Bruno Palmeira de Oliveira

Entendimento da Opinião Pública em Relação às Vacinas da Covid-19 com Base na Mineração das Informações Trocadas por Twitter

Campos dos Goytacazes, RJ

15 de fevereiro de 2023

Bruno Palmeira de Oliveira

Entendimento da Opinião Pública em Relação às Vacinas da Covid-19 com Base na Mineração das Informações Trocadas por Twitter

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação de Prof. Dr. Luis Antonio Rivera Escriba.

Universidade Estadual do Norte Fluminense Darcy Ribeiro – UENF

Centro de Ciência e Tecnologia – CCT

Laboratório de Ciências Matemáticas – LCMAT

Curso de Ciência da Computação

Orientador: Prof. Dr. Luis Antonio Rivera Escriba

Campos dos Goytacazes, RJ

15 de fevereiro de 2023

Bruno Palmeira de Oliveira
Entendimento da Opinião Pública em Relação às Vacinas da Covid-19 com Base na Mineração das Informações Trocadas por Twitter/ Bruno Palmeira de Oliveira. – Campos dos Goytacazes, RJ, 15 de fevereiro de 2023-
57 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Luis Antonio Rivera Escriba

Monografia (Bacharelado) – UENF-CCT-LCMAT-Ciência da Computação, 15 de fevereiro de 2023.

CDU 004.41 : 004.4'2 :

Bruno Palmeira de Oliveira

Entendimento da Opinião Pública em Relação às Vacinas da Covid-19 com Base na Mineração das Informações Trocadas por Twitter

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação de Prof. Dr. Luis Antonio Rivera Escriba.

Trabalho aprovado. Campos dos Goytacazes, RJ, 15 de fevereiro de 2023:

Prof. Dr. Luis Antonio Rivera Escriba
Orientador

Profa. Dr. Fermín Alfredo Tang Montané
Membro da Banca

Prof. Dr. João Luiz de Almeida Filho
Membro da Banca

Campos dos Goytacazes, RJ 15 de fevereiro de 2023

Este trabalho é dedicado a Deus que sempre me deu forças para chegar até aqui e minha família que me apoiou, em especial a minha mãe que sempre me fez acreditar que era possível.

Agradecimentos

Primeiramente, eu agradeço a Deus por ter me dado saúde, força e sabedoria para superar as dificuldades. Foram muitas noites mal dormidas e de esforço para conciliar os trabalhos que me permitiram permanecer no espaço acadêmico. Agradeço a Deus por sempre estar ao meu lado e permitir que minha dedicação me conduzisse até este momento.

Agradeço também a minha família, que sempre me apoiou e acreditou que a educação poderia transformar não só a minha vida, mas a de todos à minha volta. Principalmente a minha mãe que por muitas madrugadas enxugou meu choro de cansaço e sempre me fez acreditar que eu era capaz, quando muitas vezes pensei em desistir. Ao meu pai e minha irmã por serem para mim sinônimos de força e resiliência.

Agradeço também aos meus amigos, que fizeram essa jornada se tornar mais leve e desempenharam um papel significativo para o meu crescimento, não apenas acadêmico, mas também contribuíram para o desenvolvimento da minha visão sócio cultural. Principalmente no que diz respeito à compreensão do lugar que meu corpo negro ocupa na soceidade e como eu poderia transfomá-lo em potência. A esses amigos, minha eterna gratidão.

Por último e não menos importante, agradeço aos professores por me proporcionarem o conhecimento, não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional. Agradeço ao professor Luis Antonio Rivera Escriba pelos feedbacks sinceros, parceria e os direcionamentos nos momentos difíceis. Ao professor Fermín Alfredo Tang Montané pelo seu olhar humano e senso de justiça para com os alunos. Obrigado por todas as palavras positivas. A professora Annabell Del Real por ser essa professora amiga, que sonha os sonhos dos alunos, sendo uma grande referência de força e empoderamento para muitos.

"(..)“Se você não percebeu, você é o único representante do seu sonho na face da terra. Se isso não fizer você correr, eu não sei o que vai.”.".

(Emicida)

Resumo

O processo de vacinação contra a covid-19 foi uma alavanca fundamental para redução de mortes no Brasil e no mundo, e sua efetividade na imunização de rebanho, varia de acordo com a adesão da população e consequentemente, a visão que os indivíduos tem relação a elas. Para relatar opiniões, usuários costumam usar as redes sociais, como Twitter, para discutir sobre suas preocupações, como a vacinação contra covid-19, de forma livre e espontânea, tornando-se uma ótima fonte para investigar essas opiniões públicas. Este trabalho tem como finalidade utilizar dos princípios do campo de pesquisa de mineração de opinião, utilizando-se do método léxico, para executar análises a partir dos dados captados do twitter, de usuários brasileiros relacionados às vacinas comercializadas no Brasil, com o propósito de identificar as tendências em relação a opinião pública acerca do tema das vacinas da covid-19. O processo de classificação será realizado por um modelo onde as classes de sentimentos propostos são categorizadas em 3: sentimento positivo, sentimento negativo ou neutro.

Palavras-chaves: Análise de sentimentos, Twitter, covid-19, Vacinas, Mineiração de texto.

Abstract

The vaccination process against covid-19 was a fundamental lever for reducing deaths in Brazil and in the world, and its evolution in the immunization of the herd, varies according to the adherence of the population and, consequently, the view that individuals have in relation to they. To report opinions, users often use social networks, such as Twitter, to discuss their concerns, such as vaccination against covid-19, freely and spontaneously, making it a great source to investigate these public opinions. This work aims to use the principles of the research field of opinion mining, using the lexical method, to perform analyzes based on data captured from twitter, from Brazilian users related to vaccines marketed in Brazil, with the purpose of identifying trends in public opinion on the subject of covid-19 vaccines. The classification process will be carried out by a model where the standard sentiment classes are categorized into 3: positive sentiment, negative sentiment or neutral.

Key-words: Sentiment analysis, Twitter, Covid-19, Vaccines, Text mining.

Lista de ilustrações

Lista de abreviaturas e siglas

API	Application Programming Interface
JSON	JavaScript Object Notation
IDE	Integrated Development Environment
CSV	Comma-separated values
CRM	Customer Relationship Managemen
IDE	Integrated Development Environment
VADER	Valence Aware Dictionary and sEntiment Reasoner
URL	Uniform Resource Locator

Sumário

1 Introdução

Em dezembro de 2019 uma nova doença chamada covid-19 foi descoberta na china e rapidamente se espalhou pelo o mundo. A velocidade e a facilidade da propagação do vírus foram imensas, e isso trouxe uma grande preocupação para todos os países do planeta. As medidas de contenção e controle para evitar o avanço da doença rapidamente tiveram que ser pensadas e aplicadas. Uma dessas medidas foi o desenvolvimento de vacinas imunizantes contra a Sars-Cov-2, um dos principais mecanismos de proteção para evitar consequências devastadoras na saúde da humanidade.

Embora a vacinação e uso das máscaras sejam uma alavanca fundamental para a prevenção da doença, redução de casos e de mortes, o processo enfrentou resistência, ou até mesmo oposição, por parte da sociedade brasileira. Para que o sucesso da imunização e cuidados seja efetivo, é necessário uma grande adesão por parte da população, porém diversos fatores podem fazer com que a opinião pública em relação às vacinas mude, positivamente ou negativamente, ao longo do tempo, impactando na adesão dos imunizantes.

Com o avanço das tecnologias e das redes sociais, as informações compartilhadas livremente começaram a atingir cada vez mais a um número maior de pessoas. Sem qualquer tipo de filtragem esses conteúdos são compartilhados nas redes. Porém, grande parte deles são considerados como falsos, as chamadas fake news. No Brasil, a incidência deste desse tipo de material tem atingido números desenfreados, muito alavancado pela polarização política que vem se intensificando nos últimos anos. Segundo uma pesquisa feita pelo jornal CNN Brasil (*Guimarães and Rodrigues*, 2022) quatro em cada 10 pessoas afirmam receber notícias falsas diariamente.

Por isso se torna necessário um monitoramento constante de como as vacinas são vistas pela população, com objetivo de diminuir a rejeição e aumentar a aderência à imunização. Então, como compreender de forma inteligente e disruptiva a visão e o comportamento social em relação aos imunizantes da covid-19?

