simples, pode-se citar a junção de técnicas que foram desenvolvidas nos tipos existentes de mineração de dados para processamento de dados em diversos formatos. Ao aplicar os algoritmos de aprendizado de máquina existentes, considera-se mais casos como as atualizações frequentes de dados pela exclusão, inserção e discriminação entre os dados. Precisa-se desenvolver novos algoritmos de aprendizado de máquina para tornar estes viáveis para a mineração de big data (Sowmya & Suneetha 2017).

Considerando a relação da mineração de texto com a mineração de big data, na mineração de texto, os dados textuais são fornecidos como entrada em formato uniforme, enquanto na mineração de big data, são utilizados diversos formatos de dados. Na mineração de texto se tem como premissa que os dados são corrigidos ou atualizados com pouca frequência, enquanto a mineração de big data tem como premissa que eles são atualizados constantemente. Na mineração de texto, os dados são coletados pela Internet,

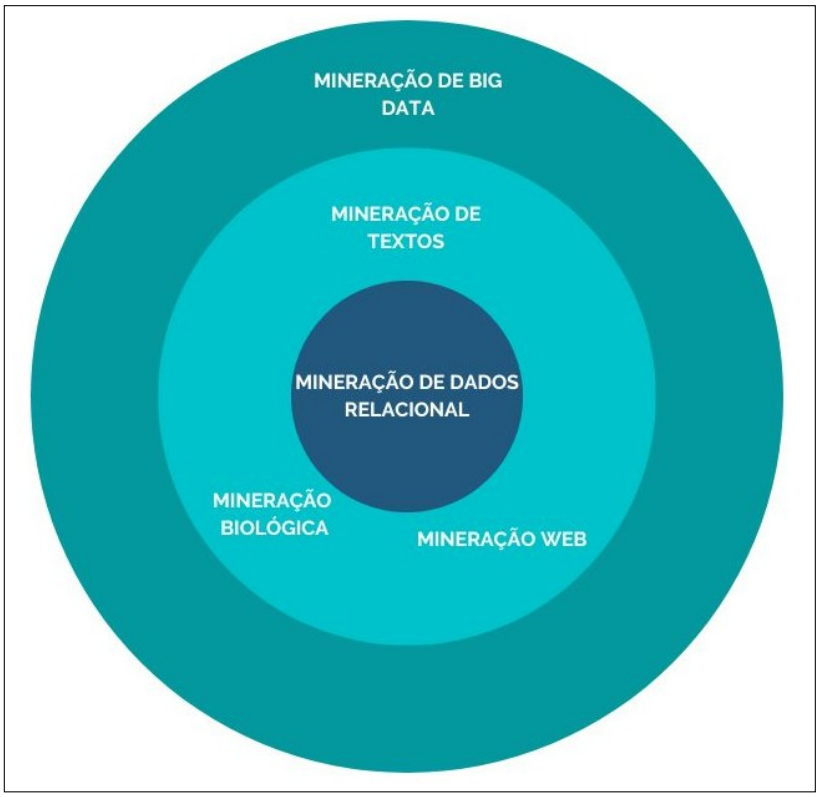

Figura 2.4: Três camadas da mineração de big data
Fonte: adaptada de (Jadhav 2013)

enquanto na mineração de big data, os dados são coletados de diversas fontes como: etiquetas RFID, sensores e smartphones. Ainda que as mensagens trocadas em smartphones pertençam ao big data, ela são consideradas fontes de mineração de texto, bem como os textos fornecidos pela Internet (Sowmya & Suneetha 2017).

2.3 Passos da indexação do texto

Esta seção trata das três etapas básicas da indexação do texto. Será explicado inicialmente a tokenização. Em seguida, será visto o stemming e a remoção de stop-words, respectivamente. Na última subseção, será coberto os esquemas de ponderação de palavras/termos. Portanto, esta seção tem como objetivo explanar o processo de indexação de um texto com as três etapas básicas e o cálculo dos pesos das palavras.

2.3.1 Tokenização

A tokenização é definida como o processo de segmentação de um texto em tokens separadas pelo espaço em branco ou sinais de pontuação. É possível aplicar a tokenização utilizando os códigos-fonte em C, C++ e Java (Aho et al. 1985), bem como aos textos escritos em linguagem natural. A análise morfológica é necessária para tokenizar textos escritos em idiomas orientais: chinês, japonês e coreano. Assim, omitindo a análise morfológica, será explicado o processo de tokenização de textos escritos em línguas de origem do latim.

A visão funcional da tokenização é ilustrada na Fig. 2.5. Um texto é fornecido como entrada e a lista de tokens é gerada como saída no processo. O texto é segmentado em tokens pelos espaço em branco ou sinais de pontuação. Como processamento subsequente, as palavras que incluem caracteres especiais ou valores numéricos são removidas e os tokens são alterados para seus caracteres minúsculos. A lista de tokens torna-se a entrada das próximas etapas da indexação de texto: o stemming ou a remoção de *stopwords* (Cai et al. 2019).

O processo de tokenização de um texto funciona do seguinte modo:

- O texto fornecido é dividido em tokens pelo espaço em branco, sinais de pontuação e caracteres especiais;
- As palavras que incluem um ou alguns caracteres especiais, como “16%”, são removidas;
- O primeiro caractere de cada frase que é dado como o caractere maiúsculo, deve ser alterado para o minúsculo;
- Palavras redundantes devem ser removidas após as etapas de indexação do texto.

Considerando o processo de tokenização do texto escrito em uma das línguas orientais, chinesa ou japonesa, as duas linguagens não permitem o espaço em branco em suas frases, então é impossível tokenizar o texto utilizando essa técnica. É necessário desenvolver e anexar o analisador morfológico que segmenta o texto dependendo das regras gramaticais para tokenização do texto. Mesmo que o espaço em branco seja permitido

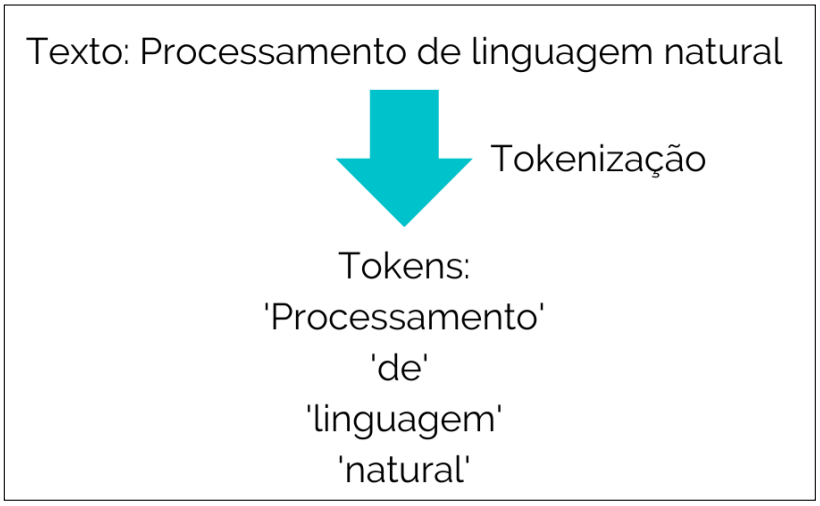

Figura 2.5: Visão funcional da tokenização
Fonte: adaptada de (Cai et al. 2019)

nos textos coreanos, os tokens que são resultados da segmentação utilizando o espaço não são completos, então o analisador morfológico também é necessário (Cai et al. 2019).

2.3.2 Stemming

Stemming refere-se ao processo de mapeamento de cada token que é gerado a partir da etapa anterior em sua forma raiz (Kowalski & Maybury 2000). As regras de stemming são as regras de associação dos tokens com sua forma raiz. O stemming geralmente é aplicável a substantivos, verbos e adjetivos, como mostrado na Fig. 2.6. A lista de formulários raiz é gerada como saída desta etapa. Portanto, nesta subseção, descrevemos o stemming que é o segundo ou terceiro passo da indexação de texto (Moral et al. 2014).

No stemming, os substantivos que são dados em sua forma plural são convertidos para sua forma singular, como mostrado na Figura 2.6. Para convertê-lo em sua forma singular, geralmente, o caractere “s” é removido do substantivo. No entanto, é preciso considerar alguns casos excepcionais no processo de stemming; para alguns substantivos que terminam com o caractere “s”, como “cores”, o sufixo, “es”, deve ser removido em vez de apenas o “s”, e também quando as formas plural e singular são diferenciadas no final uma da outra, como o caso das palavras, “opinião” e “opiniões”. Antes de fazer o Stemming nos substantivos, é preciso classificar as palavras em substantivos utilizando o POS-Tagging. Depois, serão utilizadas as regras de associação para cada substantivo com suas formas no plural para implementar o radical (Singh & Gupta 2016), (Moral et al. 2014).

Agora considerando o caso de stemming nos verbos para seus radicais, os verbos que são dados em suas formas na terceira pessoa, presente, passado ou como substantivos ou como partícula podem ser alterados para suas formas no infinitivo, removendo o sufixo. No entanto, os verbos irregulares onde sua forma infinitiva são diferentes das formas

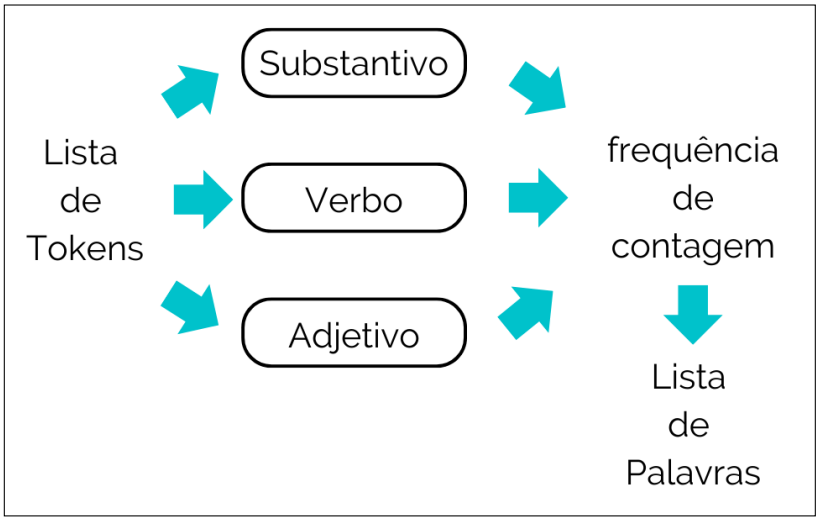

Figura 2.6: Processo do stemming
Fonte: Adaptada de (Moral et al. 2014)

conjugadas, são tratados como casos excepcionais (Singh & Gupta 2016).

2.3.3 Remoção de stopwords

A remoção de *stopwords* refere-se ao processo de remoção de palavras da lista de tokens ou palavras derivadas (Kowalski & Maybury 2000). As *stopwords* são as palavras que são irrelevantes gramaticalmente para o conteúdo do texto, então elas precisam ser removidas para se ter maior eficiência. A lista de palavras irrelevantes ou *stopwords* é carregada de um arquivo e, caso as palavras estejam registradas na lista, são removidas do texto na etapa de pré-processamento. O processo de stemming e a remoção de *stopwords* podem ser trocados; as *stopwords* são removidas antes de fazer o stemming dos tokens, por exemplo.

A *stopword* refere-se à palavra que funciona apenas como uma ligação gramatical nos textos e é irrelevante para o conteúdo do texto fornecido. Preposições, como “a”, “de”, “em”, e assim por diante, e normalmente pertencem ao grupo de *stopwords*. Conjunções como “embora”, “caso”, “conforme” e “para que” também pertencem ao grupo. Os artigos definidos, “o/os” e “a/as”, e os artigos indefinidos, “um/uma” e “umas/uns”, por exemplo, também são *stopwords* frequentes. As *stopwords* ocorrem predominantemente em todos os textos da coleção; removê-las faz com que melhorar muito a eficiência no processamento de textos (Sarica & Luo 2021).

Explicando o processo de remoção de *stopwords* no arquivo, a lista de *stopwords* é preparada como um arquivo e é carregada a partir dele. Para cada palavra do texto, caso esteja registrada na lista, ela é removida. As palavras restantes após a remoção geralmente são substantivos, verbos e adjetivos. Em vez de carregar a lista de *stopwords* de um arquivo, pode-se considerar usar um classificador que decide se a palavra é uma *stopword* ou não (Wilbur & Sirotkin 1992).

Alguns substantivos, verbos e adjetivos restantes podem repetir em vários textos. Essas palavras são chamadas de palavras comuns e não são úteis para identificar o conteúdo do texto, portanto, precisam ser removidas como as *stopwords*. O peso do TF-IDF (*Term Frequency Inverse Term Frequency*) é o critério para decidir se uma palavra é uma palavra comum, ou não, e será melhor explicado em sequência. As palavras restantes após a remoção das *stopwords* e palavras comuns são úteis na análise do texto. Pode-se considerar a remoção adicional, dependendo do tamanho do texto; a remoção é aplicada na maioria dos casos aos textos mais longos (Sarica & Luo 2021).

Consida-se a análise de sentimentos como um tipo de classificação de textos; é o processo de classificar textos em uma das três categorias: positivo, neutro e negativo. Pode ser necessário algumas *stopwords* para realizar a tarefa, em casos excepcionais. Por exemplo, as *stopwords*, "sem" e "contra", indicam uma opinião negativa. Nem todas as *stopwords* são úteis para identificar o sentimento nos textos; é preciso decidir se cada palavra é útil ou não. Ao implementar o sistema de processamento de texto para fazer a análise sentimental, o processo de indexação precisa ser modificado (Saif et al. 2014).

2.3.4 Ponderação do Termo

O termo ponderação refere-se ao processo de calcular e atribuir o peso de cada palavra com seu grau de importância. Pode-se precisar do pré-processamento para remover as palavras, como as *stopwords* para maior eficiência. A frequência do termo e o peso são os esquemas populares para ponderação de palavras, por isso serão descritos formalmente, mostrando suas equações matemáticas. O peso do termo pode ser usado como valores dos atributos na codificação de textos em vetores numéricos.

Podemos usar a frequência do termo que é a ocorrência de cada palavra no texto dado como o esquema de ponderação das palavras (Luhn 1957). Assume-se que as *stopwords* que ocorrem com mais frequência nos textos são completamente removidas no pré-processamento. As palavras são ponderadas contando suas ocorrências no texto. Existem dois tipos de frequência de termos: a frequência de termos absoluta, que é a ocorrência de palavras e a frequência de termos relativa, que é a razão máxima de suas ocorrências. A frequência relativa de termos é preferida à absoluta, a fim de evitar a superestimação e subestimação ao decorrer do texto.

O TF-IDF (*Term Frequency-Inverse Document Frequency*) é o esquema de ponderação de palavras mais popular na área de recuperação de informação, mas requer as referências de toda a coleção de texto que é chamada de corpus. Os pesos das palavras que são calculados pelo esquema TF-IDF são proporcionais às ocorrências no texto dado, mas inversamente proporcionais aos de outros textos. O peso TF-IDF, (Salton & Yang 1973).

O cálculo do TF-IDF é dividido em duas partes: 1) calcula-se separadamente o TF e o IDF; 2) depois multiplica-se as duas partes para obter o valor final, a equação 2.3 calcula o TF, que mede a probabilidade de uma palavra p ocorrer em um texto t.

$$TF_{pt} = \frac{n_{pt}}{\sum_k n_{kt}} \tag{2.3}$$

onde n_{pt} é o número de vezes que a palavra p acontece no texto t; no denominador

aparece o somatório da frequência de todas as palavras no texto. A equação 2.4 calcula a relevância geral de uma palavra em todos os textos da base de dados, IDF.

$$IDF_p = \log \frac{|T|}{|t_p : n_{pt} \neq 0|} \tag{2.4}$$

onde $|T|$ indica o número de textos total na base de dados, e o número de textos em que a palavra ocorre pelo menos uma vez.

2.4 N-gramas

Na literatura da mineração de texto, durante o processo de análise de texto, um problema fundamental é como representar o texto de forma eficaz e modelar seu tópico, não apenas sob a perspectiva do desempenho do algoritmo, mas também para que os analistas interpretem melhor e apresentem os resultados. Uma abordagem comum é usar n-grama, ou seja, uma sequência contígua de n unigramas, como unidades básicas (Cavnar et al. 1994). A Figura 2.7 mostra uma sequência de exemplo com a representação correspondente de 1-grama (unigrama), 2-grama (bigrama), 3-grama (trigrama) consolidada. No entanto, tal representação levanta preocupações de crescimento exponencial do dicionário, bem como a falta de interpretabilidade. Pode-se razoavelmente esperar um método inteligente que use apenas um subconjunto compacto de n-gramas, mas gere uma representação explicável dado um documento (Cavnar et al. 1994).

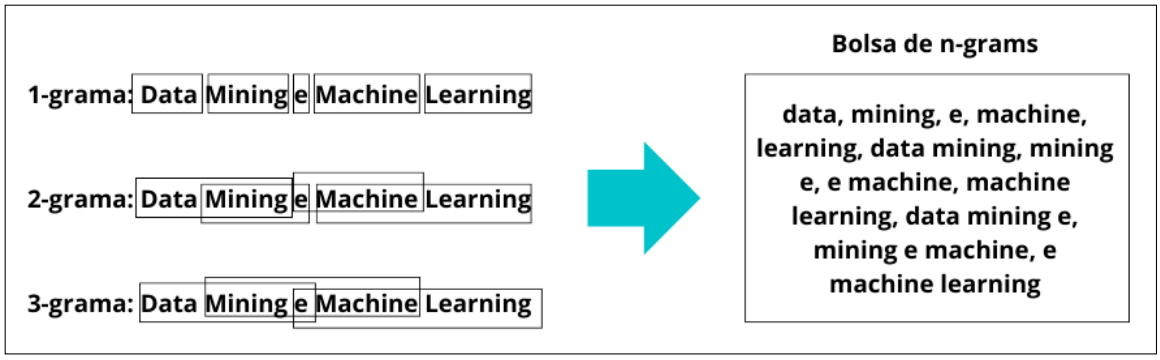

Figura 2.7: Exemplo da representação do n-grama
Fonte: adaptada de (Cavnar et al. 1994)

Capítulo 3

Trabalhos relacionados

Com o objetivo de obter trabalhos correlatos, foi realizada uma pesquisa bibliográfica como ponto de partida, apresentando o estado da arte de 2010 a 2021 e abordando os periódicos mais relevantes dentro desse período e consiste em demonstrar o que a pesquisa científica está fazendo de inovador na mineração de textos, processamento de linguagem natural, análise de conteúdo e análise de textos. Além disso, abordará também a pesquisa e experiência com ciência de dados em saúde pública que serviu de alicerce para o desenvolvimento do método. As pesquisas foram feitas nas bases de dados:

- Scielo;
- ACM;
- Science Direct;
- IEEE Xplore;
- PubMed Central.

Os artigos foram organizados utilizando uma combinação das palavras chaves como mostrado na tabela 3.1. As bases Scielo e IEEE Xplore não foram adicionadas à tabela, pois, foram encontrados poucos ou nenhum artigos relacionados ao tema.

Tabela 3.1: Palavras chaves utilizadas no estado da arte

Palavras Chaves	**ACM**	**Science Direct**	**PubMed Central**
("Natural Language Processing") AND (Syphilis) OR (“STI” OR “Sexually Transmitted Infections”)	1.178	16.123	45
("Machine Learning") AND ("Syphilis") OR (“STI” OR “Sexually Transmitted Infections”)	1.189	16.142	156
("Machine Learning") AND ("Health information systems")	28	120	263
("Natural Language Processing") AND ("Health information systems")	13	68	169
("Natural Language Processing") AND ("Health systems")	58	308	605
("Machine Learning") AND ("Health systems")	176	865	1.438
("text mining") AND (Syphilis)	2	15	35

O Estado da Arte está organizado inicialmente com o estudo do AVASUS que serviu de base de conhecimento para desenvolvimento da tese, seguido de alguns artigos mais relevantes e por último dez artigos selecionados, um de cada ano, de 2010 a 2020 na linha do tempo.

3.1 Processamento de Linguagem Natural e mineração de textos: o caso do AVASUS

O Sistema Único de Saúde (SUS) é um dos maiores e mais complexos sistemas de saúde pública do mundo, envolvendo desde a Atenção Primária à saúde com o atendimento para avaliação da pressão arterial, até o transplante de órgãos, sustentando o acesso integral e gratuito a toda a população brasileira. O SUS é formado por mais de cinco milhões de profissionais e, a necessidade de aproximar a formação desses profissionais das reais necessidades dos usuários e do sistema têm sido um desafio.

Como solução para formação em massa desses profissionais e a necessidade do uso de cursos mediados por tecnologia em ambientes online direcionadas à educação continuada desses profissionais, o Ambiente Virtual de Aprendizagem do Sistema Único de Saúde (AVASUS) foi criado integrando módulos educacionais e de extensão desenvolvidos por entidades e profissionais da saúde (Nóbrega et al. 2016). Na plataforma, até julho de 2022, mês em que foi realizada a última coleta de dados, estavam disponíveis para matrícula 324 cursos ativos, 872.540 usuários totais cadastrados, 2.261.486 matrículas realizadas

em cursos e 1.393.060 usuários com direito a certificação. Esses dados demonstram a robustez da plataforma aliando educação de qualidade e o alcance em larga escala com formação massiva de profissionais (AVASUS 2020).

Na plataforma AVASUS, vários cursos não possuem tutoria, portanto, melhorias estão sendo aplicadas para aprimorar a plataforma, tornando-a mais automatizada e reduzindo a intermediação humana ((Vieira et al. 2016), (Nóbrega et al. 2016), (Souza, Vieira, Coutinho & Valentim 2018) e (da Rocha et al. 2019)). Com esse objetivo, esse estudo utilizando PLN e mineração de textos foi desenvolvido para realizar o reconhecimento de autoria de textos digitados pelos usuários nos fóruns do curso, enriquecendo ainda mais o sistema e aumentando a automação do AVASUS, garantindo a integridade do usuário que acessa corretamente a plataforma de forma automatizada. Portanto, pode representar o início de um sistema de ensino a distância mediado por tecnologia e inteligência artificial.

A Figura 3.1 mostra todos os passos do modelo completo do sistema de reconhecimento de autoria. Inicialmente, os textos são adquiridos e formam um banco de dados que passa pela etapa de pré-processamento. Os dados pré-processados são divididos em validação e teste, formando dados para treinamento e validação. A etapa de extração de características é aplicada aos dados. Em seguida, os dados passam pela etapa de treinamento constituindo o modelo para reconhecimento de autoria.

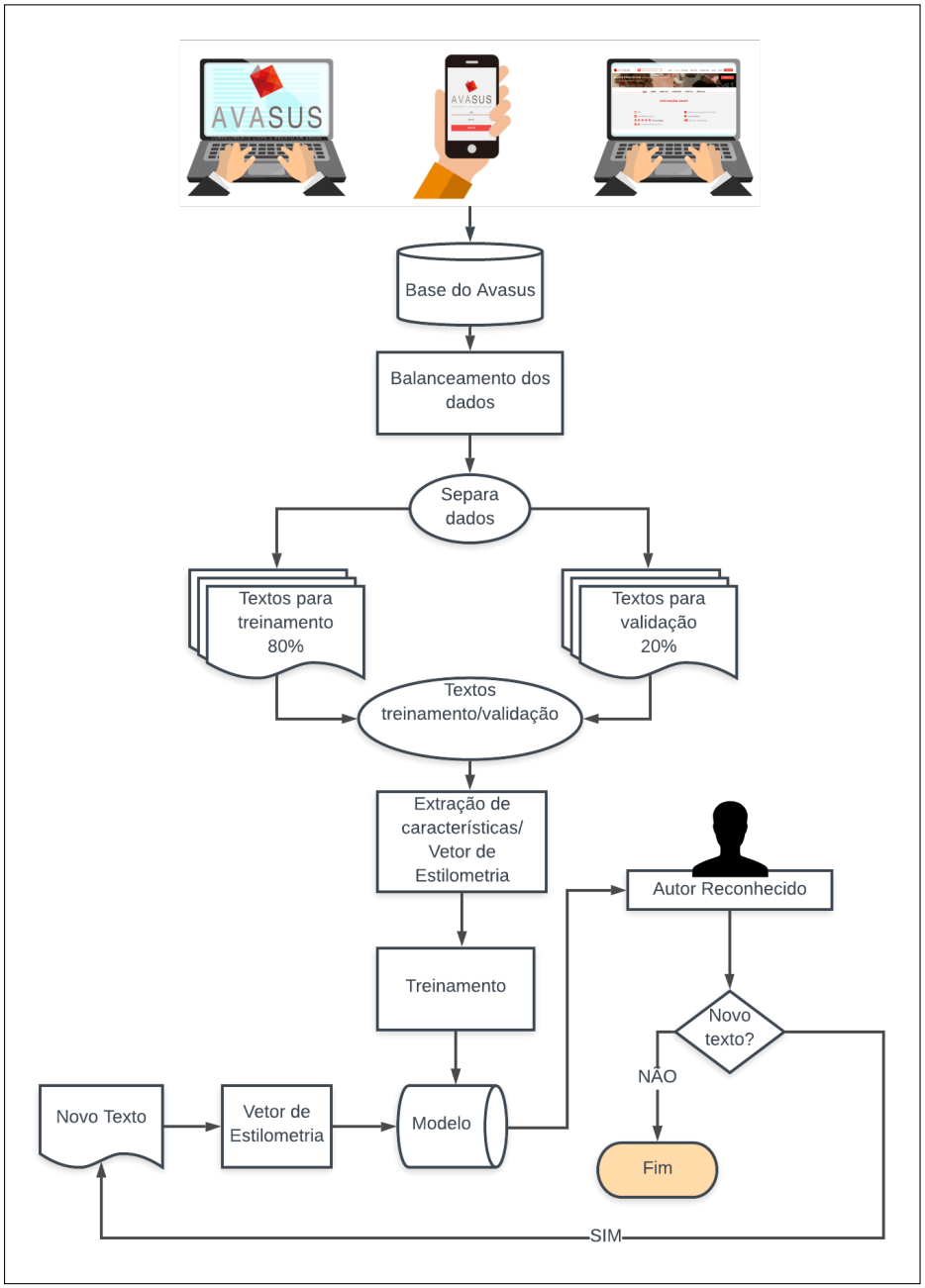

Figura 3.1: Representação esquemática do fluxo do sistema
Fonte: Autoria Própria

A base de texto AVASUS abrange alguns artigos de conclusão de curso e respostas de fóruns que os alunos postam praticamente todos os dias no sistema. O número de autores, o tamanho do documento e o grande número de entradas de texto com tendência a crescimento foram as características que influenciaram a seleção da base de dados utilizada. Uma análise inicial deste banco de dados foi realizada por (da Rocha et al. 2019) usando BoW combinado com Term Frequency-Inverse Document Frequency (TF-IDF). O objetivo desta análise foi verificar a semelhança dos padrões de escrita entre os textos de um mesmo autor em cada um dos cenários de dados considerados.

O carregamento e a preparação dos dados do AVASUS foram os primeiros passos para o modelo. O objetivo foi organizar os dados para a etapa de pré-processamento realizada posteriormente. Na etapa de pré-processamento para padronização de dados, os dados nulos, as tags HTML e os espaços em branco foram removidos. Pouco depois, as palavras irrelevantes foram removidas. A maioria delas eram palavras funcionais comumente usadas na escrita como artigos, preposições, pronomes e conjunções que não têm significado/relevância no texto e aparecem com alta frequência no corpus. Assim, eles podem ser removidos sem afetar o contexto do que foi escrito.

Uma análise de extração foi realizada para determinar a relevância das características distintas para cada configuração de teste. Seguindo um estudo anterior (El & Kassou 2014), o conjunto foi nomeado como Características Estilométricas. Essas características constituíram o Vetor de Estilometria que serviu de entrada para o modelo. Após isso, vários algoritmos de ML foram treinados e validados. Os principais algoritmos escolhidos para esta etapa foram Support Vector Machine (SVM) e Logistic Regression (LR), alguns resultados podem ser vistos na tabela 3.2.

Tabela 3.2: Resultados de classificação

Teste	**Classificadores/Precisão (%)**					
#	**SVM**	**LR**	NB	KNN	BNB	GCP
1	92	92	78	84	86	92
2	66	55	45	70	65	60
3	93	91	58	86	84	94
4	63	64	32	47	63	58
5	89	83	56	79	73	86
6	94	92	85	93	94	92
7	76	76	55	74	80	12
8	54	60	26	38	26	55
9	66	65	37	51	39	22
10	73	67	67	69	74	28
11	62	66	54	51	63	57
12	84	68	63	63	63	68
13	75	71	71	64	64	10
14	62	67	44	51	55	18

Os resultados apontados comprovam com a boa precisão do algoritmo, que a formação do vetor de estilometria utilizando métodos para extração de características pode ser aplicado para reconhecimento de autoria do texto. O SVM e o LR, entre os seis classificadores, indicaram os melhores desempenhos na média dos testes, o que não exclui os testes com outros classificadores que também alcançaram bons resultados em alguns testes. Esse estudo foi de grande importância para formação de conhecimento da base teórica para construção do método desenvolvido nesta tese de doutorado.

3.2 Técnicas computacionais aplicadas na análise de conteúdo e análise de textos em saúde pública: Uma revisão sistemática

Analisar grandes quantidades de dados a partir de métodos manuais tradicionais levariam anos. Diante desse desafio, estratégias computacionais têm surgido para interpretar essas informações. Algoritmos de mineração de textos, Big Data e PLN podem ser aplicados para essas finalidades. Analisar um grande volume de dados em formato de texto poderá auxiliar na compreensão do papel do pesquisador de campo no território, e como utilizar dessa estratégia para otimizar a redução da epidemia de sífilis no país.

Teóricos como (Laurence 2011), discutiram a utilidade e relevância da informática para complementar técnicas de análise de conteúdo, desde que os pesquisadores preparem instruções não ambíguas. Nesta perspectiva, a análise de conteúdo pode ser automatizada em diversos graus (totais ou parciais), o que implica em consequências diretas sobre a prática da análise, quais sejam, rapidez aumentada, acréscimo de rigor na organização da investigação, possibilidade de manipulação de dados complexos, dentre outros (Laurence 2011). Alguns trabalhos foram feitos em várias abordagens diferentes e os trabalhos mais relevantes estão descritos abaixo.

(Villena & Dunstan 2019) utilizaram uma técnica de processamento de linguagem natural, o algoritmo de frequência ponderada, Term Frequency-Inverse Document Frequency (TF-IDF), para encontrar as palavras chave que definiam mais satisfatoriamente as causas das consultas por especialidade de patologias não cobertas pelas Garantias Explícitas da Saúde (GES) nos hospitais públicos do Chile. O banco de dados utilizado pelos autores foi o sistema de gerenciamento do tempo de espera (SIGTE) e contém dados pessoais e de consultas dos pacientes, tal como, a especialidade e o diagnóstico relatados em textos livres não estruturados. A fim de mostrar a relevância de uma palavra dentro do texto relacionada a especialidade, (Villena & Dunstan 2019) buscou evidenciar as palavras que surgiram com mais frequência na especialidade, mas não aquelas palavras que apareceram igualmente em todas elas, ou seja, explorou-se as informações-chave, próprias das especialidades médicas. Para cada especialidade médica e odontológica da base de dados utilizada, foram estabelecidas quatro palavras chave com a maior frequência ponderada por sua relevância no texto e para adquirir uma representação visual dessas palavras foi definida uma nuvem de palavras.

(Șerban et al. 2019) desenvolveram o sistema SENTINEL utilizando aprendizado de máquina para detectar surtos de doenças em tempo real utilizando dados de vigilância do

CDC, feed de notícias e, principalmente, o Twitter. O SENTINEL detecta eventos monitorando o fluxo do Twitter, armazenando aquelas informações relacionadas a doenças e, em seguida, conta diariamente os tweets mencionando os sintomas das doenças para cada estado dos EUA. Os autores ofereceram três principais aplicações em vigilância de doenças: Detecção de alerta precoce, que fornece um aviso antecipado das ocorrências iminentes de saúde; Informações relacionadas a ocorrências de saúde que possam ter acontecido; E, previsões a curto prazo que reúnem dados de monitoramento de fluxo do Twitter sobre as doenças. Para isso, utilizaram o algoritmo Nowcasting para combinar os dados do CDC de semanas anteriores juntamente com os eventos diários do Twitter para prever o nível atual da doença.

(Luo et al. 2019) implementaram um framework utilizando processamento de linguagem natural (PLN) para analisar as opiniões reproduzidas no Twitter no período entre 2008 e 2017, em relação a vacinação em combate ao HPV. A pesquisa inclui inicialmente uma análise de sentimentos das postagens do Twitter no período abordado e, a partir dos sentimentos extraídos, faz-se uma análise de entidade para indicar a organização, entidades e localização geográfica de ocorrências relacionadas aos tweets classificados como negativos e positivos e a mineração utilizando Inteligência Artificial (IA) da associação de frases para apontar os tópicos mais relevantes retratados nos tweets positivos, negativos e detalhados.

Com o objetivo de melhorar modelos preditivos de diagnóstico de HIV, (Feller et al. 2018) utilizou o Processamento de Linguagem Natural (PLN), com os dados extraídos do Clinical Data Warehouse (CDW) do Hospital Presbiteriano de Nova Iorque - Centro médico da Universidade de Columbia, são dados clínicos de cerca de 5 milhões de pacientes desde 1995, os autores limitaram a extração de janeiro de 2007 a dezembro de 2015. Dados dos Registros Eletrônicos de Saúde incluindo dados demográficos, testes laboratoriais, códigos de diagnóstico e notas não estruturadas antes do diagnóstico de HIV foram, também, extraídos para modelagem. Os autores desenvolveram três algoritmos preditivos utilizando aprendizado de máquina, o primeiro modelo utilizando apenas os dados dos registros eletrônicos que estavam estruturados, o segundo extraindo tópicos utilizando PLN e o terceiro extraindo palavras-chave clínicas usando TF-IDF.

O vocabulário do dataset consistia em cerca de 100.000 palavras únicas e os autores utilizaram chi quadrado, teste univariado para identificar um subconjunto menor de palavras indicando comportamento de alto risco. Após isso, fizeram o teste estatístico utilizando correlação de Pearson para cada palavra e selecionaram as 300 com maior medida de associação. Por fim, 37 palavras-chave clínicas relacionadas ao fator de risco de HIV foram selecionadas manualmente para inclusão no modelo preditivo. Utilizaram também o LDA com o dataset de notas clínicas para a inferir a presença ou a ausência dos tópicos previamente aprendidos. Os modelos demonstraram uma faixa de desempenho regular e o PLN melhorou o desempenho preditivo da avaliação de risco do HIV.

A Papua-Nova Guiné possui uma das taxas mais altas de infecções sexualmente transmissíveis no mundo. Machekera et al. (2021) desenvolveu um modelo de intervenção para eliminação da sífilis chamado SITE, com o empenho de reduzir a incidência da doença no país. O SITE foi utilizado para explorar o impacto esperado e custo de cenários alternativos de aumento de intervenção da sífilis. Os autores utilizaram dados de vigilância de

rotina, pesquisas bio comportamentais, estudos de pesquisa e registros de programas para o SITE que é um modelo dinâmico que simula a transmissão da sífilis em pessoas de 15 a 49 anos. O SITE foi usado para simular quatro tipos de intervenções (tratamento clínico, rastreamento de contato, rastreamento da sífilis e promoção do preservativo), supondo que o tratamento da sífilis seja igualmente eficaz em todas as intervenções (valor padrão de 90%) para todos os grupos populacionais (Machekera et al. 2021).

O modelo desenvolvido por (Machekera et al. 2021) foi programado em C++ como um pacote complementar do R, versão 3.5.1, e projetado para uso pela equipe do programa nacional de HIV/IST. Os valores de entrada de parâmetros definidos pelo usuário são especificados em um arquivo Excel, assim como as saídas do modelo. Os resultados adquiridos pelo SITE previram um aumento de 25%-35% em 2020 para 60% em 2023 da cobertura do tratamento de casos sintomáticos em estágio primário/secundário reduzindo assim a incidência estimada em 2021-2030 em 55%, em comparação com um cenário que pressupõe cobertura constante nos níveis de 2019-2020. O modelo SITE é uma nova ferramenta que os programas nacionais de HIV/IST podem usar para informar e melhorar o planejamento estratégico de controle de IST e a otimização das intervenções (Machekera et al. 2021).

Com os avanços recentes em Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) agora é possível examinar textos relacionados ao HIV em mídias sociais, onde o conteúdo textual está publicamente distribuído em grande escala. Os autores (Young et al. 2017) construíram um classificador de aprendizado de máquina que pode identificar tweets que mencionam comportamentos de risco do HIV (atividade sexual e uso de substâncias associadas à infecção pelo HIV) com precisão semelhante à anotação manual. Até então, a maior parte das análises feitas em grandes quantidades de dados relacionados ao HIV nas mídias sociais são feitas em textos e voltadas apenas para complementar a vigilância tradicional dos resultados de saúde.

Os autores (Nobles et al. 2020) surgiram com uma abordagem diferente, dando o primeiro passo ao descrever imagens que o público compartilha no Instagram, autoidentificadas como referente ao HIV e auto marcadas com a hashtag “HIV”. As postagens públicas que formaram o dataset utilizado pelos autores abrangiam entre janeiro de 2017 e julho de 2018. Utilizando a hashtag “HIV” foi possível coletar através do software InstaLooter, imagens e legendas relacionadas à hashtag. As postagens coletadas foram limitadas àquelas com legenda e autoria em inglês para que as comparações pudessem ser feitas com os alvos de saúde dos EUA e também serem interpretadas por investigadores de língua inglesa.

(Nobles et al. 2020) descreveram a prevalência de hashtags co-ocorrentes, onde, primeiro identificaram as hashtags comuns mais usadas por sua frequência, depois separaram as hashtags mais usadas relacionadas apenas ao HIV ou risco de HIV. Agruparam as hashtags em clusters de acordo com um tópico temático principal e relataram a frequência das postagens dentro de cada tópico. Os autores analisaram o conteúdo visual das imagens das postagens do Instagram usando um software de reconhecimento de imagem automatizado para atribuir tags textuais ao conteúdo das imagens. Depois que cada imagem recebeu uma tag textual, elas foram agrupadas em tópicos usando o LDA.