Yousefinaghania et al. (2021) acreditam que, por mais que pesquisas clássicas sejam úteis para investigar o ponto de vista da população, cada vez mais as redes sociais têm sido usadas para discutir e compartilhar os pontos de vista sobre tópicos de saúde de surtos de doenças infecciosas. Tornando assim, redes como twitter uma ótima fonte para realizar entendimento de como o tema da vacinação da covid-19 é visto por usuários e consequentemente, pela sociedade brasileira. Portanto, se torna interessante constantemente identificar as tendências temporais nos tweets relacionados às vacinas covid-19, para ajudar aos órgãos públicos a entender a visão da sociedade em relação ao tema,

podendo assim esses órgãos se concentrar na criação de materiais de conscientização em mídias sociais, baseados no quanto o tema é visto positivamente ou negativamente por esses usuários, impulsionando assim a conscientização sobre os imunizantes e reduzindo as visões negativas sobre eles e ajudando consequentemente na adesão da vacinação. Assim, redes sociais como twitter se tornam ótimos fontes de dados para realizar esse método de identificação.

Através das redes sociais milhares pessoas conseguem se conectar com grupos do seu círculo próximo de relacionamento ou até mesmo conhecer novos usuários que compartilham de um mesmo interesse (*Becker and Tumitan*, 2013). Nelas, esses usuários conseguem expressar suas opiniões diariamente, de forma natural, sobre diversos assuntos como: política, economia, entretenimento, ciência, cultura, entre outros. Desse modo, há um grande volume de dados sendo gerados pela interação desses indivíduos. Esses dados contêm opiniões, críticas e relatos sobre diversos assuntos que podem ser usados para o entendimento da opinião pública e do comportamento humano em relação a temas específicos, como, por exemplo, as vacinas da covid-19. O conhecimento das opiniões da população em relação aos impactos da vacinação versus o grau das infecções da covid-19 é fundamental, pois com essas informações podem servir para as entidades governamentais da saúde e entidades locais possam buscar mecanismos de mitigação e combate aos contágios e propagação do vírus.

O processo de extrair opiniões de expressões de linguagem natural é estudado dentro do campo de mineração de opinião que também pode ser chamado de análise de sentimento, esse campo de pesquisa tem duas abordagens básicas: a de aprendizado de máquina e a baseada em dicionário (*Sunitha et al.*, 2022). A abordagem escolhida para ser utilizada nesta pesquisa é baseada em dicionário léxico. Este trabalho tem como objetivo, utilizar este campo de pesquisa como um meio de identificação de como o tema da vacinação da covid-19 é visto por usuários brasileiros do twitter. Se certificando em qual das categorias se enquadra os tweets relacionados aos imunizantes. As classes de sentimentos propostos nesta pesquisa são categorizadas em 3: sentimento positivo, sentimento negativo ou neutro.

Isso se torna muito interessante no contexto social atual onde os usuários expressam e compartilham constantemente suas opiniões nas redes sociais como Twitter, sendo uma das redes sociais mais populares do Brasil. Segundo a revista eletrônica Valor Investe (*Braun*, 2022), o Brasil é o quarto país com maior número de usuários na plataforma, com cerca de 19,05 milhões. Nele são discutidos temas bastante pertinentes e que normalmente estão em alta no momento. Um dos principais temas que os internautas brasileiros discutiram nos últimos dois anos na rede social foi a Covid-19 e suas adjacências como sintomas, ações governamentais tomadas em relação à doença e as vacinas.

1.1 Problemática

Segundo *Yousefinaghania et al.* (2021), cerca de 70% de uma população precisa ser imunizada para alcançar um nível relevante de imunização de rebanho. Apesar do encadeamento da imunização ter sido crucial para a redução dos casos de mortes e infectados no país, existe uma resistência bastante significativa da população em receber as doses seguintes das vacinas da covid-19. Segundo o jornal CNN Brasil (*Resende and Alpaca*, 2022) apenas 6,63% da população brasileira tomaram a 4ª dose de reforço da covid, sendo que 83,98% da população já foi imunizada com ao menos uma dose e 78,93% têm o esquema primário completo (segunda dose). Trazendo então, uma mudança de comportamento da população brasileira em relação a aprovação do imunizante, que está muito relacionado em como elas são vistas pelas sociedade.

Portanto, o problema está na necessidade de mapear as tendências nas opiniões das pessoas em relação à vacinação da covid-19 em função dos comentários efetuados pela sociedade de forma natural, nas redes sociais, sem pressão alguma que limite o modo de pensar das pessoas. Sendo uma alternativa para isto, utilizar a mineração de opinião como termômetro de como as vacinas da covid-19 estão sendo vistas pelos usuários do twitter, realizando o processo de classificação dos posts relacionados aos imunizantes. Isso permite que através dessa identificação os órgãos de saúde possam perceber de forma temporal como o tema está sendo visto positivamente ou negativamente nas redes e tomar ações baseadas nessas informações obtidas através desses dados.

Criando campanhas mais agressivas de conscientização nas redes, caso haja uma tendência mais negativa no resultado das classificações, por exemplo. Com o propósito de trazer uma redução de mensagens negativas e aumentar mensagens positivas e consequentemente melhorar a aceitação dos imunizantes e diminuir a hesitação e a oposição às vacinas da covid-19.

1.2 Hipótese

Utilizar o método de análise de sentimento, como estudo de investigação da opinião de usuários do twitter em relação a vacinação da covid-19 no Brasil. Realizando o processo de classificação dos posts relacionados aos imunizantes em 3 categorias: sentimento positivo, sentimento negativo ou neutro. Sendo assim, esse trabalho pode ajudar o poder público a entender como os imunizantes estão sendo vistos e ajudar as as agências de saúde a definir o melhor momento de quando ser mais agressivo em relação a divulgação de materiais de conscientização nas redes sociais como twitter e outros meios de comunicação, para melhorar a visão dos usuários e da população em relação às vacinas e aumentando por consequência a absorção da população no calendário de vacinação contra covid-19 no país. Muitos trabalhos estão sendo desenvolvidos em todo mundo com o

mesmo propósito.

Trabalhos relacionados, como o de (*Nezhad and Deihimi*, 2022) enfocam em realizar o procedimento análise de sentimentos, como mecanismo de entendimento das opiniões iranianas sobre a vacinação contra a COVID-19 e identificou durante sua pesquisa que o sentimento negativo em relação às vacinas estrangeiras e nacionais da covid-19 aumentou ao longo dos meses em que a coleta de dados foi realizada. Os pesquisadores trouxeram como proposta para solução do problema a concentração de campanhas nas mídias sociais, como o Twitter, para promover mensagens e diminuir pontos de vista negativos em relação aos imunizantes. Outro trabalho interessante é o de (*Siru and Jialin*, 2021) que tem como principal objetivo identificar tendências temporais de tweets relacionados à vacina covid-19 em nível nacional dos Estados Unidos, utilizando o método de análise de sentimentos. O grupo de pesquisadores acredita que a análise de sentimento pode fornecer informações interessantes sobre o sentimento público em relação à vacina covid-19 e orientar órgãos de saúde pública na elaboração de programas de educação sobre vacinas.

1.3 Objetivo

O principal objetivo neste trabalho é realizar a análise de sentimento dos textos gerados pelos usuários do twitter em relação a vacinação da covid-19, utilizando-se do método de dicionário léxico. Podendo assim posteriormente estabelecer padrões através da classificação de polaridade, para melhor entendimento da opinião pública acerca do tema.

Objetivos específicos:

- Estabelecer os princípios do campo de pesquisa de mineração de opinião para executar análises a partir dos dados captados do Twitter.
- Realizar melhorias no dicionário léxico escolhido com o propósito de tornar classificação de polaridade mais assertiva.
- Identificar padrões de tweets (positivos, negativos ou neutros) feitos por usuários brasileiros.