Porém, não há na literatura conhecida a estratégia de usar métodos de mineração de

textos para identificar as principais ações de intervenções em grandes projetos nacionais, como é o caso do projeto "Sífilis Não".

Assim, com auxílio de um conjunto de algoritmos, os quais compõem um método computacional aplicado à mineração de textos, é possível identificar quais ações de intervenções dos apoiadores do projeto podem ter induzido estratégias de políticas públicas de saúde no território, as quais, apontam para a redução importante dos casos de sífilis congênita.

No entanto, para além da necessidade de interpretar os dados sobre as conexões do Projeto no território, verifica-se que abordagens de pesquisa que interpretem a produção textual, a exemplo da análise de conteúdo ou temática, são desafiadoras, principalmente devido à carência de metodologias consolidadas para utilização e processamento de grande volume de textos.

3.3 Linha do tempo

A tabela 3.3 detalha alguns trabalhos de 2010 a 2020 relacionados com o tema proposto, com um resumo do que foi feito, técnicas utilizadas e banco de dados.

Tabela 3.3: Linha do tempo do estado da arte anos 2010 a 2020

Estudos	**Ano**	**Detalhamento do Trabalho**		
		Resumo	Técnicas	Banco de Dados
(Mathur & Dinakarpandian 2010)	2010	Mapear dados de sequência de proteínas para terminologias de doenças	MetaMap a anotação da doença de entradas da proteína Swissprot; e Co-anotação e hierarquia semântica para estimar a semelhança entre as doenças	Swissprot (SP)
(Olvera-Lobo & Gutiérrez-Artacho 2011)	2011	Estudo para avaliar a eficiência dos sistemas de controle de qualidade como fontes terminológicas para médicos, tradutores especializados e usuários em geral	As questões foram avaliadas como incorretas, inexatas ou corretas. Aplicaram uma série de medidas de avaliação para marcar a qualidade das respostas	MedQA, START, QuALiM e HONqa
(Xia & Yetisgen-Yildiz 2012)	2012	Criaram três corpora para estudos clínicos de PLN: um marca recomendações críticas em relatórios de radiologia e os outros dois indicam se um paciente tem pneumonia com base em relatórios de radiografia de tórax ou relatórios de UTI	Criaram um corpus para treinar e avaliar os sistemas de PLN. Os três projetos lidam com relatórios médicos de pacientes e os corpora foram anotados por médicos	Critical Recommendations in Radiology, PNA and CPIS in Chest X-ray e Pneumonia in the ICU Reports
(Pesaranghader et al. 2013)	2013	Otimizaram a medida de relação semântica do vetor de co-ocorrência de segunda ordem para um melhor desempenho	Latent Semantic Analysis (LSA) para eliminação de características insignificantes e gerar vetores menores, aplicando ambas as abordagens ao domínio biomédico	MEDLINE abstract

Estudos	Ano	Detalhamento do Trabalho		
		Resumo	Técnicas	Banco de Dados
(Sánchez et al. 2014)	2014	Método de higienização automática para registros eletrônicos de saúde capaz de proteger entidades sensíveis (doenças) e também aqueles termos relacionados semanticamente (sintomas)	Método de higienização automática de documentos textuais	Prontuário eletrônico
(Fleuren & Alkema 2015)	2015	Os autores descreveram as técnicas para mineração de texto e visão geral das ferramentas usadas atualmente e os tipos de problemas aos quais elas são normalmente aplicadas	Mineração de texto	Diversos
(Macedo et al. 2016)	2016	Desenvolveram a prova de conceito e a experimentação do Health Surveillance Software Framework (HSSF) com o objetivo de subsidiar estratégias preventivas de saúde.	Recomendação de conteúdo	registros médicos de pacientes simulados
(Castro et al. 2017)	2017	Processamento de Linguagem Natural em um Registro Médico Eletrônico (RME) para identificar com precisão os pacientes com aneurismas cerebrais e seus controles correspondentes.	NLP	CID-9 e Current Procedural Terminology

Estudos	Ano	Detalhamento do Trabalho		
		Resumo	Técnicas	Banco de Dados
(Feller et al. 2018)	2018	Examinaram se o processamento de linguagem natural melhoraria os modelos preditivos de diagnóstico de HIV.	Tópicos e palavras-chave	prontuário eletrônico
(Șerban et al. 2019)	2019	Desenvolveram o sistema SENTINEL utilizando aprendizado de máquina para detectar surtos de doenças em tempo real.	Nowcasting	Twitter
(Ščepanovic et al. 2020)	2020	Desenvolveram uma ferramenta de Deep Learning para Processamento de Linguagem Natural que extrai menções de praticamente qualquer condição médica ou doença de texto não estruturado de mídia social.	Contagem de palavras	Mídia Social

Capítulo 4

Materiais e Métodos

A base teórica que serviu como subsídio para determinar as etapas de desenvolvimento da presente tese foi inspirada em Laurence (2011), e pode ser descrita por:

1. Etapa da Pré-análise na seção 4.1;
2. Etapa da Exploração do material: Mineração de Texto, seção 4.2;
3. Etapa do Tratamento dos resultados obtidos: a inferência e a interpretação, seção 4.3.

De acordo com Laurence Bardin (2011), a análise de conteúdo é um conjunto de instrumentos de cunho metodológico, e em constante aperfeiçoamento, que se aplicam a discursos diversificados. Estes instrumentos evidenciam o desenvolvimento de categorias a partir dos dados e reconhecem a importância de se compreender o significado do contexto em que os itens analisados apareceram (Laurence 2011).

A função primordial da análise de conteúdo é desvendar o crítico em relação a um determinado discurso. Essa análise sobre os dados pode ser realizada de forma qualitativa ou quantitativa. Na primeira forma, a análise é realizada de forma sistemática e analítica, onde o pesquisador revisa os temas ou categorias de análises utilizando conceituação, coleta de dados, análise e interpretação. Já na segunda, a busca é pela quantificação do conteúdo, em termos de categorias predeterminadas, sendo realizada de forma sistemática e replicável (Vespestad & Clancy 2020), (Laurence 2011).

Nas duas formas de análise de conteúdo, o objetivo é realizar uma inferência de conhecimentos relacionados à semântica do discurso predominante nos dados ou na mensagem dada por esses dados. No presente estudo, foi utilizada uma abordagem quali-quantitativa. A escolha se deve ao fato de mesclar métodos de mineração de texto e uma conceituação do discurso extraído, com uso de uma base teórica que busca compreender as ações de intervenção, do ponto de vista da saúde pública no território.

Este capítulo detalha cada uma das etapas citadas nas próximas seções.

4.1 Etapa da Pré-análise

Nesta etapa é onde foi organizado o material a ser analisado com o objetivo de torná-lo operacional, sistematizando as ideias iniciais. Foram realizadas 3 (três) atividades: leitura flutuante, fase responsável pela compreensão geral dos documentos que foram coletados;

escolha dos documentos, que consistiu na demarcação do que foi analisado; e formulação das hipóteses e dos objetivos do trabalho a ser desenvolvido (Laurence 2011).

Na pré-análise foi necessário conhecer e entender (I) as características do projeto "Sífilis Não!", para proposta do objeto desta tese; e (II) as características dos dados (textos) para análise na etapa que se segue - Exploração do Material. Que estão descritas nas subseções em sequência.

4.1.1 Características do projeto "Sífilis Não!"

O projeto denominado "Pesquisa Aplicada para Integração Inteligente Orientada ao Fortalecimento das Redes de Atenção para Resposta Rápida à Sífilis", também nomeado de projeto "Sífilis Não!", foi desenvolvido pelo governo brasileiro, por intermédio de uma cooperação técnica entre a Universidade Federal do Rio Grande do Norte (UFRN) e o Ministério da Saúde em parceria com a OPAS, como a principal intervenção de saúde pública relacionada à sífilis no Brasil dos últimos 10 anos. (Marques dos Santos et al. 2020), (de Morais et al. 2020), (de Andrade et al. 2020).

Dentre as diversas estratégias para o combate à sífilis, este estudo se concentra na rede de apoio institucional conformada por apoiadores que atuaram *in loco* nos municípios prioritários das regiões do Brasil. Tal rede de apoio foi conformada com o objetivo maior de fortalecer os "nexos entre o projeto e os gestores de saúde no território", "articular ações programáticas pactuadas nos órgão de governança do SUS com os planos locais", e oferecer apoio para "resposta oportuna à sífilis nas redes de atenção em saúde" (Priamo et al. 2020).

O projeto contou com 52 apoiadores de pesquisa e intervenção, também nomeados pesquisadores de campo do projeto "Sífilis Não!", distribuídos em 72 dos 100 municípios prioritários, em todas as regiões brasileiras (Priamo et al. 2020). Esses pesquisadores trabalharam junto aos gestores para traçar as melhores estratégias de combate à sífilis, tendo como orientação as seguintes determinações: avaliação dos planos e programações de saúde das secretarias municipais de saúde, fortalecimento de comitês de investigação da transmissão vertical, fortalecimento dos sistemas de informação estratégica para vigilância em saúde e qualificação da notificação, seguimento laboratorial e fechamento de casos de sífilis, bem como operacionalização da linha de cuidado da sífilis em adulto e da criança exposta à sífilis congênita (de Andrade et al. 2020).

Todo o trabalho foi acompanhado por supervisores a partir de um sistema digital de gestão dos apoiadores, denominado "Plataforma LUES" (LAIS/UFRN 2022). A partir dessa plataforma, os pesquisadores produziram milhares de relatos de textos de suas experiências no território nacional (Marques dos Santos et al. 2020).

4.1.2 Características dos dados

A base de dados abrange relatos de visitas técnicas e reuniões, listas de presença, relatórios de acompanhamento do plano de trabalho e pesquisa/estudo, totalizando 4.874 documentos em arquivo de texto que totalizam 3,86 Gigabytes de espaço em disco.

O texto contido em cada arquivo possui cerca de 740 palavras e abrange o período

Tabela 4.1: Dados de indexação dos documentos

Variável	Descrição
Identificador	Número de identificação do documento
Mês	Mês de criação do documento
Ano	Ano de criação do documento
Estado	Estado do Brasil de criação do documento
Região	Região do Brasil de criação do documento

entre maio de 2018 e dezembro de 2020. A grande quantidade de dados textuais e a inviabilidade de analisá-los manualmente foram as características que influenciaram na escolha da base de textos produzida pelos diversos apoiadores e também a necessidade de encontrar fatores preditores da sífilis congênita e sífilis em gestantes no material produzido.

A base dos documentos de texto relacionada ao projeto de Intervenção "Sífilis Não!" utilizada nesta tese é de domínio público. Neste contexto, são utilizados dados secundários, sem qualquer identificação sensível dos participantes ou agentes públicos, que foram coletados após o processo de extração e anonimização dos documentos disponibilizados no repositório da "Plataforma LUES". É importante salientar que, esses dados podem ser acessados por qualquer pesquisador no link: <http://vigilanciasaude.ufrn.br/files/anonymous_reports.csv>.

4.2 Etapa da Exploração do material: Mineração de Texto

Mineração de texto pode ser definida como a descoberta automatizada de conhecimentos, até então desconhecidos, utilizando uma base de dados não estruturada (Kar & Dwivedi 2020) (Kumar et al. 2021). Um elemento-chave deste método é a vinculação das informações em conjunto para formar novos fatos ou novas hipóteses a serem exploradas.

Para a implementação do método de mineração de texto, no presente estudo, foram desenvolvidas as seguintes atividades: i) Detalhamento da Base de dados; ii) Pré-processamento; iii) Extração de palavras e dados; iv) Extração dos N-gramas. A seguir, são apresentadas em detalhe cada uma das atividades.

4.2.1 Detalhamento da Base de dados

A base de dados foi extraída da "Plataforma LUES". Os textos coletados e utilizados tratam sobre relatos de visitas técnicas e reuniões; e relatórios de acompanhamento do plano de trabalho e pesquisa. O intervalo de tempo da coleta dos documentos é de maio de 2018 a dezembro de 2020.

Para formação da base de dados, os documentos passaram por um processo de anonimização e identificação onde foram indexados conforme as informações da Tabela 4.1

Cada registro representa um documento que foi coletado. Esse catálogo de documentos foi salvo em formato Comma Separated Values (CSV) um tipo de arquivo de texto separado por vírgulas (Shafranovich 2005).

4.2.2 Pré-processamento

A atividade de pré-processamento foi dividida em 4 (quatro) tarefas definidas por: i) Padronização dos documentos; ii) Normalização dos textos; iii) Remoção dos documentos duplicados; e iv) Correções textuais.

A primeira tarefa teve como objetivo padronizar a base de dados dos textos, salvando os arquivos separadamente, onde, cada documento de texto foi tratado e renomeado com a seguinte estrutura: arquivoA_B.C, onde:

A = número da atividade do apoiador de pesquisa e intervenção na plataforma;

B = número de identificação do documento gerado pela "Plataforma LUES";

C = extensão do arquivo.

Desta forma, um exemplo de arquivo pode ser definido como: arquivo2923_22305.docx ou arquivo3924_23044.xlsx

A segunda tarefa foi a normalização dos textos. A finalidade foi uniformizar o texto em letras minúsculas; (b) remover espaços em excessos; (c) remover caracteres especiais, símbolos, palavras duplicadas em sequência; e (d) remover acentos e números. Em seguida, foi utilizado um dicionário de stopwords para remoção de aspectos não significativos das partes do idioma. Como os relatórios de textos foram elaborados na língua portuguesa do Brasil, foi utilizado o dicionário de stopwords para este idioma. Este dicionário consiste em 204 palavras exclusivas que são consideradas indesejáveis para uma análise significativa, são essas palavras comuns, preposições, pronomes e artigos, por exemplo: 'de', 'a', 'o', 'que', 'e', 'é', 'do', 'da', 'em', 'um', etc. (Uysal & Gunal 2014). O exemplo de processamento do texto pode ser visualizado na Figura 4.1.

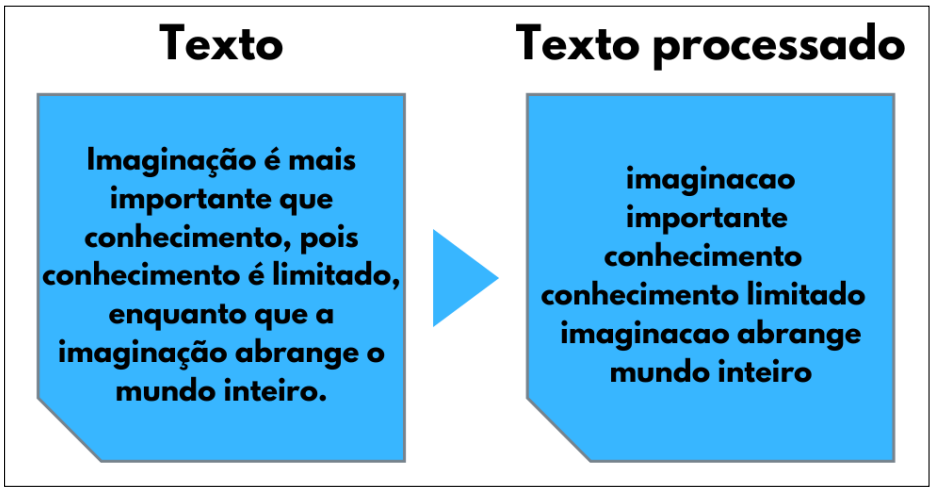

Figura 4.1: Exemplo de processamento do texto
Fonte: Autoria Própria

A terceira tarefa consistiu na remoção dos documentos duplicados que possuem o conteúdo idêntico; e por fim, a quarta tarefa foi a realização das correções textuais necessárias, tais como: correções gramaticais e adição de espaços faltantes em algumas sentenças, utilizando o Algoritmo 1.

Algoritmo 1: ALGORITMO DE CORREÇÃO TEXTUAL

```
1 início
2     correcoes = (palavra : correcao)
3     para palavra em correcoes:
4     para i, linha em base_dados.texto.contem(palavra)
5     se palavra em linha:
6     correcao = linha(texto).substitui(palavra, correcao)
7     texto = correcao
8 fim
```

4.2.3 Extração de palavras e dados

Com o objetivo de retratar a relevância das palavras dentro dos documentos de texto, foi necessário destacar aquelas sentenças que aparecem com frequência e relacionam-se com o assunto nos textos, não dando importância àquelas sentenças que aparecem em todos os documentos e trazem uma obviedade no contexto retratado.

Para obter uma representação da temática tratada nos documentos e extração das sentenças, foi aplicada inicialmente a abordagem dos N-gramas, com $N = 2, 3, 4$ em todo o conjunto de textos e, em seguida, foi feito o agrupamento destes N-gramas utilizando o método de análise de conteúdo. Por fim, foi realizada a contabilização do total de apoiadores e total dos relatórios por região do país.

4.2.4 Extração dos N-gramas: bigramas, trigramas e quadrigramas

Nesta etapa foram feitas extrações automáticas de palavras-chave formadas por uma sequência de duas palavras, as quais o algoritmo julgou, também automaticamente relevantes nos documentos de texto dos pesquisadores de intervenção, como os bigramas (N-grama com $N = 2$).

O primeiro passo para extrair os bigramas foi obter uma lista das palavras ordenadas do texto (*tokens*), utilizando a abordagem Bag of Word (BoW). Em seguida, o BoW foi convertido em uma matriz com a contagem de duas palavras e por fim, foi aplicada a métrica *Term Frequency - Inverse Document Frequency* (TF-IDF) (Salton & Buckley 1988). Essa técnica mede estatisticamente a importância de uma palavra dentro de um texto em relação a outros textos dentro da mesma base de dados. O valor da importância da palavra aumenta proporcionalmente enquanto se aumenta o total de ocorrências dessa palavra no texto, o valor é compensado pela frequência dessa palavra na base de dados. Isso ajuda a discriminar a ocorrência de algumas palavras que são muito comuns, mas não importantes dentro do texto, ver seção 2. Os bigramas também foram explorados

e agrupados por região do Brasil. O mesmo procedimento foi aplicado para os outros n-gramas extraídos.

Na extração dos trigramas (N-grama com N=3) as palavras-chave foram formadas por uma sequência de três palavras e utilizou-se o mesmo procedimento demonstrado para o bigrama, porém com a matriz de contagem adquirida com a formação de três palavras.

Para a extração dos quadrigramas, N-grama com N=4, as palavras-chave foram adquiridas por uma sequência de quatro palavras, utilizando-se o mesmo método mostrado para o bigrama, onde a matriz de contagem foi adquirida com a formação de quatro palavras.

Os N-gramas foram extraídos e agrupados por região do Brasil e por ano (2018-2020). Esta etapa foi realizada com base na identificação das estruturas derivadas dos níveis de atenção à saúde, ações de resposta rápida à sífilis e outras atividades predominantes observadas nos relatos. Em seguida, foi calculada a proporção de termos considerando os dois grupos (região e ano), de acordo com a Equação 4.1 a seguir:

$$Proporcao = \frac{\sum Ngrama\_por\_relatorio}{N^{\circ} relatorios\_por\_regiao/ano} \tag{4.1}$$

4.3 Etapa do tratamento dos resultados obtidos: a inferência e a interpretação

Compreende-se que a produção textual dos apoiadores possui um conteúdo importante para destacar as conexões do projeto "Sífilis Não!" no território, especialmente no contexto da sífilis congênita. No caso em questão, trata-se da área de saúde, especificamente da resposta local à sífilis e dos dados ou subsídios sobre as observações e análises dos apoiadores durante a operacionalização do projeto. Neste estudo, considera-se o projeto "Sífilis Não!" como uma intervenção de saúde pública, a partir da abordagem de programas de saúde pública de Zulmira Hartz, que analisa tais intervenções como ações que favorecem "comportamentos adaptativos nas diferentes áreas ou atividades humanas" (Hartz 1999).

Segundo Bardin, no método de análise de conteúdo, as deduções são propostas a partir dos resultados significativos e em função dos objetivos da pesquisa (Laurence 2011). Nesta pesquisa, após a mineração de textos com a formação de bigramas, trigramas e quadrigramas mais importantes, utilizou-se uma sequência lógica para o agrupamentos conceitual, em função dos diferentes processos para as ações de enfrentamento da sífilis que foram utilizadas pelos apoiadores. Nos agrupamentos conceituais dos termos, procurou-se identificar as estruturas dos níveis de atenção, ações programáticas de resposta à sífilis nas redes de atenção à saúde e as etapas mais prevalentes, identificadas nos relatórios que foram produzidos.

4.4 Análise do Estado da Arte

Utilizando o método desenvolvido nessa pesquisa, foi possível extrair dos artigos de 2010 a 2021 informações relevantes sobre o que estava sendo pesquisado a cada ano

que estavam associados a PLN, saúde, HIV e Sífilis (critério de seleção para o Estado da Arte, ver capítulo 3). A base de dados para essa análise foi construída através do conteúdo textual de cada artigo selecionado, foram selecionados três artigos para cada ano, constituindo 36 artigos, maior parte deles já foram descritos com detalhes no capítulo 3. A figura 4.2 mostra os 5 bigramas referentes a cada ano. Os gráficos gerados mostram a porcentagem proporcional ao valor do TF-IDF, a barra completamente cheia equivale a 100% e representa o maior valor de TF-IDF entre as palavras.

Figura 4.2: Bigramas de 2010 a 2021
Fonte: Autoria Própria

Todos os trabalhos de 2010 a 2021 utilizados no estado da arte e nesta análise com o algoritmo de mineração de dados desenvolvido nesta pesquisa, usaram alguma técnica de Processamento de Linguagem Natural, como pode ser visto na figura 4.2. Praticamente todos os anos tiveram o bigrama "Processamento Linguagem" ou "Linguagem Natural" entre os 5 (cinco) bigramas mais relevantes extraídos dos artigos. Com exceção dos anos 2010 e 2019, que tiveram entre os 5 bigramas "Machine Learning", "Sistema Informação", "Coleta Dados" e "Análise Sentimento". A análise de sentimentos é uma técnica de Processamento de Linguagem Natural aplicada a muitos textos, inclusive textos de mídias sociais para identificação, extração, quantificação e estudo de elementos subjetivos e sentimentos como o trabalho de (Luo et al. 2019) que envolve análise de sentimentos em postagens do Twitter, já citado anteriormente. Em 2010, os trabalhos abordados ((Wright et al. 2010), (Mao et al. 2010) e (Mathur & Dinakarpandian 2010)) tratavam de técnicas de PLN aplicados a Sistemas de Informação em Saúde, Coleta de Dados ou Resultados de Laboratório, os bigramas se destacaram devido a isso.

Em 2011 e 2013 os trabalhos que se destacaram foram de PLN aplicados ao prontuário eletrônico ((Ohno-Machado 2011), (Kohane 2011), (Xu et al. 2011) e (Liu et al. 2013)), principalmente o trabalho de (Liu et al. 2013) que abordavam um framework de extração de informações para identificar coortes usando dados de prontuário eletrônico de saúde que foi desenvolvido sob a Unstructured Information Management Architecture (UIMA). Para avaliar o framework os autores implementaram dois algoritmos de PLN e demonstraram que esse framework pode ser usado para extração de informações clínicas específicas. O ano de 2012 destacou o bigrama "Nódulos Pulmonares" devido ao trabalho de (Danforth et al. 2012) que desenvolveu um método para identificar pacientes com nódulos pulmonares, os autores combinaram cinco códigos de diagnóstico, quatro de procedimentos e uma técnica de PLN que pesquisava em transcrições de radiologia feitas em texto livre.

Os anos 2014, 2016 e 2017 abordaram os bigramas "Saúde Mental" e "Sintomas Psiquiátricos", os trabalhos foram mais voltados para esses temas, os autores (Coppersmith et al. 2014) desenvolveram uma análise de fenômenos de saúde mental com dados do Twitter utilizando métodos de PLN para fornecer informações sobre distúrbios específicos extraindo evidências de sinais linguísticos de saúde mental de forma rápida, os autores (Cook et al. 2016) usaram PLN e AM para prever pensamentos suicidas e sintomas psiquiátricos aumentados entre adultos liberados de salas de emergência e internações psiquiátricas. As variáveis alvo foram pensamentos suicidas e sintomas psiquiátricos e variáveis preditoras incluíram itens estruturados, por exemplo, relacionados aos sono e bem estar e os autores (Calvo et al. 2017) revisaram como as mídias sociais e outras fontes de dados foram utilizadas para detectar emoções e identificar pessoas que podem precisar de assistência psicológica com as técnicas computacionais de PLN.

O ano de 2021 tiveram mais trabalhos envolvendo PLN, Sifilis e Covid19, já 2020 os termos mais destacados foram relacionados a "Conteúdo Visual" e "Saúde Pública". O termo "Conteúdo Visual" foi devido ao trabalho de (Nobles et al. 2020), mencionado anteriormente. Os anos de 2015 e 2018 evidenciaram os termos "Mídia Social" e "Termos Médicos". Em 2015 os trabalhos foram voltados a PLN e extração de dados em Mídias Sociais (Rastegar-Mojarad et al. 2015) e (Nikfarjam et al. 2015), e em 2018 o termo se

destacou devido ao trabalho de (Chen et al. 2018) que implementaram métodos computacionais para supervisão à distância adaptado para priorizar termos médicos importantes para compreender o registro eletrônico de saúde.

Foram extraídos também os bigramas de toda a base de artigos, como mostrado na figura 4.3

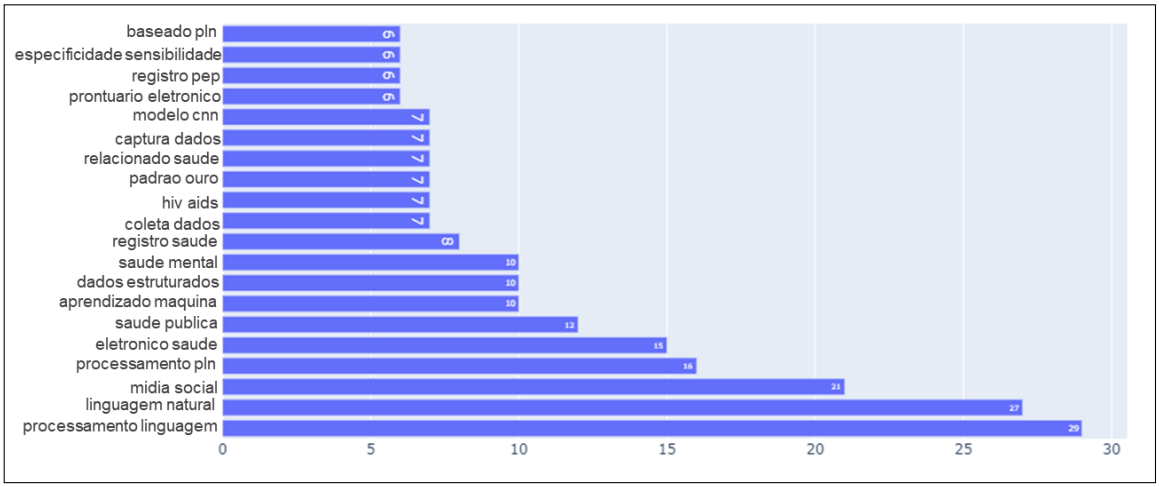

Figura 4.3: Bigramas dos artigos
Fonte: Autoria Própria

Os bigramas “processamento linguagem”, “linguagem natural”, “processamento pln” e “baseado em pln” indicam que a busca de artigos foi totalmente voltada para PLN. Alguns artigos envolviam também AM, os bigramas “aprendizado máquina”, “modelo cnn” e “especificidade sensibilidade” são termos mais comuns em aprendizado de máquina. A maior parte dos bigramas são termos de saúde, pois a busca de artigos envolveu essa temática.

Capítulo 5

O Caso da Epidemia de Sífilis no Brasil: Resultados, análises e discussões sobre intervenções nos municípios prioritários

Neste capítulo estão os resultados obtidos nesta tese, utilizando os métodos abordados no capítulo anterior.

Dos 4874 arquivos inseridos pelos apoiadores na "Plataforma LUES" no período de 2018-2020, observa-se que a maior quantidade são do tipo relatório, o que totaliza 1.019, como demonstrado pela Figura 5.1. Outras produções foram inseridas no sistema tais como imagem (14,42%), resumo (12,84%), informe técnico (6,38%), comunicados em geral (4,76%), programação de eventos ou reuniões (3,97%), lista de presença (1,34%) e outros tipos de arquivos de texto (artigos, boletim epidemiológico e etc.).

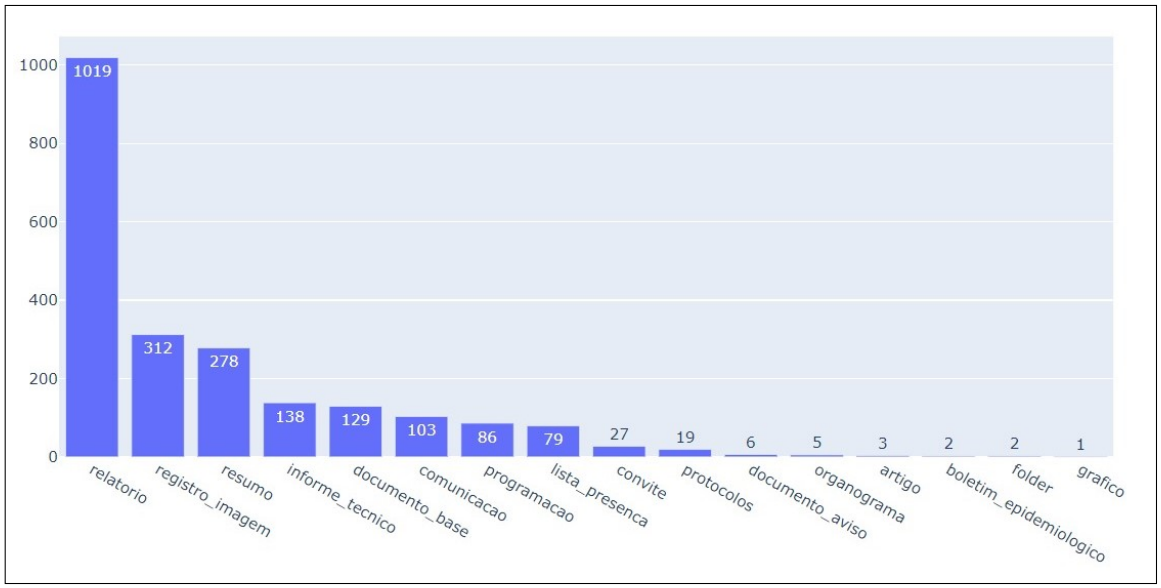

Figura 5.1: Tipos de arquivos da Plataforma LUES
Fonte: Autoria Própria

Os relatórios descrevem as ações de intervenção dos apoiadores nos municípios e, devido a sua importância, foram escolhidos como base ou corpus para fins de aplicação do método de análise de conteúdo e mineração de textos no presente estudo. Observa-se, conforme a Figura 5.2, que no ano de 2018 foram produzidos 430 relatórios; no ano 2019, 355; e no ano de 2020, 234 relatórios em documentos de texto.

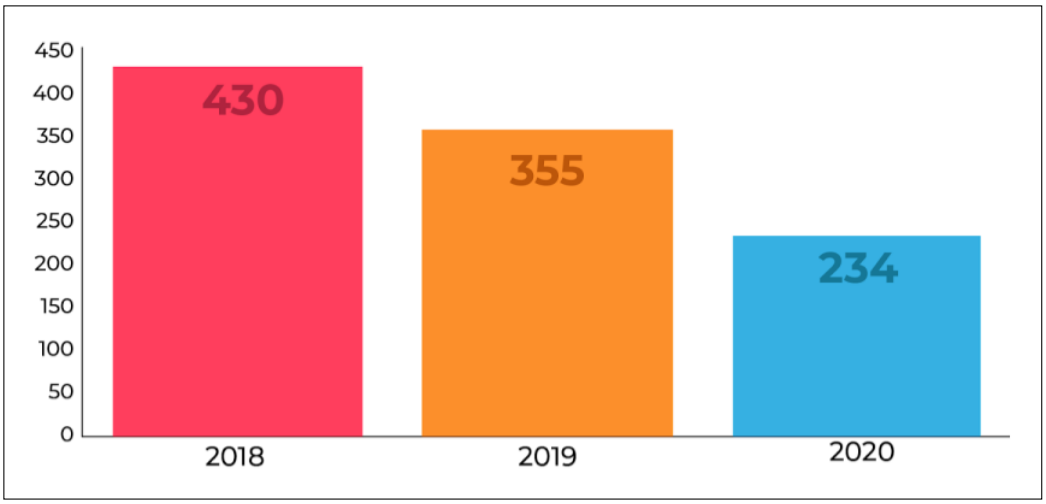

Figura 5.2: Quantidade de relatórios produzidos anualmente no país
Fonte: Autoria Própria

A Figura 5.3 apresenta o mapa do Brasil com a distribuição geográfica dos apoiadores, o total de relatórios produzidos durante os três anos de atuação dos apoiadores no projeto, e o tamanho da população por região. A região com mais apoiadores é a Sudeste totalizando 22, também é a região que mais produziu relatórios (629), sendo a maior região, em termos populacionais do país, com 80.364.410 de pessoas, conforme o último censo realizado (Censo 2000).

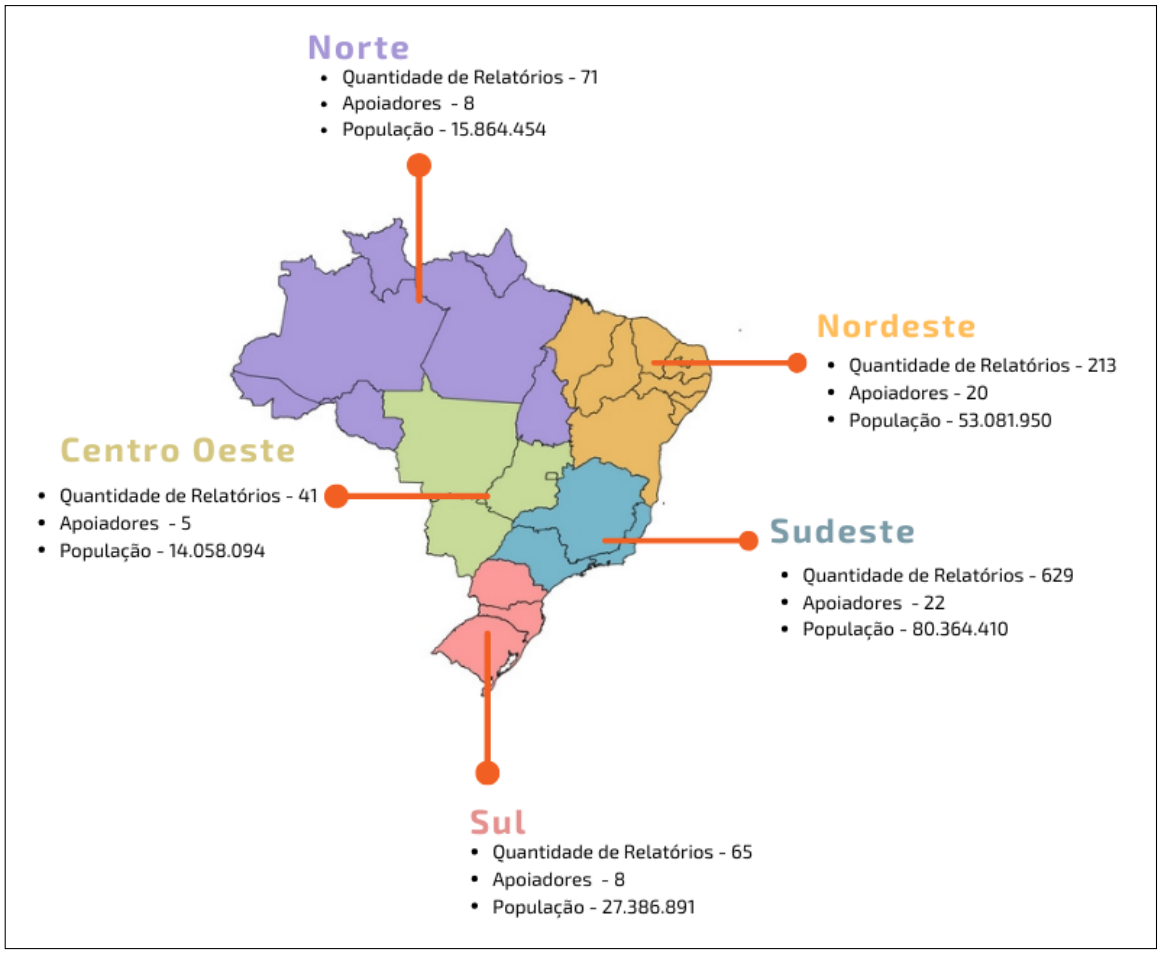

Figura 5.3: Mapa do Brasil com a distribuição geográfica dos apoiadores
Fonte: Autoria Própria

A segunda região que mais produziu relatórios foi a Nordeste, com um total de 213. Também é a segunda maior em relação ao número de apoiadores (20) e em número populacional (53.081.950). A região Sul produziu 65 relatórios; e a Região Centro-Oeste, 41.

A proporção de relatórios em documentos de texto válidos por região pode ser visto na Figura 5.4, a região sudeste teve uma maior proporção de relatórios produzidos devido a quantidade de apoiadores nessa região ser em maior número.