1.4 Justificativa da pesquisa

O sucesso da adesão dos imunizantes da Sars-Cov-2, está muito vinculado à visão que a sociedade tem deles, sendo assim torna-se interessante os órgãos públicos realizarem constante uma análise inteligente da percepção da sociedade em relação ao tema. Uma maneira de resolver essa questão, é a utilização de dados estruturados de plataformas de

mídia social, tendo em vista que o tema da vacinação da covid-19 tem gerado uma grande quantidade de discussões nas redes sociais e em plataformas de mídia, sobre diversos fatores relacionados às vacinas, incluindo reações, proteção e a eficácia (*Shahriar et al.*, 2022). Como objetivo analisar textos e classificar a sua polaridade, ajudando a compreender sentimentos expressados neles, o campo de pesquisa mineração de opinião ou análise de sentimentos, como também é chamado, possibilita classificar avaliações, opiniões ou relatos em relação a um tema em específico, e a partir daí entender o quanto esse tópico, assunto ou produto é aceito pelo público em questão. Podendo ser amplamente usado em questões emergentes relacionadas à saúde pública (*Nezhad and Deihimi*, 2022). *Benevenuto et al.* (2018) acreditam que as opiniões das redes sociais colhidas e devidamente tratadas, podem ajudar a compreender e explicar diversos fenômenos sociais complexos, como a visão populacional em relação às vacinas da covid-19.

2 Informações de redes sociais e análise sentimentos

Com o avanço das redes sociais, pessoas do mundo todo tiveram seus alcances de escuta, relacionamento e interação ampliados. Através dessas redes, os usuários conseguem expressar opiniões, desejos, críticas e ideias. Diversos desses relatos estão em formato de texto e geram um grande volume de dados. Nos últimos anos, entender as informações contidas nesses dados tornou-se uma ferramenta interessante no entendimento de comportamentos de grupos de pessoas e da opinião desses grupos em relação a temas específicos.

Diversas organizações já utilizam de métodos para entendimento de opiniões de usuários nas redes sociais sobre sua marca ou produto, na intenção de tomar medidas cada vez mais ágeis e assertivas a respeito do público que almejam (*Soares*, 2017). Todo esse movimento de entendimento e tomada de ações pautada em informações contidas em dados, chamamos de data-driven. Ser data-driven faz com que seja possível entender tendências de mercado, comportamentos de consumidores ou até mesmo comportamentos de concorrentes.

No entanto, o grande volume de dados gerados faz com que seja necessário também metodologias e processos robustos, inteligentes e automatizados, que consigam captar os insumos das publicações feitas por esses usuários, como também o sentimento contido nelas (*Sampaio*, 2021). Dito isto, o campo de análise de sentimento vem exatamente nessa vertente, com o intuito de detectar e classificar a polaridade das opiniões contidas em um texto, relacionadas a um tema específico, neste caso o covid-19.

Existem diversas abordagens para a realização de análise de sentimentos e classificação de polaridade, as principais delas são léxicas e aprendizado de máquina. Cada abordagem tem sua característica específica.

2.1 Data driven: dados, informação e ação

Os dados são registros ou símbolos desconexos que na sua essência não apresentam sentidos isolados. Esses dados, neste contexto, podem ser classificados como quantitativos, qualitativos e categóricos. Em *Team* (2022), os quantitativos estão relacionados aos números, como por exemplo idade, preço e quantidade. Já os qualitativos são dados que representam a qualidade ou característica de algo, como por exemplo cor. E, por último, os categóricos indicam uma categoria, como por exemplo usuários novos e antigos. Em sîntese, (*Viana* (2014) – p.14) "define dado como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado".

Os dados se tornam insumos para se conseguir as informações. Através da organização e estruturação desses dados, consegue-se obter o significado contido neles. Assim podendo entender, contextualizar e interpretá-los. No mundo de hoje as informações têm um peso cada vez maior para o conhecimento e consequentemente para as tomadas de ações. Através das informações obtidas nos dados, consegue-se entender melhor a realidade que um produto, tema ou problema está inserido (*Soares*, 2017). Esse movimento de tomada de ação baseada em dados é conhecido como "orientado a dados". O conceito de data-driven tem a ver com o uso dos dados e as informações contidas neles, como principal matéria-prima, para auxiliar nas tomadas de decisão (*Soares*, 2017). Sendo assim, o processo data driven é construído através das estruturações dos dados, obtendo-se as informações que nos auxiliam na tomada de ações, como ilustra a Figura 1

Figura 1 – Etapas data-driven

Fonte: *Freaza* (2018)

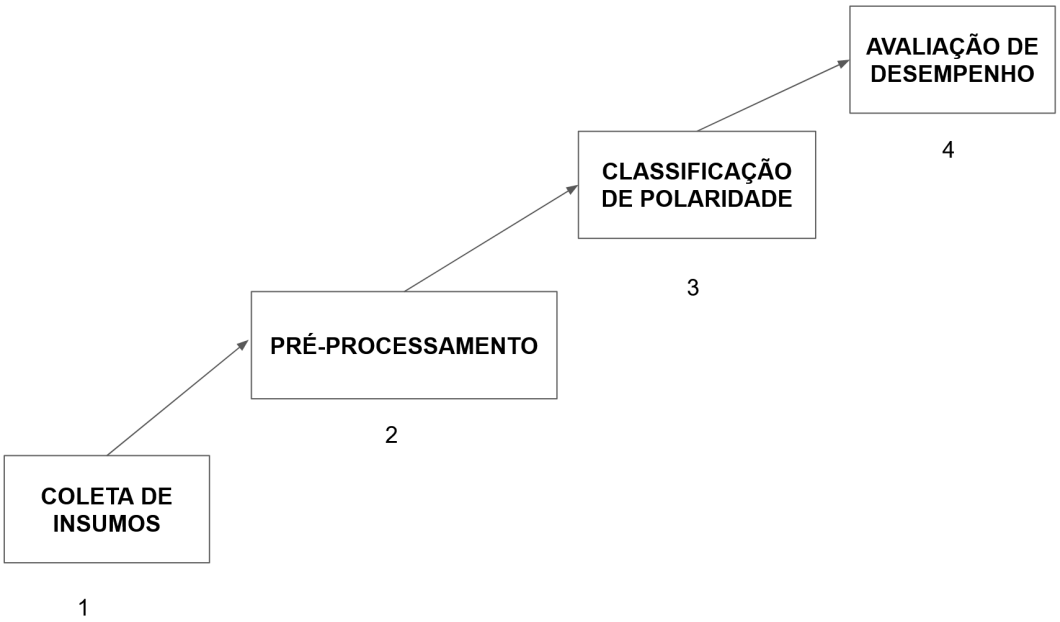
2.2 Análise de sentimentos

O campo de estudo de análise de sentimentos ou como também é chamado mineração de opinião tem como objetivo extrair opiniões, informações ou sentimentos de expressões de linguagem natural, utilizando-se de técnicas automatizadas. Segundo *Pezzini* (2016), a mineração de opinião contém várias etapas, mas quatro delas são consideradas básicas para todos os processos (*Sampaio*, 2021), tal como ilustrada pela Figura 2, que são: a coleta de insumos, pré-processamento, classificação de polaridade e avaliação de desempenho.

Becker and Tumitan (2013) afirma que existem dois tipos de abordagens básicas para realizar a classificação de polaridade ou extração de sentimentos: aprendizado de máquina e baseado em dicionário. O método baseado em dicionário, como o próprio nome sugere, faz uso de um dicionário léxico que possui em seu conteúdo um significado quantitativo de cada palavra, onde cada termo é associado a uma pontuação previamente rotulada. Já o método de aprendizado de máquina, torna-se necessário realizar o processo

de treinamento de um modelo computacional utilizando um algoritmo de aprendizagem com amostras previamente classificadas.

Figura 2 – Etapas mineração de opinião



Fonte: O autor

(A) **Coleta de insumos**

A fase de coleta de material é uma das fases mais importantes, porque aqui é montada a base de dados voltada para o tema alvo. A qualidade dos dados coletados, e a relação desses dados com o assunto que procura-se fazer a análise, é um fator significativo para o sucesso das fases posteriores *Sampaio* (2021). Existem dois tipos de coleta de dados: a realizada por API 's e as manuais.

- **Coleta por API -** API (Interface de programação de aplicações) é um conjunto de regras e protocolos que ajudam componentes de softwares a se comunicar e interagir uns com os outros. Atualmente diversas plataformas oferecem API's para coleta de dados como por exemplo Youtube e Twitter.
- **Coleta Manual -** Como o próprio nome já sugere, são aqueles dados coletados de forma manual. Normalmente a manipulação é feita a partir de dados gerados por sistemas internos de uma corporação, como por exemplo dados capturados manualmente de um CRM (Customer Relationship Manageme) ou também inseridos manualmente em arquivos CSV.