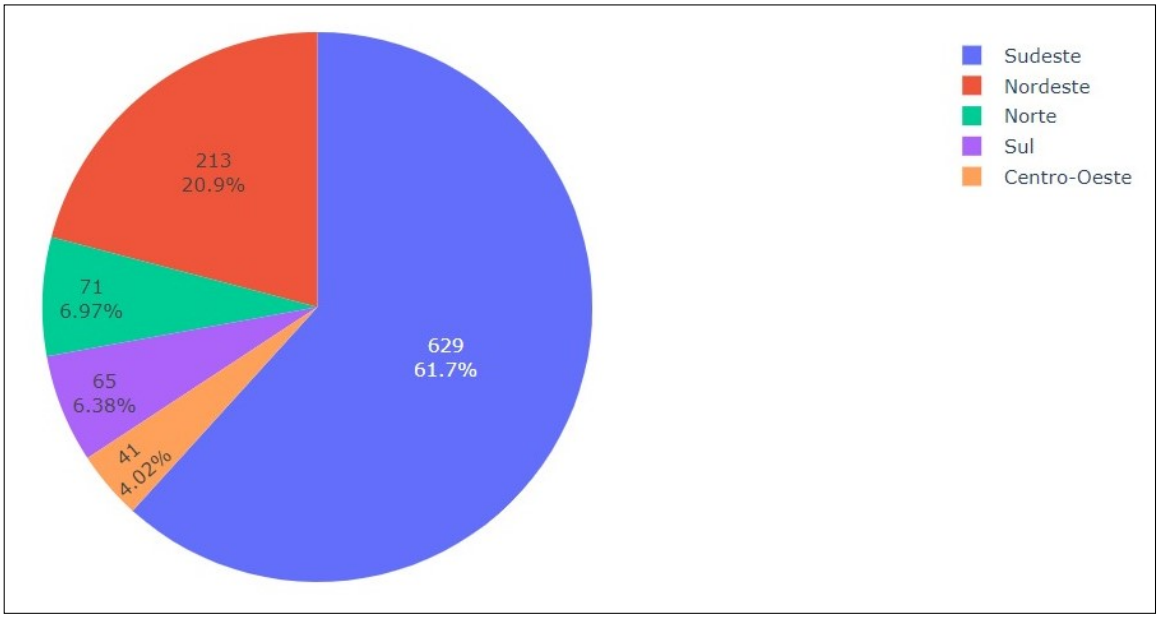

Figura 5.4: Proporção de relatórios por região
Fonte: Autoria Própria

A quantidade de relatórios em documentos de texto produzidas por mês e região do país está demonstrado na Figura 5.5, onde, pode-se observar que os meses de julho/2018 e julho/2020, na região sudeste, foram os de maior produção dos relatórios, e na Figura 5.6 está a quantidade de relatórios válidos do país produzidos mensalmente, os meses de maior produção foram junho/2018, maio/2020 e julho/2020.

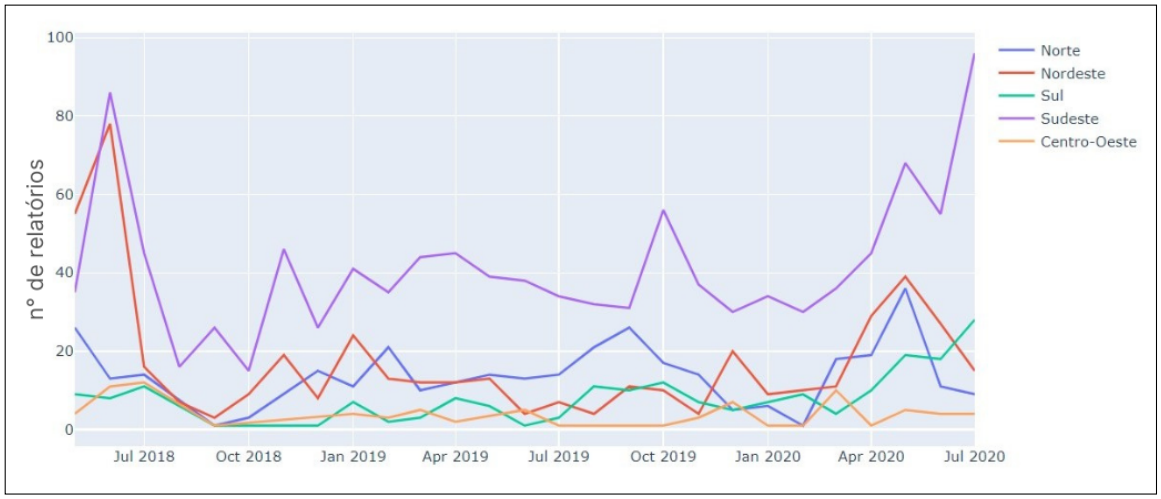

Figura 5.5: Quantidade de relatórios produzidos mensalmente por região
Fonte: Autoria Própria

Figura 5.6: Quantidade de relatórios produzidos mensalmente no país
Fonte: Autoria Própria

5.1 Análise dos Bigramas

A extração das palavras foi realizada para explorar o que de relevante foi desenvolvido pelos apoiadores dentro do projeto. Por isso existe a necessidade de uma investigação de cada palavra extraída da base de documentos de texto.

A análise de conteúdo com base nos bigramas permitiu verificar quais os termos do projeto "Sífilis Não!" foram mais presentes nos relatórios dos apoiadores. O termo "sífilis congênita" foi o de maior frequência, sinalizando que este agravo foi o de maior importância nos relatórios de trabalho, em detrimento da "sífilis em gestante", que teve uma frequência bem menor, e da "sífilis adquirida", que não apareceu entre os 20 bigramas mais presentes nos relatórios (Figura 5.7).

O termo "atenção básica" foi o segundo mais frequente nos bigramas de documentos de texto, sinalizando o componente do sistema de saúde que foi mais relevante para os apoiadores no que concerne ao "enfrentamento da sífilis" (terceiro termo de maior frequência), sendo o "pré natal", o serviço de saúde de maior relevância, a "saúde da mulher" a área programática mais importante e a "gestante" o público alvo mais mencionado. Dentre os demais termos, o "transmissão vertical", o "investigação transmissão", o "casos sífilis" sugerem a relevância do trabalho de vigilância epidemiológica, para a resposta à sífilis congênita.

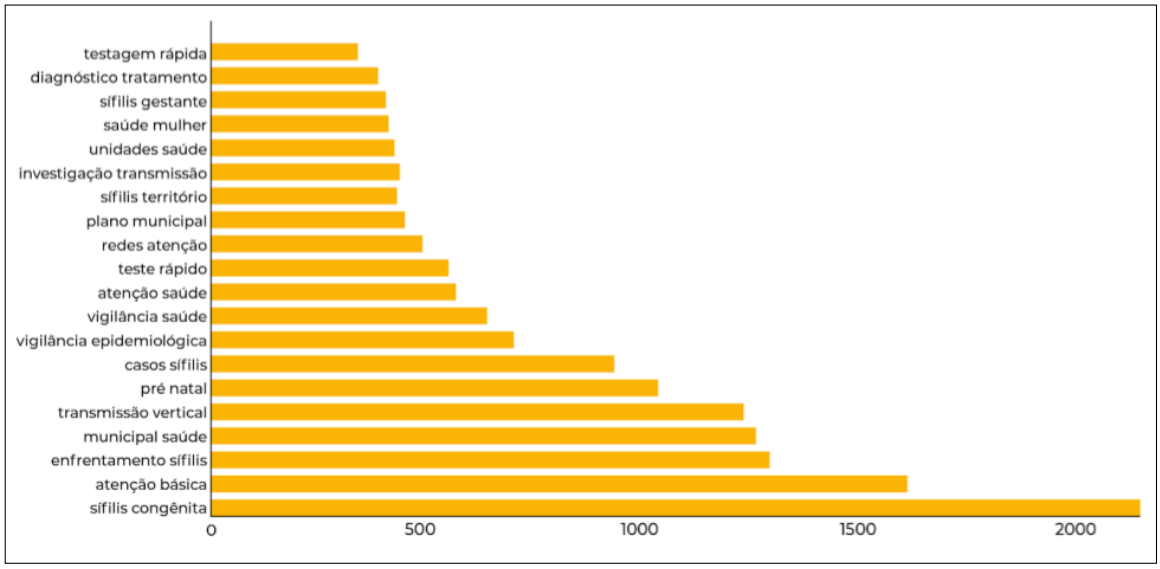

Figura 5.7: Os 20 bigramas mais presentes nos relatórios de documentos de texto
Fonte: Autoria Própria

A análise dos bigramas permitiu o agrupamento dos mesmos em categorias exploratórias relacionadas à resposta à sífilis congênita e frequência destes termos nas redes de atenção, fazendo referência ao nome original do projeto "Sífilis Não!", a Figura 5.8 apresenta a distribuição das categorias exploratórias por ano e região.

Os termos conceituais foram formados combinando os bigramas, trigramas e quadrigramas, utilizando análise de conteúdo para formar sentenças com informação interpre-

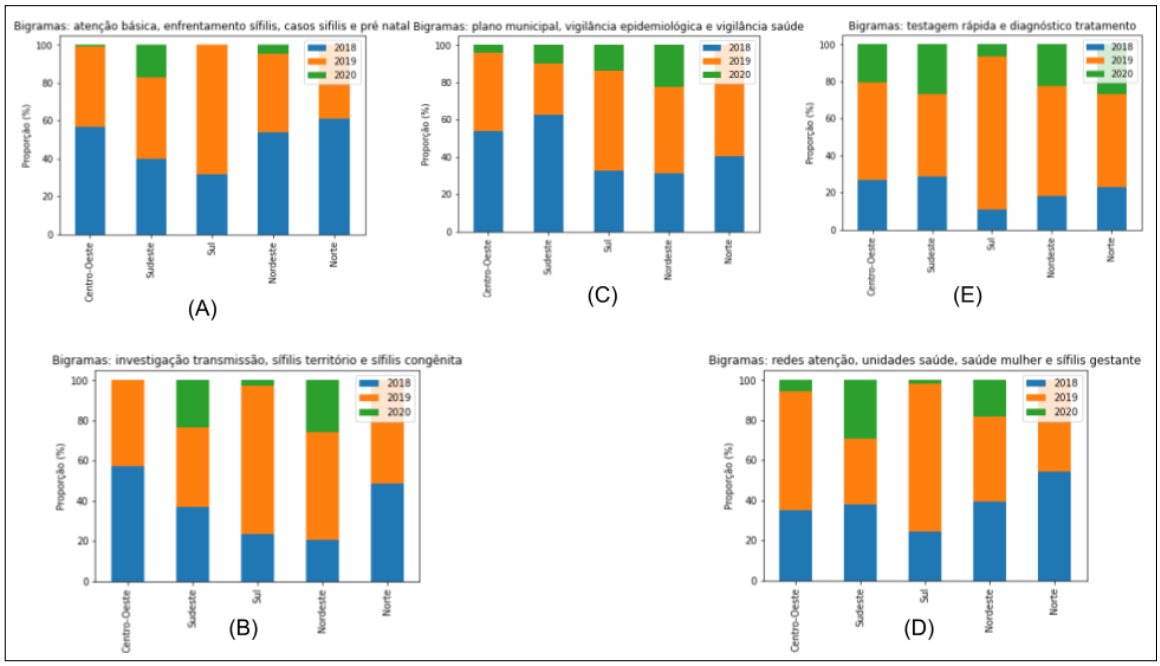

Figura 5.8: Proporção dos termos relevantes separados por região
Fonte: Autoria Própria

tável. São considerados termos conceituais a partir dos bigramas: "Enfrentamento da sífilis no pré-natal da atenção primária em saúde" (Figura 5.8A); "investigação da transmissão de sífilis congênita no território" (Figura 5.8B); "vigilância epidemiológica no plano municipal de saúde" (Figura 5.8C); "sífilis em gestantes na atenção à saúde da mulher"(Figura 5.8D); "testagem rápida, diagnóstico e tratamento"(Figura 5.8E).

O "Enfrentamento da sífilis no pré-natal da atenção primária em saúde" (Figura 5.8A) demonstra que as maiores atividades ocorreram em 2018 na maior parte das regiões, exceto para a região Sul e Sudeste. Para o "investigação da transmissão de sífilis congênita no território" (Figura 5.8B), as regiões Centro-oeste e Norte apresentaram proporções elevadas no início do projeto. O "vigilância epidemiológica no plano municipal de saúde"(Figura 5.8C) obteve sua melhor distribuição no ano de 2018 nas regiões Centro-oeste e Sudeste. Para as "sífilis em gestantes de atenção à saúde da mulher"(Figura 5.8D), a região sul apresentou a maior atividade no ano de 2019. Por fim, a maior proporção da "testagem rápida, diagnóstico e tratamento"(Figura 5.8E) foi no ano de 2019 em todas as regiões, com destaque para a região Sul.

Ainda na análise de bigramas, foram extraídos os dez bigramas mais presentes nos relatórios de texto por região do país, Figura 5.9.

5.2 Análises dos Trigramas

Os 20 trigramas de documentos de texto mais relevantes extraídos na análise foram destacados na Figura 5.10, com maior relevância para "investigação transmissão verti-

Figura 5.9: Os 10 bigramas mais presentes nos relatórios por região
Fonte: Autoria Própria

cal", "casos sífilis congênita", "transmissão vertical sífilis" e "sífilis redes atenção". Estes trigramas apontam que a resposta à transmissão vertical da sífilis congênita, com investigação desta transmissão nas redes de atenção foi considerada prioritária nos relatórios.

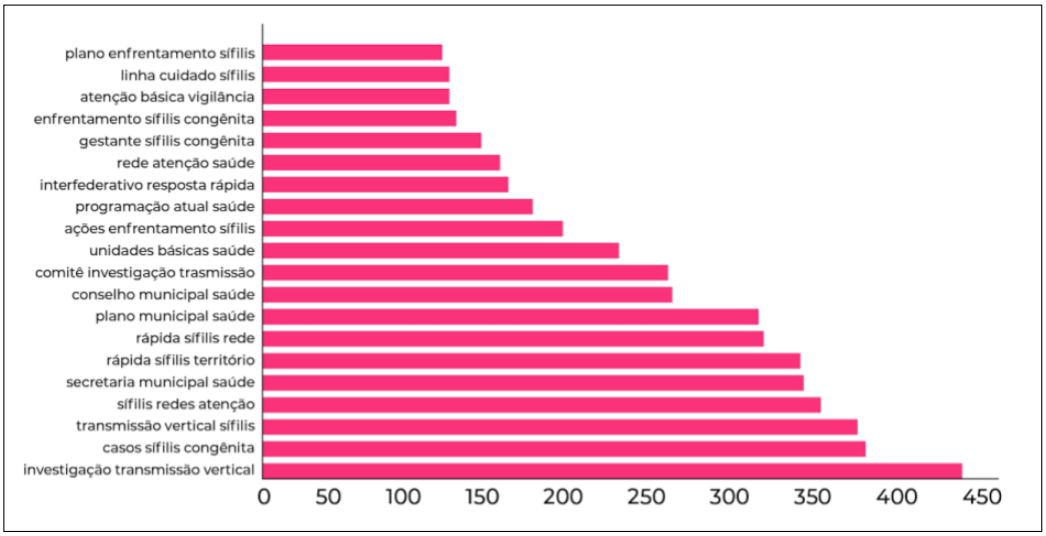

Figura 5.10: Os 20 trigramas mais presentes nos relatórios de documentos de texto
Fonte: Autoria Própria

São considerados termos conceituais a partir dos trigramas: "Comitê de investigação da transmissão vertical da sífilis congênita" (Figura 5.11A), "sífilis congênita no planejamento e programação anual de saúde dos municípios" (Figura 5.11B) e "vigilância e atenção nas linhas de cuidado da sífilis congênita e da sífilis em gestante nas unidades básicas das redes de atenção à saúde" (Figura 5.11C).

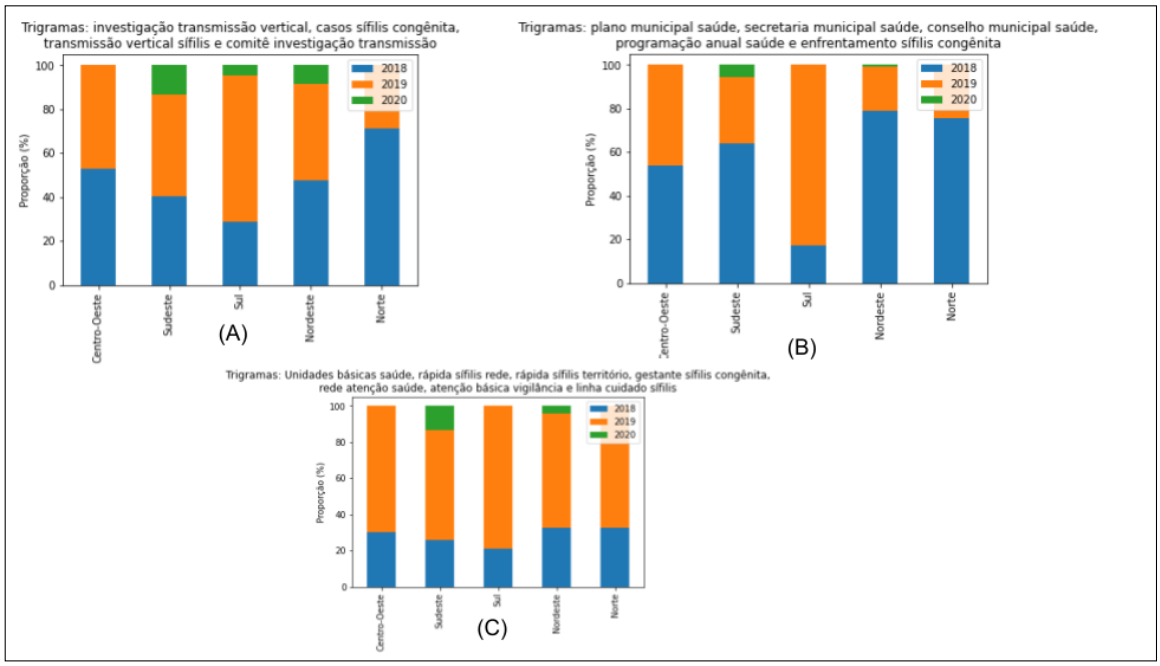

Figura 5.11: Proporção dos trigramas relevantes separados por região
Fonte: Autoria Própria

A análise com base nos agrupamentos de trigramas confirma a tendência do trabalho dos apoiadores no âmbito da "investigação da transmissão vertical" da "sífilis congênita" (Figura 5.11A, 5.11B e 5.11C), permanece as "gestantes" como a população alvo mais importante (Figura 5.11C), e aparece o "plano municipal de saúde" como instrumento relevante, agregando a ele a "programação anual de saúde" e os "conselhos municipais de saúde".

Ainda na Figura 5.11, as maiores atividades para o "Comitê de investigação da transmissão vertical da sífilis congênita" (Figura 5.11A) e para a "sífilis congênita no planejamento e programação anual de saúde dos municípios" (Figura 5.11B) ocorreram em 2018 para quase todas as regiões. Por fim, o termo conceitual "vigilância e atenção nas linhas de cuidado da sífilis congênita e da sífilis em gestante nas unidades básicas das redes de atenção à saúde" (Figura 5.11C) teve sua maior atividade no ano de 2019 em todas as regiões.

Os trigramas também foram extraídos por região do Brasil, como mostrado na Figura 5.12.

Figura 5.12: Os 8 trigramas mais presentes nos relatórios por região
Fonte: Autoria Própria

Nas regiões Centro-Oeste e Norte, o termo "Resposta rápida sífilis" foi o mais relevante extraído. Na região Nordeste, o trigrama mais relevante foi "transmissão vertical sífilis" e o menos relevante dentre os dez foi "plano municipal saúde".

5.2.1 Análise dos Quadrigramas

Os quadrigramas extraídos podem ser visualizados na Figura 5.13, com maior frequência para parte do nome original do projeto "Sífilis Não!" ("projeto resposta rápida sífilis").

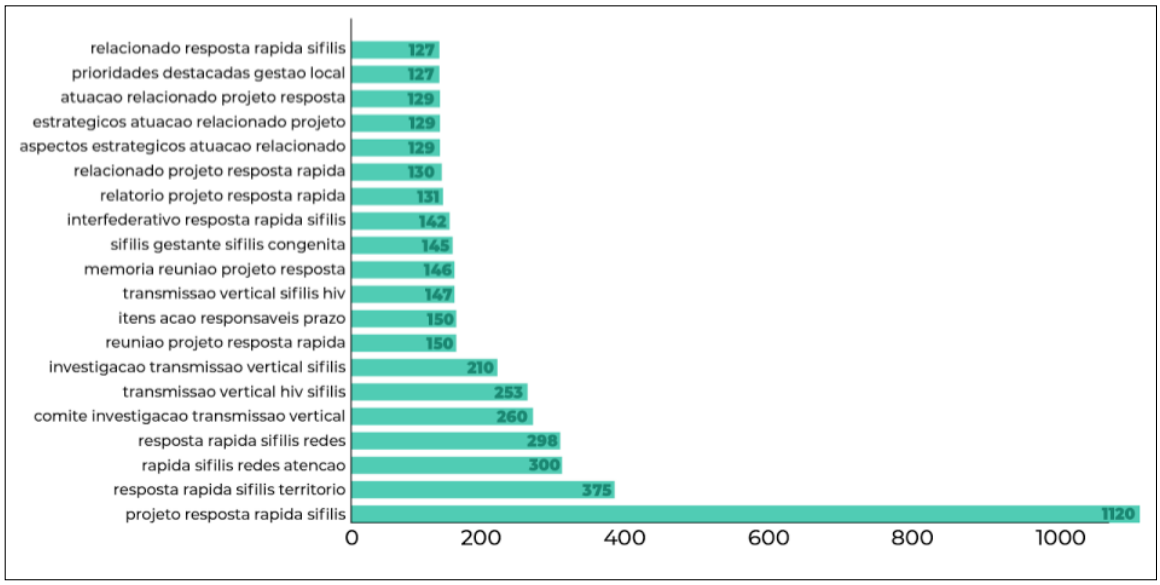

Figura 5.13: Os 20 quadrigramas mais presentes nos relatórios dos documentos de texto
Fonte: Autoria Própria

Os demais quadrigramas confirmam a priorização dos apoiadores quanto a investigação da transmissão vertical e o comitê como meio ou serviço importante para a resposta à sífilis congênita no território. A partir destes foram gerados quatro termos conceituais, distribuídos por ano e região, conforme a Figura 5.14.

Na Figura 5.14A, prevaleceu o conceito de "implantação do projeto de resposta rápida através de ações estratégias do apoiador de pesquisa e intervenção". Esse termo obteve sua maior distribuição em todas as regiões no ano de 2019. O segundo termo conceitual gerado foi o "projeto de resposta rápida à sífilis conduz avanços nas redes de atenção à saúde na sífilis congênita", com distribuição predominante em quase todas as regiões no início do projeto "Sífilis Não!" (Figura 5.14B).

Na Figura 5.14C, o termo conceitual foi "prioridades da gestão local para o combate à sífilis em gestante e congênita", com sua distribuição quase totalitária no ano de 2019 em quase todas as regiões. Por último, o termo conceitual "comitê para investigação da transmissão vertical conduzido pela resposta rápida à sífilis" teve sua maior distribuição no início do projeto em quase todas as regiões (Figura 5.14D). Verificou-se que os quadrigramas destacam ainda mais as possibilidades de correlações entre as ações dos

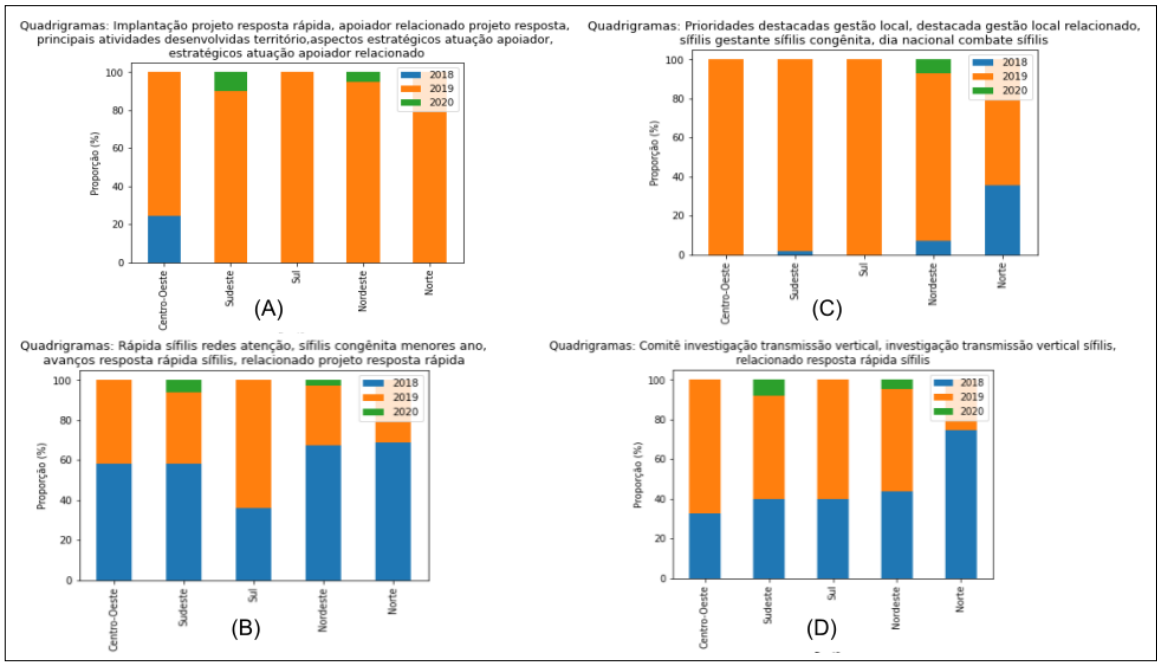

Figura 5.14: Proporção dos quadrigramas relevantes separados por região
Fonte: Autoria Própria

apoiadores quanto à sífilis congênita e os eixos de Governança e Gestão, Vigilância e Atenção Integral, do projeto "Sífilis Não!".

5.3 Inter-relação entre bigramas, trigramas e quadrigramas

A Figura 5.15 apresenta a inter-relação entre os termos conceituais que foram formados à luz dos objetivos do projeto "Sífilis Não!" para os bigramas, trigramas e quadrigramas. A interpretação dos mesmos estabelece pelo menos cinco categorias analíticas consideradas representativas das conexões entre o trabalho de intervenção do apoiador de pesquisa e intervenção e os objetivos do projeto "Sífilis Não!":

1. Enfrentamento da sífilis durante o pré-natal na atenção básica;
2. Comitê de investigação de sífilis congênita no território;
3. Plano municipal para monitoramento e investigação dos casos de sífilis a partir da vigilância em saúde;
4. Redes de atenção à saúde da mulher para sífilis em gestantes;
5. Diagnóstico e tratamento com ênfase na testagem rápida.

Assim, pode-se inferir que o trabalho dos apoiadores do projeto "Sífilis Não!" foi prioritariamente de intervenção para o enfrentamento da resposta à sífilis congênita no território trabalhado, seguindo a linha de eliminação da sífilis congênita pelos municípios.

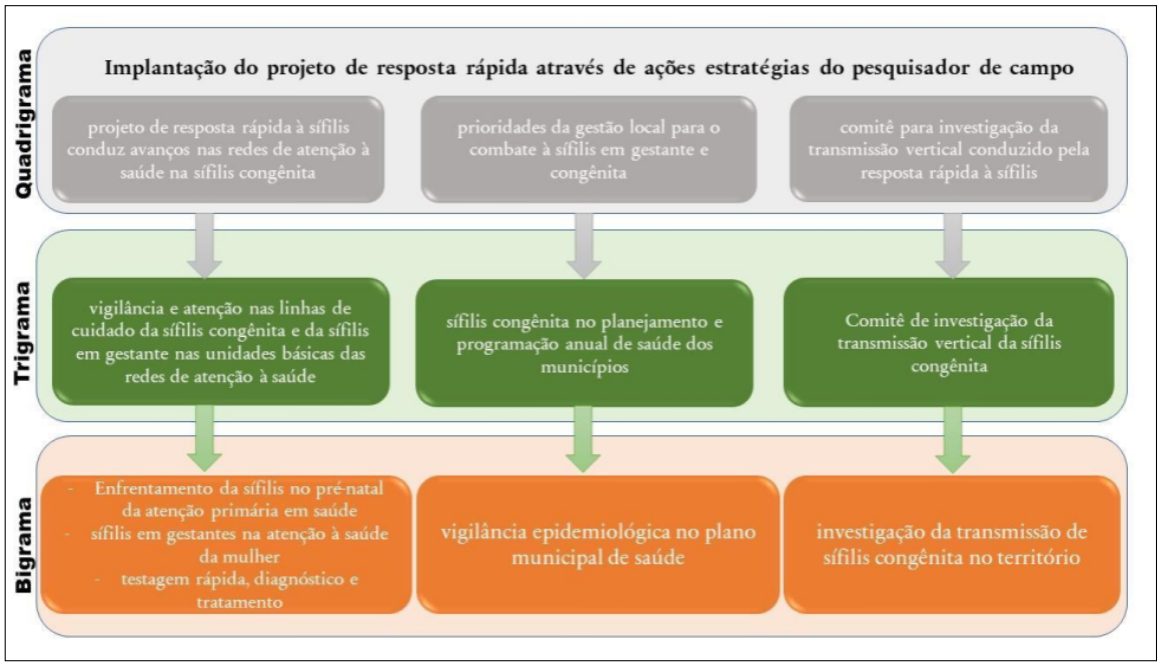

Figura 5.15: Inter-relações dos termos conceituais gerados a partir da interpretação dos bigramas, trigramas e quadrigramas.
Fonte: Autoria Própria

5.4 Discussão

A pesquisa demonstrou que a utilização da técnica de mineração de textos identificou padrões relacionados às ações ou intervenções do trabalho dos apoiadores em relação à sífilis nos municípios prioritários, e possibilitou medir e comparar a atuação dos pesquisadores por região do país. Foram determinadas 6 categorias analíticas que demonstram conexões importantes e representativas do trabalho de articulação do apoiador de pesquisa e intervenção e os objetivos do projeto "Sífilis Não!". São elas:

1. Enfrentamento da sífilis durante o pré-natal na atenção básica;
2. Comitê de investigação de sífilis congênita no território;
3. Plano municipal para monitoramento e investigação dos casos de sífilis a partir da vigilância em saúde;
4. Redes de atenção à saúde da mulher para sífilis em gestantes;
5. Diagnóstico e tratamento com ênfase na testagem rápida;
6. Fortalecimento de ações para enfrentamento da sífilis.

Cabe mencionar que a atuação dos apoiadores no projeto "Sífilis Não!" surgiu com a finalidade de intervir e auxiliar as equipes técnicas e de gestores locais, uma das principais ações era a redução das altas taxas de sífilis congênita em municípios prioritários (BRASIL 2018). No Brasil, no período de 2010 a 2017, a taxa de sífilis congênita foi de 2 casos por 1.000 nascidos vivos para 8,8 por 1.000 nascidos vivos, respectivamente, um

incremento de 338%. Nesse período, houve uma tendência significativa de crescimento da sífilis congênita, com AAPC de 15,75%, sem perspectiva de mudança de tendência (Marques dos Santos et al. 2020).

No ano de 2018, a taxa de sífilis congênita atingiu o seu maior valor histórico, como pode ser visto na Figura 5.16, chegando a surpreendentes 9 casos por 1000 nascidos vivos e 26441 casos no total. Porém, após o início do projeto "Sífilis Não!", as taxas de sífilis congênita reduziram para 8,2 casos por 1000 nascidos vivos, segundo o último boletim epidemiológico (da S. 2020). Com isso, houve uma mudança de tendência de crescimento de 15,75% para uma redução de -8,8% na taxa anual de incidência de sífilis congênita, com perspectiva de queda para os próximos anos, podendo chegar na meta de redução de 90% estabelecida pela OPAS em quase quatro anos (OPAS 2019). A redução das taxas de sífilis congênita a nível nacional, em um intervalo curto de tempo, pode ser creditada ao modelo de intervenção do projeto "Sífilis Não!", cujo impacto epidemiológico se mostrou efetivo, principalmente a partir da identificação de eixos temáticos de intervenção presentes nos relatórios dos apoiadores.

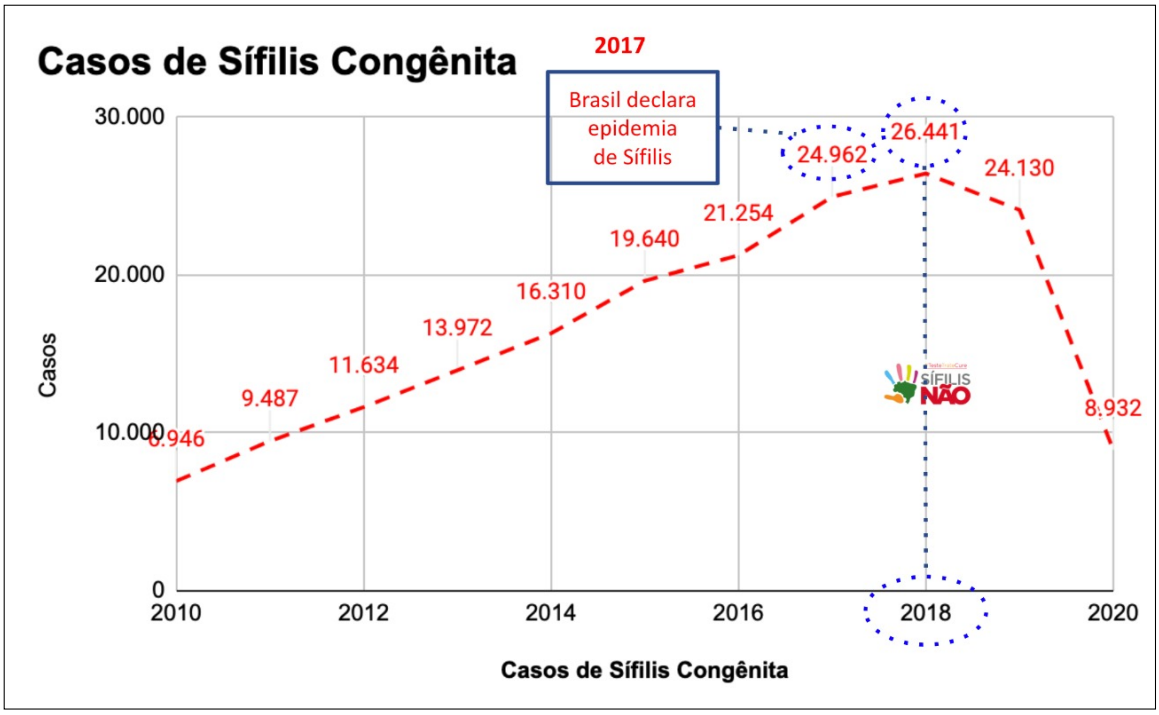

Figura 5.16: Casos de Sífilis Congênita entre os anos 2010 e 2020
Fonte: Autoria Própria

A partir dos resultados, foi possível identificar que houve uma tendência à agrupamentos de termos dirigidos às ações programáticas relacionadas à sífilis congênita nos municípios prioritários. Os trigramas e quadrigramas consolidaram essa tendência observada nos bigramas e avançaram com informações relacionadas ao trabalho dos apoiadores por eixos do projeto "Sífilis Não!", principalmente no eixo "gestão e governança" e "educação comunicação".

O Framework da Organização Panamericana da Saúde (OPAS) para a eliminação da Transmissão Materno Infantil das ISTs na Região das Américas tem como um dos eixos do marco conceitual a integração das políticas direcionadas à mulher e a criança com os serviços de saúde correspondentes, desde as etapas de prevenção que são prévios à gestação, até a saúde da mulher e da criança, após a etapa de atenção perinatal e ao parto (OPAS 2019). Os resultados desta pesquisa evidenciaram que a integração entre a vigilância e a atenção está presente no trabalho dos apoiadores, corroborado pela presença dos comitês de investigação da transmissão vertical, mostrando que a atividade de resposta à sífilis no Brasil se utiliza de uma estratégia que não está prevista no ETMI Plus da OPAS, portanto, inovadora. Os dados mostram que a principal intervenção dos apoiadores foi no trabalho dos comitês e indica a importância estratégica desta política de vigilância e atenção integral para eliminação da sífilis congênita.

O sistema de saúde brasileiro possui protocolos específicos para investigar e intervir nos casos identificados de sífilis congênita. Quem faz esse trabalho são os comitês de investigação da transmissão vertical, ou grupos de especialistas multiprofissionais, instituídos com a finalidade de mapear os problemas relacionados à transmissão vertical e propor soluções a partir de um protocolo de investigação pré-estabelecido para reduzir a sífilis congênita. Os comitês se integram ao SUS como espaços de atuação técnica, sigilosa, não-coercitiva ou punitiva, podendo estar vinculados à gestão municipal ou estadual de saúde (da Saúde 2014), (Ministério da Saúde & de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis 2019).

Os resultados demonstraram que esses comitês de investigação tiveram uma atuação forte no projeto "Sífilis Não!", citado principalmente em 2018, com destaque para a região Norte. Além disso, a investigação da transmissão vertical demonstrado no bigrama teve sua distribuição distinta, apresentando as maiores proporções para as regiões nordeste e sudeste, mantendo quase constante essas atividades no período de 2018-2020.

A constatação dessas diferenças por regiões podem ser atribuídas à própria complexidade organizacional das redes de atenção à saúde brasileira (Oliveira et al. 2017), cuja implementação dos comitês de investigação da transmissão vertical dependerá de cada governo local de maneira descentralizada. É provável que as regiões com maior fragilidade de consolidação desses comitês tenham exigido mais atenção por parte dos apoiadores no início do projeto, como foi constatado na região Norte (Garnelo et al. 2017), o que pode ter influenciado na efetividade do combate à sífilis congênita naquela região.

A mineração dos relatórios das ações de intervenção dos apoiadores também destacou as ações de enfrentamento da sífilis no pré-natal, da atenção primária em saúde e da sífilis em gestantes na atenção à saúde da mulher. No Brasil, dentre as experiências do Ministério da Saúde que são consideradas específicas de prevenção da transmissão vertical que foram anteriores ao projeto "Sífilis Não!", destacavam-se o Projeto Nascer Maternidades

(de 2002) e o Plano Operacional para Redução do HIV e da Sífilis (de 2007).