(B) **Pré-processamento**

Após a etapa de coleta, se torna fundamental realizar a de pré-processamento, pois os dados coletados na maioria das vezes não estão naturalmente estruturados, sendo necessário realizar a remoção de conteúdos irrelevantes para facilitar o processo de classificação de polaridade. No pré-processamento o texto é reestruturado para torná-lo mais digerível na etapa de classificação, tornando-se uma etapa muito importante para aprimorar a qualidade dos dados brutos (*Sunitha et al.*, 2022).

No pré-processamento é feita limpeza e padronização dos dados que foram coletados e estão sendo pré-processados com o objetivo de preparar esses insumos para fazer a classificação (*Sampaio*, 2021). É nessa fase também que deixamos apenas o material que será utilizado na pesquisa, por exemplo, se o intuito da análise é classificar e polarizar texto oriundos do twitter, logo é feita a remoção de todas palavras irrelevantes, imagens, símbolos e tags que podem vir através da coleta. Essa etapa de pré-processamento é muito importante para tornar a identificação e a classificação das palavras cada vez mais assertivas.

(C) **Classificação de polaridade**

Para *Benevenuto et al.* (2018), a polaridade é a representação do grau do quando um texto é negativo ou positivo. Esta etapa tem como propósito realizar uma rotulagem dos textos utilizando de alguns dos métodos de abordagem, aprendizado de máquina ou baseado em dicionário. Esses métodos podem retornar uma polaridade, considerando que Pi representa a polaridade do texto Ti, como:

- Resultado discreto binário (positivo ou negativo)

$$T_i = P_i > 0, \rightarrow \ Positivo; \ P_i < 0, \rightarrow \ Negativo \tag{2.1}$$

- Resultado discreto ternário (positivo, negativo ou neutro)

$$T_i = P_i > 0, \rightarrow \ Positivo; \ P_i = 0, \rightarrow \ Neutro; \ P_i < 0, \rightarrow \ Negativo \tag{2.2}$$

Na classificação da polaridade é feita a tokenização, que basicamente consiste no processo de separar um texto em unidades menores; ou seja, separar as palavras contidas no texto, para posteriormente utilizar um dos tipos de abordagem para executar o processo de classificação de polarização dos textos contidos no conjunto de dados coletados (*Sampaio*, 2021). A cada palavra contida no texto é atribuída um valor correspondente, resultando assim uma polaridade final para o texto.

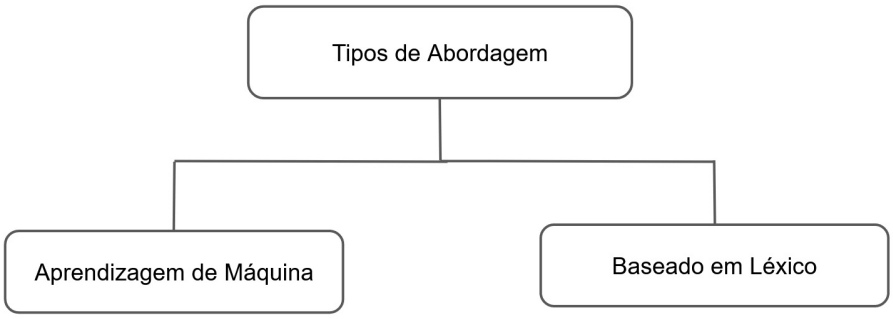
(D) **Avaliação de desempenho**

A última etapa temos a de avaliação de desempenho, onde é feita a validação do classificador. Nesta etapa é reunido os dados classificados e os transformados em informações, encontrando possíveis correlações feitas pelos grupos de dados polarizados. Esses resultados podem ser representados através de gráficos ou até mesmo tabela para que seja possível retirar insights de como um tema ou tópico é visto por um conjunto de usuários (*Sampaio*, 2021).

2.3 Tipos de abordagem

Existem na literatura diferentes técnicas utilizadas para realização da etapa de classificação de polaridade, dentro do campo de pesquisa de mineração de opinião. Os dois tipos de abordagens são: "machine learning"e a baseada em dicionário (*Becker and Tumitan*, 2013), como ilustrado na Figura 3. São constituídas pelas 4 etapas básicas: coleta de insumos, pré-processamento, classificação de polaridade e avaliação de desempenho, já demonstradas. Porém, cada uma delas tem suas particularidades.

Figura 3 – Análise de sentimentos



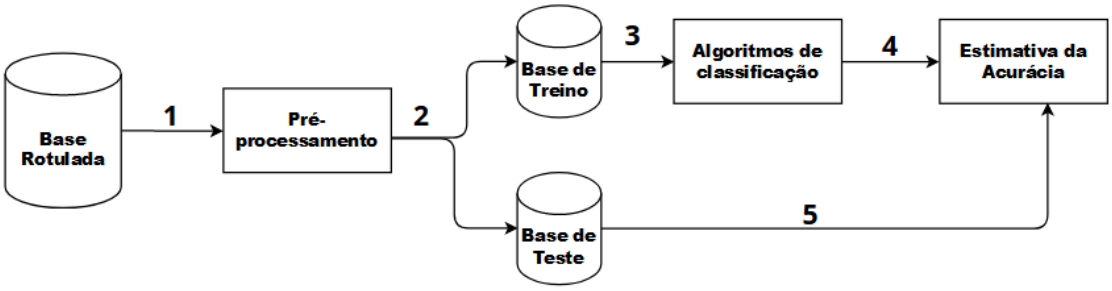
Fonte: O autor

A abordagem baseada em aprendizagem de máquina normalmente necessita de uma base de dados previamente rotulada chamada de material de treino, que é usada para fazer a etapa de treinamento do algoritmo, que será utilizado para fazer a processo de classificação dos textos (*Benevenuto et al.*, 2018). Já, as abordagens baseadas em dicionários têm como ponto central a utilização de dicionários de palavras associada à respectiva polaridade dos termos, onde cada palavra é relacionada a uma pontuação previamente rotulada.

2.3.1 Machine learning

A abordagem de machine learning ou aprendizagem de máquina consiste no treinamento de um modelo com textos rotulados e utilização do modelo de forma que ele seja capaz de identificar o sentimento em sentenças de forma automática (*Benevenuto et al.*, 2018). Nesse tipo de abordagem é utilizado um algoritmo de aprendizagem para criação de um modelo computacional que aprende os padrões para que posteriormente conseguir realizar a classificação de forma automática (*Maurosso et al.*, 2017). Para realizar manipulação da abordagem de aprendizado de máquina é necessário seguir as etapas ilustradas na Figura 4.

Figura 4 – Etapas da Abordagem de Aprendizagem de máquina



Fonte: O autor

(A) A primeira fase constitui em coletar o dado e preparar um material, que consiste em um conjunto de dados categorizado, ou seja, o texto é atrelado a uma classificação prévia, construindo o que chamamos de base rotulada (*Maurosso et al.*, 2017).

(B) Após o processo de criação da base rotulada, é necessário realizar o procedimento de pré-processamento dessa base, com o objetivo de preparar o texto rotulado para fase posterior que é o treino do algoritmo de classificação (*Benevenuto et al.*, 2018).

(C) Logo após o pré-processamento a base rotulada é dividida em dois tipos de materiais: os de treino e os de teste.

(D) Os materiais de testes são utilizados nos algoritmos de classificação para gerar o modelo que será usado para realizar a classificação dos textos, posteriormente. Já o material de teste é usado na fase de acurácia para verificação e validação do modelo classificador (*Benevenuto et al.*, 2018).

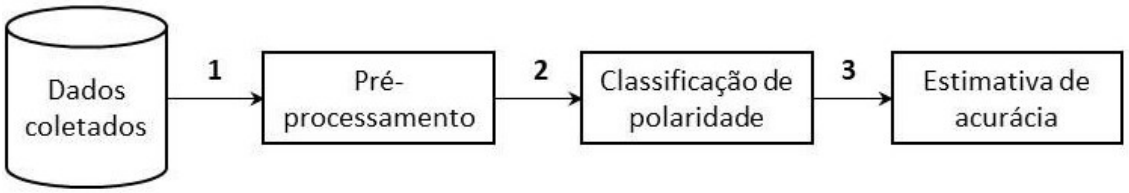
(E) A última fase, é a estimativa da acurácia ou também chamada de avaliação de desempenho, consiste em realizar os testes, e identificar se o que foi treinado está funcionando efetivamente, com o auxílio do material de teste (*Maurosso et al.*, 2017).