O projeto Nascer Maternidades foi direcionado ao aproveitamento do momento do parto para garantir o conhecimento do status sorológico da mulher para o HIV (Santos et al. 2010) e o Plano Operacional para Redução do HIV e da Sífilis visava ampliar a cobertura de testagem tanto do HIV, como da sífilis, no pré-natal e garantir ações pactuadas desde os anos 1990 pelo Programa Nacional de HIV/Aids com os estados e municípios. Tais ações estavam bastante relacionadas à melhoria dos indicadores do HIV/Aids, tendo como principal estratégia a aquisição e distribuição de insumos de medicamentos antirretrovirais, fórmula infantil para aleitamento de crianças expostas ao HIV e a determinação de rotinas de manejo do HIV durante a gestação, parto e pós-parto (da Saúde & de Vigilância em Saúde 2007).

A indução dessas políticas era realizada através da compra dos principais insumos e da prestação de contas sobre a sua utilização pelos três entes federativos do SUS. Observa-se que tais estratégias se de um lado avançaram com a prevenção da transmissão vertical do HIV, do outro não conseguiram conter o avanço da epidemia de sífilis no país, conforme os indicadores epidemiológicos anteriores ao projeto "Sífilis Não!" (da S. 2020).

O projeto "Sífilis Não!" parece ter avançado em relação aos seus antecessores, ao trazer uma proposta de indução de política pública de saúde de maneira integrada e cooperativa entre a vigilância e atenção à sífilis no território, sendo o apoiador de campo um elemento articulador fundamental. Os dados apontam para a possibilidade de que estes apoiadores tenham conseguido alinhar ações e intervenções de projetos relativos à saúde da mulher que já existiam nos locais de intervenção, com os objetivos do projeto "Sífilis Não!", pois nesta pesquisa, verificou-se que as intervenções de maior relevância foram observadas em relação ao público alvo das gestantes nas redes de atenção.

No que se refere a distribuição das intervenções dos apoiadores por região, destaca-se que as ações de prevenção e assistência foram menos relatadas na região Sul, no início do Projeto, mas que aumentaram no ano de 2019. Cabe mencionar que a região Sul possui as maiores tendências de crescimento de sífilis congênita no Brasil (Marques dos Santos et al. 2020), e esse fato pode ter contribuído para uma maior resistência dos gestores locais no início da implementação do Projeto. Pode-se inferir que os apoiadores da região Sul tiveram maior dificuldade, comparado às demais regiões, para sua inserção no território.

A testagem rápida para sífilis foi outro ponto da assistência em saúde que obteve destaque na produção dos apoiadores. Estudo desenvolvido por Santos e colaboradores (Santos et al. 2021), com a finalidade de identificar fatores que influenciaram o combate à sífilis, destacaram que a testagem rápida tem um papel fundamental para combater a sífilis na atenção primária em saúde. Além disso, estudo desenvolvido por Roncalli e colaboradores demonstrou que o período de 2018-2019 houve a maior disponibilidade de teste rápido para combate à sífilis congênita em toda série histórica, independente de outros fatores de confusão (Roncalli et al. 2021). Dessa forma, pode-se aventar que as ações intervencionistas dos apoiadores contribuíram para o fortalecimento das medidas de *Point of care rapid test* (Organization 2021), (Angel-Müller et al. 2018) para prevenção da sífilis congênita, o que se manteve quase constante na produção dos relatórios na maioria das regiões no período de 2018-2020.

As ações intervencionistas dos apoiadores do projeto "Sífilis Não!" podem ter influen-

ciado também na redução das internações da sífilis congênita em todo país. Foi verificado que as taxas de internações para tratamento da sífilis congênita tiveram uma redução significativa no período de maio de 2018 a dezembro de 2019 (de Andrade et al. 2020). Esses dados corroboram com nossos resultados, cujas ações de assistência e vigilância foram presentes na produção dos relatórios de campo, bem como na formação de bigramas e trigramas.

Capítulo 6

Considerações Finais

Este capítulo tem como objetivo expor as considerações finais, mencionando as conclusões obtidas com a aplicação do método desenvolvido, como também os trabalhos futuros. Os resultados obtidos são motivadores, levando em conta as dificuldades para o problema em questão.

6.1 Conclusões

O estudo comprovou que a mineração de textos, ao ser integrada ao tradicional método de análise de conteúdo, é capaz de atender objetos de pesquisa de saúde pública que contemplem grandes volumes de dados. Este método computacional permitiu extrair do Projeto "Sífilis Não", tanto as ações de intervenção dos pesquisadores de campo, como também subsidiou as inferências sobre como as estratégias do Projeto podem ter incidido na redução de casos de sífilis congênita nos municípios prioritários.

A mineração de textos dos relatórios apontou que houve uma articulação entre as ações dos pesquisadores de campo quanto à sífilis congênita em três, dos quatro eixos do Projeto "Sífilis Não" (Governança e Gestão; Vigilância e Atenção Integral). A pesquisa confirmou a hipótese de que o trabalho do apoiador de pesquisa e intervenção demonstra os nexos entre os objetivos do Projeto e o enfrentamento da sífilis no território, ao demonstrar por meio de seis categorias analíticas como se deu a atuação prioritária dos apoiadores do projeto, além de sua contribuição com os indicadores de diminuição de sífilis congênita no Brasil.

Assim, as ações dos pesquisadores de campo destacados nos relatórios podem ter sido um indutor de políticas públicas de saúde no território ao contribuírem para manter as ações de prevenção e promoção à saúde no combate à sífilis congênita como prioridade pelos gestores de saúde locais.

6.2 Contribuições

A principal contribuição deste trabalho é o método computacional utilizando mineração de dados para extrair informações de grandes quantidades de textos de forma automatizada, além de algumas contribuições secundárias:

- Base de dados consolidada de acesso aberto com informações anonimizadas;
- Artigo A1: "The Text Mining Technique Applied to the Analysis of Health Interventions to Combat Congenital Syphilis in Brazil: The Case of the "Syphilis No!" Project", publicado na Frontiers in Public Health seção Digital Health;
- Artigo A1: "A text as unique as a fingerprint: Text analysis and authorship recognition in a Virtual Learning Environment of the Unified Health System in Brazil", publicado na Expert Systems with Applications.

6.3 Limitações

Algumas limitações podem ser identificadas na construção desta tese. Por exemplo, a formação de bigramas, trigramas e quadrigramas foram mineradas a partir de um algoritmo que destacou as ações mais importantes dos apoiadores de pesquisa e intervenção. Isso pode excluir ações que foram desenvolvidas com a gestão local e que podem ter provocado algum impacto significativo para as mudanças dos processos de trabalho para prevenção da sífilis, a exemplo de palavras relacionadas ao eixo de educação e comunicação do Projeto. Mesmo assim, a pesquisa conseguiu identificar quais esforços foram mais importantes a ponto de contribuir com as mudanças atuais dos indicadores epidemiológicos de sífilis no Brasil. Existe também a limitação do algoritmo de não extrair o contexto em que essas palavras estão ou extrair os tópicos em que esses termos se relacionam que poderiam ser interessantes e facilitar na análise de conteúdo.

6.4 Trabalhos Futuros

Para as próximas etapas, já em desenvolvimento, será feita a inclusão da produção textual de relatórios a análise de LDA para extração de tópicos conceituais e agrupamento dos N-gramas. Com isso, os tópicos gerados pelo LDA poderão dar mais consistência na análise de conteúdo dos N-gramas. Além disso, os tópicos gerados serão analisados a partir dos dados do território de sífilis, por região. Assim, será possível identificar quais tópicos da produção dos apoiadores foram mais significativos na redução da sífilis ou na melhoria de indicadores de produção. Outra etapa importante é avaliar a relação dos tópicos do LDA com os questionários de autoavaliação dos apoiadores, avaliação dos supervisores e gestores e identificar quais ações dos apoiadores de pesquisa e intervenção modificaram também outros indicadores importantes para o combate à sífilis.

Referências Bibliográficas

aaa (n.d.).

Aggarwal, Charu C (2015), Data classification, *em* 'Data mining', Springer, pp. 285–344.

Agrawal, Rakesh, Ramakrishnan Srikant et al. (1994), Fast algorithms for mining association rules, *em* 'Proc. 20th int. conf. very large data bases, VLDB', Vol. 1215, Citeseer, pp. 487–499.

Aho, AV, R Sethi & JD Ullman (1985), 'Compilers: Principles, techniques, and tools'.

Alshaikh, Fahdah, Farzan Ramzan, Salman Rawaf & Azeem Majeed (2014), 'Social network sites as a mode to collect health data: a systematic review', *Journal of medical Internet research* **16**(7), e171.

Amaral-Rosa, Marcelo Prado (2019), 'Considerations on the use of iramuteq software for qualitative data analysis', *Revista da Escola de Enfermagem da USP* **53**.

Angel-Müller, Edith, Carlos F Grillo-Ardila, Jairo Amaya-Guio, Nicolas A Torres-Montañez & Luisa F Vasquez-Velez (2018), 'Point of care rapid test for diagnosis of syphilis infection in men and nonpregnant women', *The Cochrane Database of Systematic Reviews* **2018**(5).

AVASUS (2020), 'Base de dados do cadastro nacional de estabelecimentos de saude - profissionais do sus, ministério da saúde, br.'.

BRASIL, M da S (2018), 'Seminário apresenta projeto "resposta rápida à sífilis nas redes de atenção" a profissionais de saúde [internet]. 2018 p. 1'.

Cai, Tianrun, Luwan Zhang, Nicole Yang, Kanako K Kumamaru, Frank J Rybicki, Tianxi Cai & Katherine P Liao (2019), 'Extraction of emr numerical data: an efficient and generalizable tool to extend clinical research', *BMC medical informatics and decision making* **19**(1), 1–7.

Calvo, Rafael A, David N Milne, M Sazzad Hussain & Helen Christensen (2017), 'Natural language processing in mental health applications using non-clinical texts', *Natural Language Engineering* **23**(5), 649–685.

Castro, Victor M, Dmitriy Dligach, Sean Finan, Sheng Yu, Anil Can, Muhammad Abd-El-Barr, Vivian Gainer, Nancy A Shadick, Shawn Murphy, Tianxi Cai et al. (2017), 'Large-scale identification of patients with cerebral aneurysms using natural language processing', *Neurology* **88**(2), 164–168.

Cavnar, William B, John M Trenkle et al. (1994), N-gram-based text categorization, *em* 'Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval', Vol. 161175, Citeseer.

Censo, IBGE (2000), 'Instituto brasileiro de geografia e estatística. 2010', *Sinopse por Setores. Disponível em:< https://censo2010. ibge. gov. br/sinopseporsetores> .*

Chen, Jinying, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, Hong Yu et al. (2018), 'A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews', *Journal of medical Internet research* **20**(1), e8669.

Cook, Benjamin L, Ana M Progovac, Pei Chen, Brian Mullin, Sherry Hou & Enrique Baca-Garcia (2016), 'Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid', *Computational and mathematical methods in medicine* **2016**.

Coppersmith, Glen, Mark Dredze & Craig Harman (2014), Quantifying mental health signals in twitter, *em* 'Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality', pp. 51–60.

Crosby Jr, Alfred W (1969), 'The early history of syphilis: a reappraisal', *American Anthropologist* **71**(2), 218–227.

Da Rocha, Marcella A, Marquiony M Dos Santos, Raphael S Fontes, Andréa SP de Melo, Aliete Cunha-Oliveira, Angélica E Miranda, Carlos AP de Oliveira, Hugo Gonçalo Oliveira, Cristine MG Gusmão, Thaísa GFMS Lima et al. (2022), 'The text mining technique applied to the analysis of health interventions to combat congenital syphilis in brazil: The case of the "syphilis no!" project', *Frontiers in Public Health* **10**.

da Rocha, Marcella Andrade, Philippi Sedir Grilo de Morais, Daniele Montenegro da Silva Barros, João Paulo Queiroz dos Santos, Ricardo Alexsandro de Medeiros Valentim et al. (2022), 'A text as unique as a fingerprint: Text analysis and authorship recognition in a virtual learning environment of the unified health system in brazil', *Expert Systems with Applications* p. 117280.

da Rocha, Marcella Andrade, Roberto Douglas da Costa, Ricardo Alexsandro de Medeiros Valentim & Aline de Pinho Dias (2019), 'Um texto tão singular quanto a impressão digital: reconhecimento de autoria com um olhar para o avasus', *Brazilian Journal of Development* **5**(12), 32960–32973.

da S., Brasil M (2020), 'Boletim epidemiológico de sífilis. secr vigilância em saúde [internet]. 2020'.

da Saúde, Ministério (2014), 'Protocolo de infestigação de transmissão vertical'.

da Saúde, Ministério & Secretaria de Vigilância em Saúde (2007), 'Programa nacional de dst/aids. plano operacional. redução da transmissão vertical do hiv e da sífilis'.

Danforth, Kim N, Megan I Early, Sharon Ngan, Anne E Kosco, Chengyi Zheng & Michael K Gould (2012), 'Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing', *Journal of Thoracic Oncology* **7**(8), 1257–1262.

de Andrade, Ion Garcia Mascarenhas, Ricardo Alexsandro de Medeiros Valentim & Carlos Alberto Pereira de Oliveira (2020), 'The influence of the no syphilis project on congenital syphilis admissions between 2018 and 2019', *DST j. bras. doenças sex. transm* pp. 1–6.

de Morais, Philippi Sedir Grilo, Rodrigo Dantas da Silva, José Arilton Pereira Filho, Ricardo Alexsandro de Medeiros Valentim, Karilany Dantas Coutinho, Carlos Alberto Pereira de Oliveira, Azim Roussanaly & Anne Boyer (2020), Strategies for content recommendation in the brazilian rapid response to syphilis project, *em* 'Proceedings of the 10th Euro-American Conference on Telematics and Information Systems', pp. 1–6.

Duran, Benjamin S & Patrick L Odell (2013), *Cluster analysis: a survey*, Vol. 100, Springer Science & Business Media.

Džeroski, Sašo (2009), Relational data mining, *em* 'Data mining and knowledge discovery handbook', Springer, pp. 887–911.

El, Sara El Manar & Ismail Kassou (2014), 'Authorship analysis studies: A survey', *International Journal of Computer Applications* **86**(12).

Feldman, Ronen, James Sanger et al. (2007), *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge university press.

Feller, Daniel J, Jason Zucker, Michael T Yin, Peter Gordon & Noémie Elhadad (2018), 'Using clinical notes and natural language processing for automated hiv risk assessment', *Journal of acquired immune deficiency syndromes (1999)* **77**(2), 160.

Fleuren, Wilco WM & Wynand Alkema (2015), 'Application of text mining in the biomedical domain', *Methods* **74**, 97–106.

Forrai, Judit (2011), 'History of different therapeutics of venereal disease before the discovery of penicillin', *Syphilis-Recognition, Description and Diagnosis, In Tech* pp. 37–58.

Garnelo, Luiza, Amandia Braga Lima Sousa & Clayton de Oliveira da Silva (2017), 'Health regionalization in amazonas: progress and challenges', *Ciencia & saude coletiva* **22**, 1225–1234.

Gartner, Gideon (2022), 'Gartner, inc. [internet]. 2022', urlhttps://www.gartner.com/en.

Gianfrancesco, Milena A, Suzanne Tamang, Jinoos Yazdany & Gabriela Schmajuk (2018), 'Potential biases in machine learning algorithms using electronic health record data', *JAMA internal medicine* **178**(11), 1544–1547.

Hart, Peter E, David G Stork & Richard O Duda (2000), *Pattern classification*, Wiley Hoboken.

Hartz, Zulmira Maria de Araújo (1999), 'Avaliação dos programas de saúde: perspectivas teórico metodológicas e políticas institucionais', *Ciência & saúde coletiva* **4**, 341–353.

IDC (2022), 'International data corporation [internet]. 2022', urlhttps://www.idc.com/.

Jadhav, Deepali Kishor (2013), 'Big data: the new challenges in data mining', *International Journal of Innovative Research in Computer Science & Technology* **1**(2), 39–42.

Jo, Taeho (2006), The implementation of dynamic document organization using the integration of text clustering and text categorization, Tese de doutorado, University of Ottawa (Canada).

Jo, Taeho (2019), 'Text mining', *Studies in Big Data* .

Jo, Taeho & Malrey Lee (2007), The evaluation measure of text clustering for the variable number of clusters, *em* 'International Symposium on Neural Networks', Springer, pp. 871–879.

Kar, Arpan Kumar & Yogesh K Dwivedi (2020), 'Theory building with big data-driven research–moving away from the "what" towards the "why"', *International Journal of Information Management* **54**, 102205.

Kohane, Isaac S (2011), 'Using electronic health records to drive discovery in disease genomics', *Nature Reviews Genetics* **12**(6), 417–428.

Kojima, Noah & Jeffrey D Klausner (2018), 'An update on the global epidemiology of syphilis', *Current epidemiology reports* **5**(1), 24–38.

Korenromp, Eline L, Jane Rowley, Monica Alonso, Maeve B Mello, N Saman Wijesooriya, S Guy Mahiané, Naoko Ishikawa, Linh-Vi Le, Morkor Newman-Owiredu, Nico Nagelkerke et al. (2019), 'Global burden of maternal and congenital syphilis and associated adverse birth outcomes—estimates for 2016 and progress since 2012', *PloS one* **14**(2), e0211720.

Kowalski, Gerald J & Mark T Maybury (2000), *Information storage and retrieval systems: theory and implementation*, Vol. 8, Springer Science & Business Media.

Kumar, Sunil, Arpan Kumar Kar & P Vigneswara Ilavarasan (2021), 'Applications of text mining in services management: A systematic literature review', *International Journal of Information Management Data Insights* **1**(1), 100008.

LAIS/UFRN (2022), 'Plataforma lues [internet]. 2022', urlhttps://lues.lais.ufrn.br/login.

Laurence, BARDIN (2011), 'Análise de conteúdo', *São Paulo: Edições* **70**, 276.

Lee, Hyeyong, Rie Shimotakahara, Akimi Fukada, Sumiko Shinbashi & Shigemitsu Ogata (2019), 'Impact of differences in clinical training methods on generic skills development of nursing students: A text mining analysis study', *Heliyon* **5**(3), e01285.

Li, Yuelin, Bruce Rapkin, Thomas M Atkinson, Elizabeth Schofield & Bernard H Bochner (2019), 'Leveraging latent dirichlet allocation in processing free-text personal goals among patients undergoing bladder cancer surgery', *Quality of Life Research* **28**(6), 1441–1455.

Liu, Hongfang, Suzette J Bielinski, Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, KE Ravikumar, Stephen T Wu, Iftikhar J Kullo & Christopher G Chute (2013), 'An information extraction framework for cohort identification using electronic health records', *AMIA Summits on Translational Science Proceedings* **2013**, 149.

Luhn, Hans Peter (1957), 'A statistical approach to mechanized encoding and searching of literary information', *IBM Journal of research and development* **1**(4), 309–317.

Luo, Xiao, Gregory Zimet & Setu Shah (2019), 'A natural language processing framework to analyse the opinions on hpv vaccination reflected in twitter over 10 years (2008-2017)', *Human vaccines & immunotherapeutics* **15**(7-8), 1496–1504.

Macedo, Alessandra Alaniz, Juliana Tarossi Pollettini, José Augusto Baranauskas & Julia Carmona Almeida Chaves (2016), 'A health surveillance software framework to deliver information on preventive healthcare strategies', *Journal of biomedical informatics* **62**, 159–170.