2.3.2 Baseado em léxico

Diferente da abordagem de aprendizagem de máquina, as baseadas em dicionário léxico não necessitam de sentenças previamente rotuladas e treinos para a criação de um modelo, sendo essa uma das principais vantagens desse tipo de abordagem pois a aplicação não fica restrita ao contexto para o qual foram treinados (*Benevenuto et al.*, 2018). No método léxico é utilizado um dicionário que contém palavras previamente classificadas com as suas polaridades, relacionando uma palavra a um valor numérico, através desse dicionário é designado um valor, onde cada termo está relacionado a uma pontuação previamente rotulada, o que é chamado de polaridade prévia (*Maurosso et al.*, 2017).

Nessa abordagem também são realizadas as etapas básicas de mineração de opinião como coleta de insumos, pré-processamento, classificação de polaridade e avaliação de desempenho como ilustrado na Figura 5. Porém, o que diferencia a abordagem léxica da de aprendizagem de máquina são basicamente dois pontos: Não necessitam de material rotulado para treino do modelo e utilizam como base para a classificação de polaridade um dicionário léxico de sentimento.

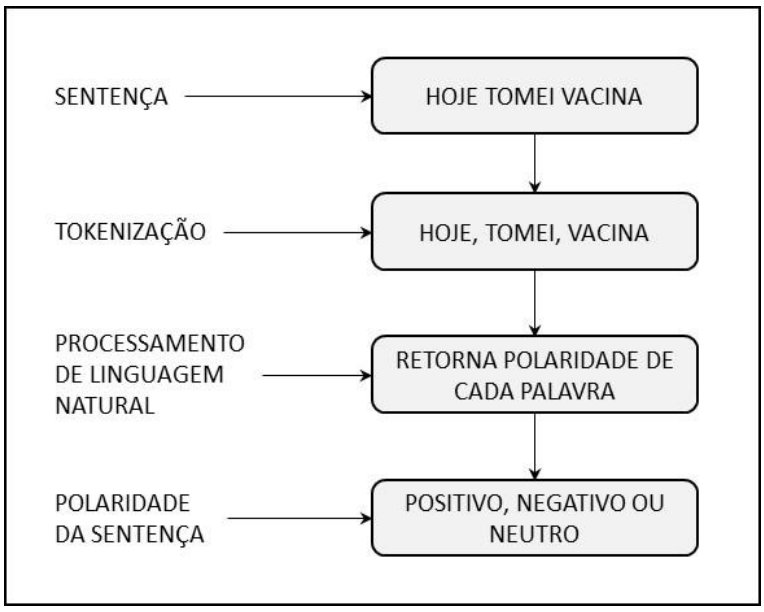
Figura 5 – Etapas da Abordagem Baseada em Léxico



Fonte: O autor

A Figura 6 mostra de forma generalizada mostra o funcionamento dessa abordagem na classificação de polaridade. Segundo *Benevenuto et al.* (2018) o processo consiste da seguinte maneira: O processo de classificação inicia quando o método recebe um texto de entrada. Em seguida, a sentença a ser classificada é tokenizada, que é a técnica de separar as palavras do texto em várias unidades. Logo, é realizado o processamento onde é feito uma busca das palavras no dicionário e retornando a polaridade correspondente dos termos. E por último o método é capaz de inferir qual é a polaridade ou sentimento implícito na sentença de entrada, retornando resultados discretos binários (positivo ou negativo) ou ternários (positivo, negativo ou neutro).

Figura 6 – Abordagem dicionário Léxico



Fonte: O autor

2.4 Redes sociais e covid-19

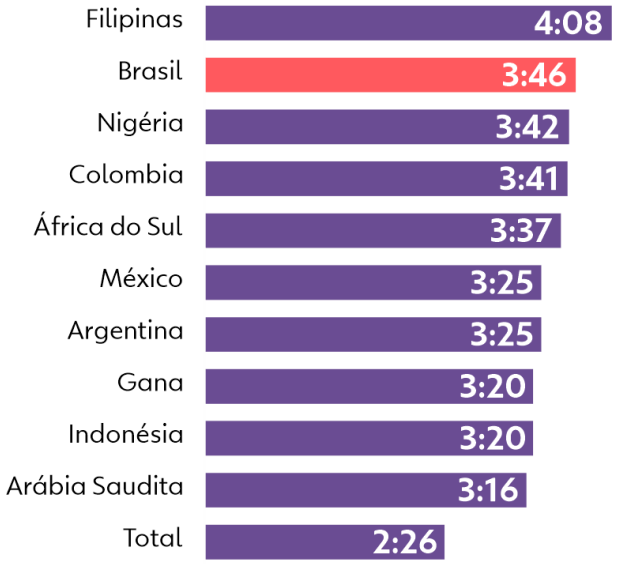
Com o avanço da tecnologia e dos meios de comunicação, as redes sociais se tornaram um método de facilidade de interação entre as pessoas. Através delas é possível se relacionar com usuários de todo mundo com poucos cliques e em tempo real (*Henrique et al.*, 2021). Essa facilidade de interação fez com que esses espaços se tornassem um lugar extremamente propício para que as pessoas expressassem suas opiniões sobre temas variados e esse movimento vem acontecendo em diversos países e culturas. Por exemplo, no período da pandemia, a população brasileira utilizou diferentes tipos de redes sociais para informar e comentar aasuntos relacionados com a covid-19 (*Dhein et al.*, 2022). Nesse período, diversas medidas sanitárias para a contenção da propagação do vírus foram tomadas e o isolamento social foi uma delas. Impedindo o funcionamento do comércio em geral, gerando fechamento de escolas, bares e diversas empresas adotaram home office. Isso fez com que as pessoas tivessem a necessidade de novos métodos para se conectar, interagir e se informar.

As redes sociais se tornaram um ambiente de debate onde os usuários puderam se atualizar sobre o avanço da doença, expressar suas angústias, se informar sobre as medidas tomadas em relação à doença. Inclusive em relação ao processo de vacinação que foi amplamente discutido nas diversas plataformas.

2.4.1 Redes sociais

Segundo *Henrique et al.* (2021), com o surgimento da internet e das ferramentas da web 2.0, despontou também um novo meio de se comunicar que vai além dos limites geográficos permitindo a interação independente de tempo ou espaço, as chamadas redes sociais. As redes sociais são sites ou muitas das vezes aplicativos que permitem usuários realizarem o compartilhamento de informações, sejam elas conteúdos de humor, utilidades ou até mesmo notícias. Em uma pesquisa feita pelo Centro de expertise setorial em telecom, o Brasil se encontra na 2º posição dos países com maior tempo médio diário gasto usando redes sociais, atrás apenas da Filipinas (*Lourenco and Lontra* , 2022). Como ilustrado na Figura 6.

Figura 7 – Países em tempo médio diário gasto usando mídias sociais



Fonte: *Lourenco and Lontra* (2022)

Atualmente existe uma grande gama de tipos de redes sociais que abrangem temas como relacionamento, profissional e entretenimento. Se tornando muita das vezes o principal meio de comunicação por onde as pessoas interagem. Dentro desses nichos existem diversos cases de sucesso como Tinder, Linkedin e Twitter. Através do twitter por exemplo, que foi a rede social escolhida como fonte de dados para obtenção dos insumos que serão utilizados para análise nesta pesquisa, são gerados grandes volumes de dados constantemente, que podem ser utilizados para gerar valor.

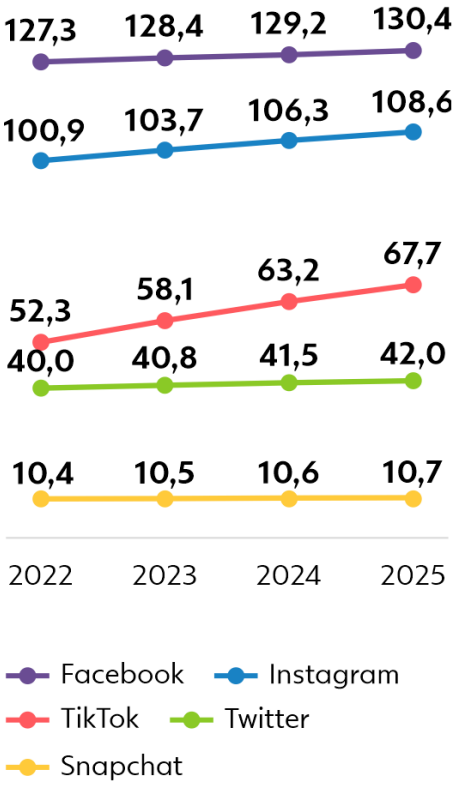
2.4.2 Twitter

A rede social twitter é uma ferramenta de micro mensagens criada em 21 de março de 2006, tendo seu primeiro tweet publicado em outubro do mesmo ano pelo seu CEO e cofundador, Jack Dorsey. A rede tem como principal dinâmica a publicação de pequenas

mensagens, onde muitas das vezes geram interações e conversações entre os usuários da plataforma (*Logghe et al.*, 2016). Através do twitter é possível que as pessoas compartilharem o que estão pensando em textos curtos (280 caracteres) e com possibilidade também de inserção de mídias como imagens e vídeos.

No Brasil, segundo uma pesquisa realizada pelo Centro de expertise setorial em telecom *Lourenco and Lontra* (2022), o Twitter é a 4º rede social mais utilizada no país e já conta com cerca de 40 milhões de usuários em 2022. Além de ter a projeção de crescimento de sua base de usuários para 42 milhões até o ano de 2025, ficando atrás apenas de Facebook, Instagram e TikTok. A Figura 8 mostra a quantidade de usuários de Redes Sociais no Brasil por Plataforma.

Figura 8 – Quantidade de usuários de Redes Sociais no Brasil por Plataforma



Fonte: *Lourenco and Lontra* (2022)

Pela característica da rede social de fácil compartilhamento de mensagens, e uma formação concentrada de opiniões e discussão gerada pela interação dos usuários da plataforma, produzindo assim um volume massivo de dados, o Twitter se tornou uma grande ferramenta de estudo e pesquisa de mineração de opinião. Criando inclusive, uma API que facilita a extração desses dados da plataforma para realização de pesquisas de objetivos variados. Um exemplo são pesquisas para o entendimento do debate online de temas relacionados a covid-19.

2.5 Covid-19

Um dos principais temas em alta nas redes sociais dos últimos anos, a Covid-19 é uma infecção viral causada pela síndrome respiratória aguda grave por coronavírus 2 (SARS-CoV-2) (*Shereen et al.*, 2020). Tendo como origem a cidade de Wuhan, na China, rapidamente se espalhou para todo o globo. Proveniente de um surto de pneumonia de causa desconhecida e tendo como principais afetados pessoas que frequentavam mercado de frutos do mar de Wuhan onde se vendiam animais vivos como morcegos, sapos, entre outros *Duarte* (2020).

2.5.1 Sintomas

No início do ano de 2020 a organização mundial da saúde declarou pandemia, devido ao novo coronavírus altamente contagioso chamado síndrome respiratória aguda grave –coronavírus-2 (SARS-CoV-2)–. Entre humanos a transmissão da doença pode acontecer por ações como tosse, espirro ou até mesmo a fala de indivíduos contaminados.

A melhor forma de evitar a propagação da doença é entender como a doença é causada e como se espalha, podendo assim tomar iniciativas que impacte de forma eficiente no controle da doença. As principais delas são o processo de imunização e o distanciamento social no caso de pessoas já infectadas, fazendo com que esses indivíduos evitem contato e espalhem a doença por meio de gotículas de saliva ou secreção nasal. Assim que entra em contato com o vírus e é contaminado, alguns indivíduos apresentam sintomas característicos. Porém outros não manifestam nenhum tipo de sintoma, essas pessoas são chamadas de assintomáticos. Portanto é importante ficar atento aos meios de transmissibilidade para evitar a contaminação.

Segundo *Petr et al.* (2020), cerca de 90% dos pacientes normalmente apresentam mais de um sintoma resultante da doença e aproximadamente 15% apresentam sintomas como febre, tosse e falta de ar. Após a passagem dos sintomas provenientes da covid-19 estima-se que a eliminação das partículas virais demoram cerca de 8 a 30 dias.

O Covid-19 tem como principais sintomas: Febre, Tosse, Dispneia, Mialgia (Dor Muscular), Fadiga, Anosmia/Disgeusia. Em alguns casos mais graves, também pode apresentar: Confusão, Tontura, Cefaleia, Dor torácica, Hemoptise, Diarreia, Náuseas/vômitos, Dor abdominal, Manifestações cutâneas, entre outros.

2.5.2 Transmissibilidade

Na maioria das vezes jovens e pessoas com nenhum histórico de doenças crônicas como diabetes e doenças cardíacas apresentam sintomas leves ou moderados, e em alguns casos nenhum sintoma (*Gilliam et al.*, 2020). Além disso, alguns estudos apontam que

determinados indivíduos podem ser contagiosos mesmo durante o período de incubação do vírus (estimado entre 1 a 14 dias), sendo chamado de transmissão pré-sintomática (*Petr et al.*, 2020).

Isso traz uma preocupação extra, pois esses grupos que não apresentam sintomas podem servir de meio de contaminação efetivo da doença quando entram em contato com outras pessoas. Sendo assim, se torna extremamente importante a necessidade de uma ampla adesão do calendário de imunização da doença e também do isolamento social, no caso de indivíduos contaminados.

2.5.3 Complicações e sequelas

O Sars-Cov-2 como várias outras doenças graves podem causar diversos impactos no corpo humano a curto e longo prazo. Mas por ser uma doença extremamente recente tem seus sintomas e sequelas após a crise aguda ainda um pouco desconhecidos, porém baseado em outras doenças pulmonares virais espera-se que em casos graves ocorra em algum grau sequela em certos pacientes (*Islam et al.*, 2020). Como por exemplo, uma das maiores epidemias já vistas no mundo, a gripe espanhola que teve uma taxa de mortalidade bem alta, matando cerca de 27-50 milhões de pessoas em todo o globo terrestre. Os sobreviventes relataram que um dos sintomas mais presentes no período pós-infeccioso foi a fatiga extrema. Tendo sido identificado em um estudo que 28% dos indivíduos apresentaram níveis incomuns de fadiga após a crise aguda da doença. Logo espera-se que uma das sequelas presentes nos contaminados com sintomas do coronavírus seja a fadiga pós viral (*Islam et al.*, 2020).

Sendo assim, *Petr et al.* (2020) relata que até o momento as principais complicações causadas pelo covid-19 são:

- Complicações cardiovasculares - Identificados em 7% a 20% dos pacientes.
- Insuficiência respiratória aguda - identificados em 8% dos pacientes.
- Choque séptico - Relatada em 4% a 8% dos pacientes.
- Lesão renal aguda - Relatada em 3% a 8% dos pacientes.

  Além disso, alguns indivíduos doentes mostram complicações neurológicas, como doença cerebrovascular aguda, convulsões, neuralgia, lesão nos músculos esqueléticos, entre outras. Há relatos que o vírus já foi detectado no cérebro e no líquido cefalorraquidiano, chegando ao sistema nervoso central *Petr et al.* (2020). Outras sequelas identificadas entre vários pacientes que tiveram a doença covid-19 foram a falta de olfato e paladar, devido a níveis muito altos da enzima que está presente também na região do nariz chamada ACE-2 (enzima conversora de angiotensina II),

que permite a entrada do coronavírus nas células do corpo causando infecção. E em alguns casos, mesmo após a fase mais crítica da doença, os sentidos não voltam impossibilitando que diversos indivíduos retornem às suas atividades normais do dia a dia.

2.5.4 Prevenção

Medidas para contenção do vírus são extremamente necessárias, para evitar o aumento do número de casos e consequentemente o de mortes. Durante todo o ano de 2020 o mundo teve que se readaptar a práticas educacionais cruciais e eficazes para evitar a propagação do vírus. Uma delas é higienizar sempre bem as mãos com água e sabão por pelo menos 20 segundos e sempre que tocar superfícies em locais públicos utilizar álcool em gel (*Petr et al.*, 2020). Além de praticar sempre a etiqueta da tosse, cobrindo nariz e boca ao tossir e espirrar, utilizando obrigatoriamente máscaras faciais em locais públicos com o propósito de evitar a propagação e contaminação por partículas infecciosas que ficam suspensas no ar por gotículas liberadas em ações como tosse e espirro. Outra medida de prevenção é o distanciamento social com o objetivo de reduzir a transmissão do vírus, continua Petr et al. Uma pesquisa feita pela BBC mostra que se reduzir em 50% do contato social, um indivíduo infectado reduz seu potencial de contágio de 406 pessoas em um mês para apenas 15 pessoas como ilustrado na Figura 9 (*BBC*, 2020).