Machekera, Shepherd, Peniel Boas, Poruan Temu, Zimmbodilion Mosende, Namarola Lote, Angela Kelly-Hanku, S Guy Mahiane, Robert Glaubius, Jane Rowley, Anup Gurung et al. (2021), 'Strategic options for syphilis control in papua new guinea–impact and cost-effectiveness projections using the syphilis interventions towards elimination (site) model', *Infectious Disease Modelling* **6**, 584–597.

Mao, Yurong, Zunyou Wu, Katharine Poundstone, Changhe Wang, Qianqian Qin, Ye Ma & Wei Ma (2010), 'Development of a unified web-based national hiv/aids information system in china', *International journal of epidemiology* **39**(suppl_2), ii79–ii89.

Marques dos Santos, Marquiony, Ana Karla Bezerra Lopes, Angelo Giuseppe Roncalli & Kenio Costa de Lima (2020), 'Trends of syphilis in brazil: a growth portrait of the treponemic epidemic', *Plos one* **15**(4), e0231029.

Mathur, Sachin & Deendayal Dinakarpandian (2010), 'Automated ontological gene annotation for computing disease similarity', *Summit on translational bioinformatics* **2010**, 12.

Ministério da Saúde, Secretaria de Vigilância em Saúde & Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis (2019), 'Protocolo clínico e diretrizes terapêuticas para atenção integral às pessoas com infecções sexualmente transmissíveis (ist)'.

Ministério da Saúde, Brasil. (2021), 'Boletim epidemiológico de sífilis. secr. vigilância em saúde [internet]. 2021'.

Moral, Cristian, Angélica de Antonio, Ricardo Imbert & Jaime Ramírez (2014), 'A survey of stemming algorithms in information retrieval.', *Information Research: An International Electronic Journal* **19**(1), n1.

Mullen, Tony & Nigel Collier (2004), Sentiment analysis using support vector machines with diverse information sources, *em* 'Proceedings of the 2004 conference on empirical methods in natural language processing', pp. 412–418.

Nikfarjam, Azadeh, Abeed Sarker, Karen O'connor, Rachel Ginn & Graciela Gonzalez (2015), 'Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features', *Journal of the American Medical Informatics Association* **22**(3), 671–681.

Nobles, Alicia L, Eric C Leas, Carl A Latkin, Mark Dredze, Steffanie A Strathdee & John W Ayers (2020), '# hiv: alignment of hiv-related visual content on instagram with public health priorities in the us', *AIDS and Behavior* **24**(7), 2045–2053.

Nóbrega, Giovani Ângelo Silva da, Gustavo Fontoura de Souza, Jânio Gustavo Barbosa, Karilany Dantas Coutinho & Ricardo Alexsandro de Medeiros Valentim (2016), Uma análise estatística do ambiente virtual de aprendizagem do sistema único de saúde: Descrição estatística dos usuários do avasus, *em* 'Uma análise estatística do ambiente virtual de aprendizagem do Sistema Único de Saúde: descrição estatística dos usuários do AVASUS', pp. 24–24.

Ohno-Machado, Lucila (2011), 'Realizing the full potential of electronic health records: the role of natural language processing', *Journal of the American Medical Informatics Association* **18**(5), 539–539.

Oliveira, Ana Paula Cavalcante de, Mariana Gabriel, Mario Roberto Dal Poz & Gilles Dussault (2017), 'Challenges for ensuring availability and accessibility to health care services under brazil's unified health system (sus)', *Ciencia & saude coletiva* **22**, 1165–1180.

Olvera-Lobo, María-Dolores & Juncal Gutiérrez-Artacho (2011), 'Open-vs. restricted-domain qa systems in the biomedical field', *Journal of Information Science* **37**(2), 152–162.

OPAS (2019), 'Politica para um enfoque integrado e sustentável visando as doenças transmissiveis nas americas [internet]. 2019'.

Organization, World Health (2021), 'World health organization. point-of-care diagnostic tests (pocts) for sexually transmitted infections (stis) [internet]. 2021'.

Pereira, Gilberto Carvalho, Ricardo Coutinho & Nelson Francisco Favila Ebecken (2008), 'Data mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of arraial do cabo', *Brazilian Journal of Oceanography* **56**, 1–12.

Pesaranghader, Ahmad, Ali Pesaranghader & Azadeh Rezaei (2013), Applying latent semantic analysis to optimize second-order co-occurrence vectors for semantic relatedness measurement, *em* 'Mining Intelligence and Knowledge Exploration', Springer, pp. 588–599.

Piai, Silvia & Massimiliano Claps (2013), 'Bigger data for better healthcare', *IDC Health Insights* **8**, 1–24.

Priamo, Vania, Sofia Campos dos Santos & Jamile Soares dos Santos (2020), 'Pistas para o trabalho do apoio no projeto "sífilis não"', *Revista Brasileira de Inovação Tecnológica em Saúde-ISSN: 2236-1103* **10**(4), 11–11.

Rastegar-Mojarad, Majid, Zhan Ye, Daniel Wall, Narayana Murali, Simon Lin et al. (2015), 'Collecting and analyzing patient experiences of health care from social media', *JMIR research protocols* **4**(3), e3433.

Rocha, Marcella A da, Giovani Ângelo S da Nóbrega, Ricardo A de Medeiros Valentim & Luca Pareja CF Alves (2020), A text as unique as fingerprint: Avasus text analysis and authorship recognition, *em* 'Proceedings of the 10th Euro-American Conference on Telematics and Information Systems', pp. 1–8.

Rocha, Marcella A da, Marquiony M Dos Santos & Ricardo A de Medeiros Valentim (2021), Automação usando deep learning para avaliação de risco de sífilis adquirida nos municípios brasileiros, *em* 'XIII Congresso da Sociedade Brasileira de DST - IX Congresso Brasileiro de AIDS - IV Congresso Latino Americano de IST/HIV/AIDS'.

Rocha, Marcella Andrade da (2019), Um texto tão singular quanto a impressão digital: o uso de sistemas inteligentes para reconhecimento de autoria, Dissertação de mestrado, Brasil.

Roncalli, Angelo Giuseppe, Tatyana Maria Silva de Souza Rosendo, Marquiony Marques dos Santos, Ana Karla Bezerra Lopes & Kenio Costa de Lima (2021), 'Efeito da cobertura de testes rápidos na atenção básica sobre a sífilis em gestantes no brasil', *Revista de Saúde Pública* **55**.

Rothschild, Bruce M (2005), 'History of syphilis', *Clinical Infectious Diseases* **40**(10), 1454–1463.

Saif, Hassan, Miriam Fernández, Yulan He & Harith Alani (2014), 'On stopwords, filtering and data sparsity for sentiment analysis of twitter'.

Salton, Gerard (1989), 'Automatic text processing: The transformation, analysis, and retrieval of', *Reading: Addison-Wesley* **169**.

Salton, Gerard & Christopher Buckley (1988), 'Term-weighting approaches in automatic text retrieval', *Information processing & management* **24**(5), 513–523.

Salton, Gerard & Chung-Shu Yang (1973), 'On the specification of term values in automatic indexing', *Journal of documentation* .

Sánchez, David, Montserrat Batet & Alexandre Viejo (2014), 'Utility-preserving privacy protection of textual healthcare documents', *Journal of biomedical informatics* **52**, 189–198.

Santos, Elizabeth Moreira dos, Ana Cristina Reis, Suzanne Westman, Rosane Gomes Alves et al. (2010), 'Avaliação do grau de implantação do programa de controle da transmissão vertical do hiv em maternidades do"projeto nascer"'.

Santos, Marquiony Marques dos, Tatyana Maria Silva de Souza Rosendo, Ana Karla Bezerra Lopes, Angelo Giuseppe Roncalli & Kenio Costa de Lima (2021), 'Weaknesses in primary health care favor the growth of acquired syphilis', *PLoS neglected tropical diseases* **15**(2), e0009085.

Sarbu, I, C Matei, V Benea & SR Georgescu (2014), 'Brief history of syphilis', *Journal of medicine and life* **7**(1), 4.

Sarica, Serhad & Jianxi Luo (2021), 'Stopwords in technical language processing', *Plos one* **16**(8), e0254937.

Ščepanovic, Sanja, Luca Maria Aiello, Ke Zhou, Sagar Joglekar & Daniele Quercia (2020), 'The healthy states of america: Creating a health taxonomy with social media'.

Shafranovich, Yakov (2005), 'Common format and mime type for comma-separated values (csv) files'.

Singh, Jasmeet & Vishal Gupta (2016), 'Text stemming: Approaches, applications, and challenges', *ACM Computing Surveys (CSUR)* **49**(3), 1–46.

Souza, Bruno RG, Ruana TP Vieira, Karilany D Coutinho & Ricardo AM Valentim (2018), Avaliação sobre o nível de satisfação dos usuários inativos com a plataforma avasus, *em* 'Anais da VI Escola Regional de Computação aplicada à Saúde', SBC.

Souza, Marli Aparecida Rocha de, Marilene Loewen Wall, Andrea Cristina de Morais Chaves Thuler, Ingrid Margareth Voth Lowen & Aida Maris Peres (2018), 'The use of iramuteq software for data analysis in qualitative research', *Revista da Escola de Enfermagem da USP* **52**.

Sowmya, R & KR Suneetha (2017), Data mining with big data, *em* '2017 11th International Conference on Intelligent Systems and Control (ISCO)', IEEE, pp. 246–250.

Tan, Pang-Ning, Michael Steinbach & Vipin Kumar (2016), *Introduction to data mining*, Pearson Education India.

Uysal, Alper Kursat & Serkan Gunal (2014), 'The impact of preprocessing on text classification', *Information processing & management* **50**(1), 104–112.

Vespestad, May Kristin & Anne Clancy (2020), 'Exploring the use of content analysis methodology in consumer research', *Journal of Retailing and Consumer Services* p. 102427.

Vieira, Geir Veras, Natanael de Freitas Neto, Karla Mônica Dantas Coutinho, Lidyane Alves da Cunha Laranjeiras, Ricardo Alexsandro de Medeiros Valentim & Karilany Dantas Coutinho (2016), 'Uma metodologia para otimizar o sistema de melhoria continuada do avasus com foco nas experiências do usuário', *Revista Brasileira de Inovação Tecnológica em Saúde-ISSN: 2236-1103* .

Villena, Fabián & Jocelyn Dunstan (2019), 'Obtención automática de palabras clave en textos clínicos: una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en chile', *Revista médica de Chile* **147**(10), 1229–1238.

Wiener, ED (1993), 'A neural network approach to topic spotting in text', *The Thesis of* .

Wilbur, W John & Karl Sirotkin (1992), 'The automatic identification of stop words', *Journal of information science* **18**(1), 45–55.

Wright, Adam, Elizabeth S. Chen & Francine L. Maloney (2010), 'An automated technique for identifying associations between medications, laboratory results and problems', *Journal of Biomedical Informatics* **43**(6), 891–901.
URL: *https://www.sciencedirect.com/science/article/pii/S1532046410001413*

Wu, Xindong, Xingquan Zhu, Gong-Qing Wu & Wei Ding (2013), 'Data mining with big data', *IEEE transactions on knowledge and data engineering* **26**(1), 97–107.

Xia, Fei & Meliha Yetisgen-Yildiz (2012), Clinical corpus annotation: challenges and strategies, *em* 'Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey', p. 67.

Xu, Hua, Min Jiang, Matt Oetjens, Erica A Bowton, Andrea H Ramirez, Janina M Jeff, Melissa A Basford, Jill M Pulley, James D Cowan, Xiaoming Wang et al. (2011), 'Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin', *Journal of the American Medical Informatics Association* **18**(4), 387–391.

Young, Sean D, Wenchao Yu & Wei Wang (2017), 'Toward automating hiv identification: machine learning for rapid identification of hiv-related social media data', *Journal of acquired immune deficiency syndromes (1999)* **74**(Suppl 2), S128.

Șerban, Ovidiu, Nicholas Thapen, Brendan Maginnis, Chris Hankin & Virginia Foot (2019), 'Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification', *Information Processing & Management* **56**(3), 1166–1184.

Apêndice A

O caminho percorrido

Neste apêndice será descrita a trajetória para construção da tese como também, os artigos publicados durante esse percurso.

A.1 Trajetória

Na Figura A.1 está uma linha do tempo com a trajetória e em seguida a descrição dela.

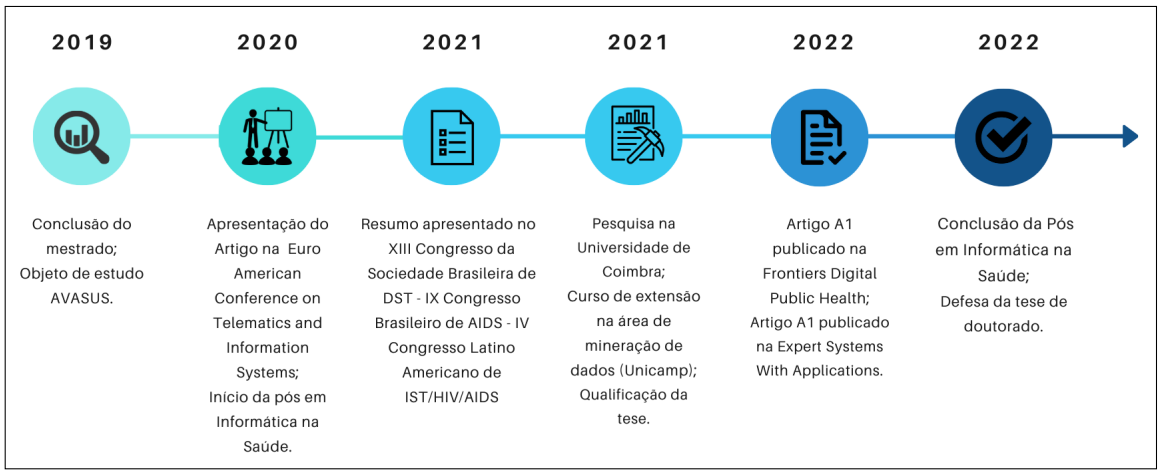

Figura A.1: Linha do tempo da trajetória para conclusão da tese (Autoria Própria)

Uma trajetória iniciada em 2019 com a conclusão do mestrado, que tinha como base científica a área de Processamento de Linguagem Natural (PLN) e o objeto de estudo, de feitio transversal, aplicado inicialmente na Plataforma AVASUS, tecnologia utilizada no Projeto "Sífilis Não" para formação massiva de profissionais de saúde no Brasil (Rocha 2019). Com isso, utilizou-se o conhecimento em PLN para o desenvolvimento de técnicas de mineração de textos para avaliação de políticas públicas de saúde, implementado nesta tese.

Neste contexto, a pesquisa desenvolvida no mestrado e parte no doutorado, teve como fruto uma publicação que faz parte do escopo do Projeto "Sífilis Não", trata-se também dos resultados esperados do projeto. A referida publicação é também um produto da

cooperação internacional com a Universidade de Coimbra e foi aceito no ano de 2022, para publicação na revista Expert Systems With Applications, ISSN: 0957-4174, fator de impacto: 6.954, Qualis A1 na Capes. Esta importante revista, dedica-se a publicar pesquisas em diversas áreas da Ciência da Computação que fazem interfaces com outras áreas, como por exemplo, a área da saúde (da Rocha, de Morais, da Silva Barros, dos Santos, de Medeiros Valentim & others 2022).

Durante o período de curso das disciplinas do doutorado, a disciplina Projeto de Pesquisa II foi cursada e nela foi desenvolvido um artigo que foi apresentado na Euro American Conference on Telematics and Information Systems em novembro de 2020, intitulado: "A text as unique as fingerprint: AVASUS Text Analysis and Authorship Recognition". Essa conferência aconteceu em Portugal, na cidade de Aveiro mas, devido a pandemia covid19 o evento foi online, através de uma videoconferência e o artigo foi desenvolvido dentro da área de PLN (Rocha et al. 2020).

Após concluir as disciplinas do doutorado, o foco foi na continuidade da pesquisa da tese para a qualificação e durante esse período, foi desenvolvido juntamente com um pesquisador da área da saúde, Marquiony Marques dos Santos, um algoritmo para previsão do *Annual Average Percent Change* (AAPC) da sífilis adquirida associado às variáveis sociodemográficas e epidemiológicas. Esse trabalho ainda está em execução e um resumo dele foi apresentado no congresso de DST. O trabalho foi intitulado: "Automação usando Deep Learning para avaliação de risco de sífilis adquirida nos municípios brasileiros" em junho de 2021 no XIII Congresso da Sociedade Brasileira de DST - IX Congresso Brasileiro de AIDS - IV Congresso Latino Americano de IST/HIV/AIDS (Rocha et al. 2021).

Neste mesmo período, ocorreu uma viagem para Portugal para o desenvolvimento de um artigo, como consequência da pesquisa, em parceria com a Universidade de Coimbra, um artigo original que foi fruto dessa parceria e que foi submetido e aprovado na revista Digital Public Health, Qualis A1 em 2022 (Da Rocha, Dos Santos, Fontes, de Melo, Cunha-Oliveira, Miranda, de Oliveira, Oliveira, Gusmão, Lima & others 2022). Após retorno de Portugal, aconteceu a qualificação e a tese ficou intitulada: "Mineração de Texto aplicada às análises de intervenção de Políticas Públicas de Saúde: o caso da epidemia de sífilis no Brasil".

A Universidade de Campinas surgiu com a oportunidade de um curso de extensão na área de mineração de dados, e como esse curso ajudaria no desenvolvimento da tese, foi feita a matrícula e o curso foi concluído em dezembro de 2021. O curso trouxe muito conhecimento relevante para a área de mineração de dados como, análise de dados, recuperação de informação, aprendizado de máquina não supervisionado e supervisionado, visualização de informação, Big Data e Deep Learning.

A aula inaugural do curso de pós-graduação em informática na saúde foi em novembro de 2020 e após um processo seletivo restrito foi conquistada a aprovação. Essa foi mais uma das grandes conquistas e, ao fazer esse curso foi percebido o quanto a computação pode se tornar uma ferramenta de apoio em outras áreas, principalmente na área da saúde. Essa formação ajudou bastante na pesquisa de doutorado que é especialmente relacionada à Sífilis no Brasil. O curso de especialização foi concluído no mês de junho de 2022 e o trabalho de conclusão de curso foi intitulado: "O uso de Processamento de Linguagem Natural para compreender o impacto da rede de apoio no Projeto Sífilis Não", onde foi

feita a análise dos relatórios produzidos pelos apoiadores utilizando um algoritmo de PLN chamado Latent Dirichlet Allocation (LDA).

A.2 Artigos publicados

Citações dos artigos publicados:

- DA ROCHA, Marcella Andrade et al. A text as unique as a fingerprint: Text analysis and authorship recognition in a Virtual Learning Environment of the Unified Health System in Brazil. Expert Systems with Applications, v. 203, p. 117280, 2022.
- DA ROCHA, Marcella A. et al. The Text Mining Technique Applied to the Analysis of Health Interventions to Combat Congenital Syphilis in Brazil: The Case of the "Syphilis No!" Project. Frontiers in Public Health, v. 10, 2022.
- ROCHA, Marcella A. da et al. A text as unique as fingerprint: AVASUS text analysis and authorship recognition. In: Proceedings of the 10th Euro-American Conference on Telematics and Information Systems. 2020. p. 1-8.