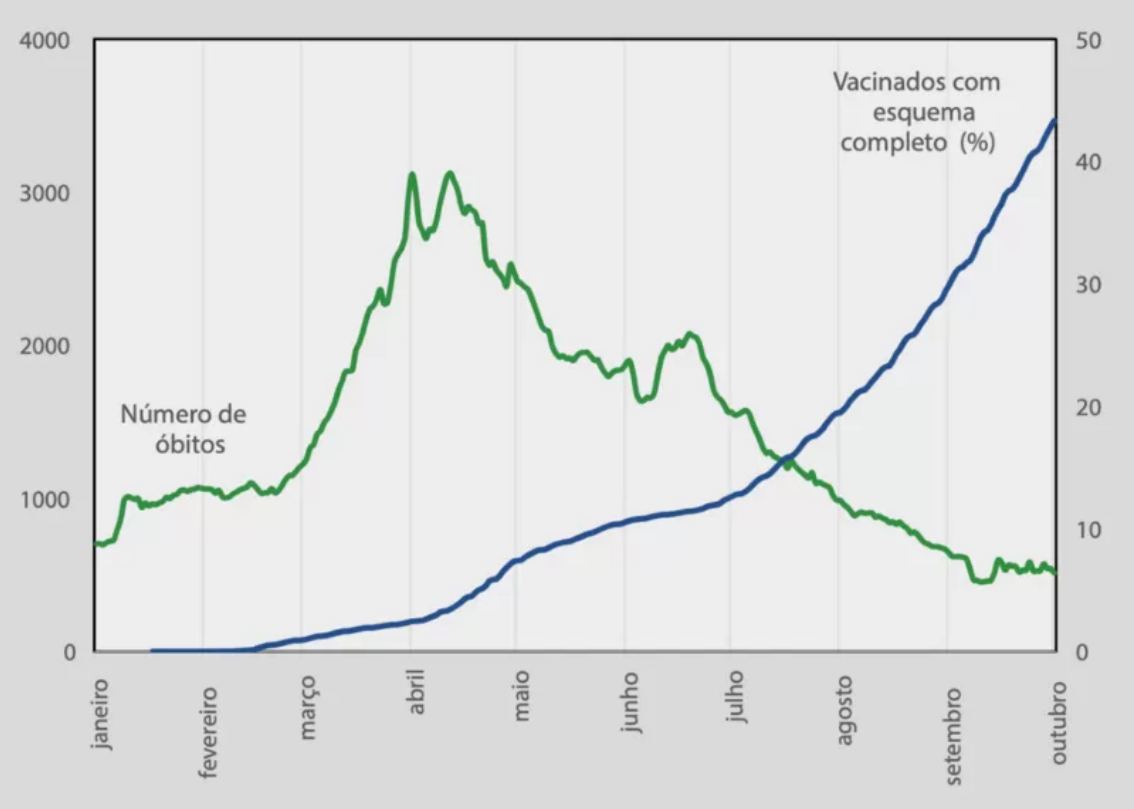
Figura 9 – Importância do distanciamento social

Fonte: *BBC* (2020)

2.5.5 Vacinas

Outra ferramenta fundamental para contenção e controle da doença covid-19 são as vacinas imunizantes. Criadas em tempo recorde devido a todo impacto econômico e social que a doença e as medidas do isolamento trouxeram para sociedade. As vacinas da covid no brasil foram essenciais para a redução do caso de mortes no brasil, como mostra o gráfico disponibilizado no boletim observatório covid-19, da Fundação Oswaldo Cruz (FioCruz) na Figura 10, onde demonstra a correlação do avanço da vacinação e os casos de óbitos relacionados à doença no brasil.

Figura 10 – Gráficos de mortes e vacinação no ano de 2021



Fonte: *Fiocruz* (2021)

Tendo em vista todo esse contexto, torna bastante relevante pensar no desenvolvimento de novas soluções para o entendimento da percepção da sociedade em relação às vacinas da covid-19. E a utilização de mecanismos tecnológicos pode se tornar grande aliado para ajudar nesse propósito, como por exemplo o modelo proposto neste projeto.

2.6 Trabalhos relacionados

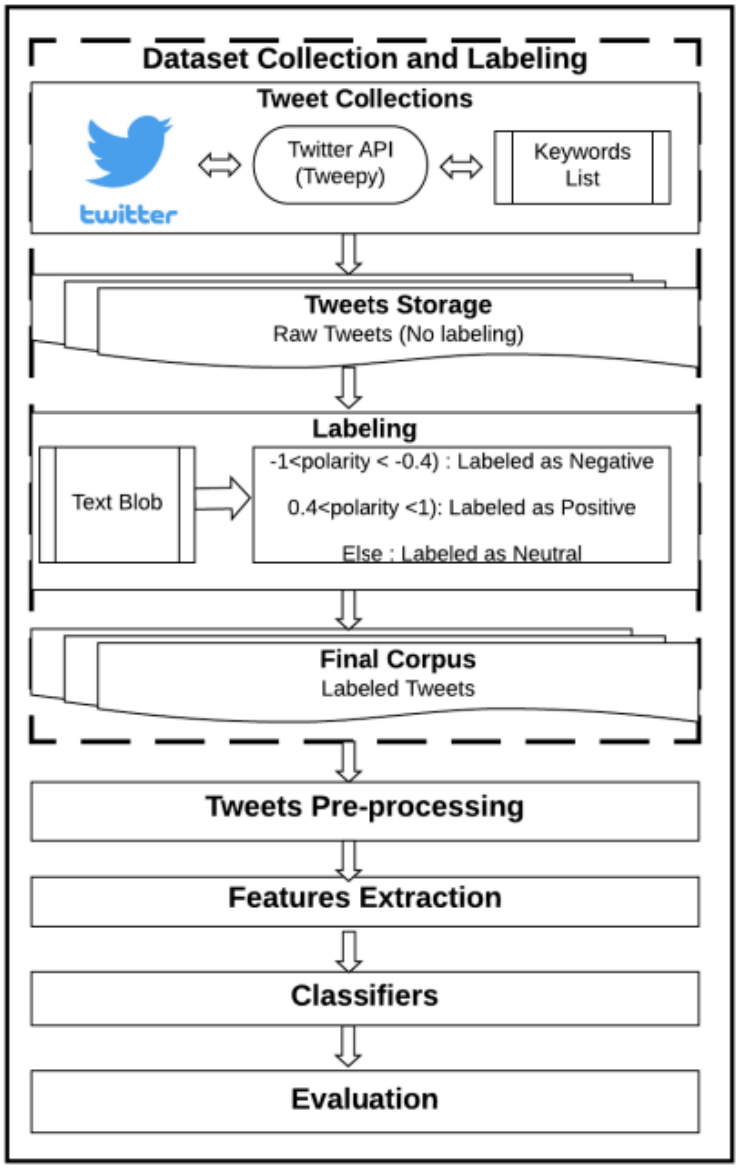
Há muitos anos a aplicação do conhecimento tecnológico vem ajudando na melhora de vida da humanidade e andando junto com a medicina com o propósito de melhorar o controle de doenças que assolam a humanidade há décadas. No cenário de pandemia a união dessas forças se tornaram mais evidentes e necessárias, tendo em vista a rapidez de propagação do vírus e as incertezas sobre ele e das informações geradas acerca do tema (*Barbosa*, 2020). Todo esse panorama gerou grandes fóruns de discussões principalmente nas redes sociais, espaço onde as pessoas podiam se conectar e comunicar de forma segura diante das medidas sanitárias tomadas. Muitos trabalhos têm surgido no campo de pesquisa de análise de sentimentos aplicada em diversos contextos e que pode também ser utilizada para a identificação de padrões de relatos feitos pelos usuários do twitter em relação às vacinas da da covid-19.

Por exemplo, *Xu et al.* (2022) diz que os internautas frequentemente expressam seus pontos de vista nas redes sociais e o surto da covid-19 resultou em um grande volume de opiniões e emoções sobre os eventos da pandemia em redes como Twitter, Facebook, entre outras. Logo, seu trabalho teve como objetivo analisar o sentimento de todos os posts realizados na plataforma twitter, relacionados à vacina covid-19 que continham atributos de idioma marcados como inglês. Na pesquisa foram realizadas as etapas de obtenção de dados usando a API do twitter, pré-processamento onde foi removido símbolos não ingleses, como endereços da web, IDs do twitter, convertendo sinais de pontuação em espaços e todas as letras em inglês para minúsculas. Também foi realizada a etapa de classificação de polaridade, onde o VADER que é um modelo de aprendizado não supervisionado foi usado para classificar o valor dos sentimentos(positivo, neutro, negativo) do conjunto de dados. E por último análise e comparação dos resultados onde foi constatado que durante o período da pesquisa as pessoas tinham sentimentos diferentes entre a vacina chinesa e as de outros países e o valor do sentimento pode ser afetado pelo número de notícias diárias e de casos e mortes.

Ansari et al. (2020) entendem que as redes sociais têm se tornado um grande espaço de debate, gerando um alto volume de dados podendo então ajudar no contexto eleitoral, tendo em vista pode ser interessante para os candidatos entenderem a percepção do público em relação a sua campanha e propostas políticas. Sendo assim, ela e seu grupo usam em sua pesquisa a análise de sentimento como método de analisar as postagens feitas pelos usuários no twitter referentes às eleições gerais da Índia em 2019, em relação aos principais partidos políticos nacionais que participaram do processo eleitoral. Para realizar a classificação dos tweets, o grupo utilizou como métodos de abordagem o aprendizado de máquina.

Naseem et al. (2021), utilizaram o método de análise de sentimentos para realizar a classificação de polaridade, usando a ferramenta TextBlob com a finalidade de rotular o sentimento emocional em positivo, negativo e neutro, dos tweets coletados. Tendo como principal objetivo analisar as opiniões de posts realizados no Twitter sobre o covid-19. O grupo de pesquisadores acredita que o processo de mineração de opinião pode ajudar a saúde pública desenvolver uma presença proativa e ágil para combater a disseminação de sentimentos negativos nas mídias sociais após uma pandemia. A Figura 11 mostra visão geral da estrutura proposta pelos pesquisadores, que também utilizaram a biblioteca Tweepy para realizar a coleta do conjunto de dados oriundos do twitter e um dicionário léxico para realizar o processo de classificação de polaridade.

Figura 11 – Visão geral da estrutura de análise de sentimentos



Fonte: *Naseem et al.* (2021)

Além do trabalho *Sarsam et al.* (2021), que recorreu aos dicionários léxicos NRC (Affect Intensity Lexicon e SentiStrength) para extrair estados para cada tweet com a proposta de dectar emoções das mensagens do Twitter em relação a conteúdos relacionados ao suicídio, as classificações geradas pelos diconairos eram (raiva, medo, tristeza, alegria, positiva e negativa). Foi aplicado neste estudo método de aprendizado de máquina semi-supervisionado com objetivo de reconhecer com eficiência tweets relacionados ao suicídio.

Já *Viteri* (2021), acredita que o método de análise de sentimentos é ferramenta fundamental para o sucesso das atividades voltadas para o público e que as redes sociais por gerarem um grande volume de dados contendo opiniões de forma espontânea, têm se firmado como um cenário válido para a realização de análises. Em sua pesquisa, realizou

todas as etapas principais para realizar a mineração de opinião, obtenção de insumos utilizando a biblioteca do python Tweepy, pré-processamento implementando métodos para eliminar símbolos que não contribuem para a análise, na etapa de classificação de polaridade foi baseada em dois dicionários léxicos Vader e TextBlob classificando os tweets entre positivo ou negativo, já na etapa de análise de resultados foi realizada a comparação entre os resultados obtidos dos dois dicionários usados na pesquisa.

Huerta et al. (2021) compreendem que existe uma necessidade de observação contínua do sentimento do público, com o objetivo de fazer com que órgãos governamentais de segurança pública tomem ações mais efetivas para o benefício da população. Em seu trabalho, os pesquisadores tinham como objetivo explorar os sentimentos de tweets relacionados a covid-19, realizados por moradores de Massachusetts. Explorando assim, padrões temporais e identificando as mudanças na polaridade de sentimentos em postagens do Twitter. O método utilizado para realizar a classificação de polaridade foi o dicionário léxico VADER que é voltado especificamente para a polaridade de sentimentos.

Os autores, *Sumitro et al.* (2021) realizaram uma pesquisa com o objetivo de descobrir opinião pública sobre a política de vacinas indonésia contra a covid-19 no Twitter, utilizando o método análise de sentimentos. Em seu trabalho os tweets foram classificados em 5 categorizados, sentimento muito positivo, sentimento positivo, sentimento negativo, sentimento um tanto negativo ou neutro. Na pesquisa foram realizadas as etapas principais para realização da análise de sentimentos, coleta de dados, pré-processamento, classificação de polaridade e análise dos resultados. A classificação de polaridade foi baseada no dicionário Vader que é um método de análise baseado em léxico de análise de sentimento baseado em regras.

3 Mineração de opinião

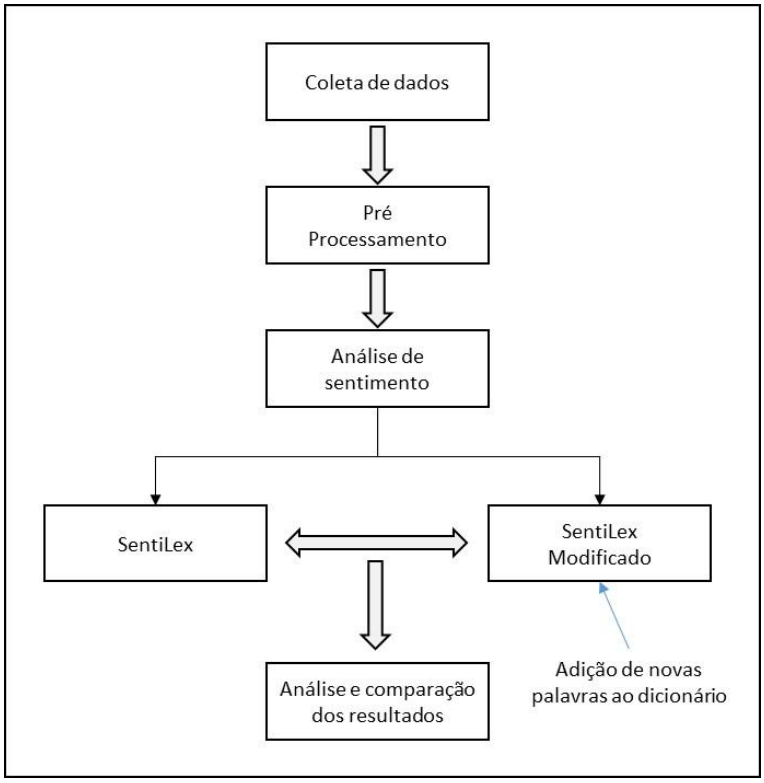
Esse trabalho tem como finalidade utilizar dos princípios do campo de pesquisa de mineração de opinião para executar análises a partir dos dados captados do twitter. Examinando os possíveis debates realizados acerca da vacinação da covid-19 na rede social, tendo em vista os diversos tópicos que são gerados diariamente em relação ao tema na plataforma, sejam eles em relação aos sintomas causados pelo imunizante ou questionamentos da eficácia do mesmo.

Para isso, o método escolhido para realização da análise de sentimento foi o léxico, utilizando como base o dicionário SentiLex, para a classificação e realização das análises. Elaborando também melhorias no dicionário léxico escolhido, gerando assim o dicionário SentiLex modificado, com o propósito de tornar classificação de polaridade mais assertiva.

3.1 Estrutura da sequência

O sistema tem como estrutura básica colher dados oriundos do twitter através de API (Interface de programação de aplicações) e armazená-los em um banco de dados localizado no Google Driver. Logo após, o algoritmo realiza a etapa pré-processamento onde é realizada toda parte de limpeza, padronização e normalização dos dados coletados. E em seguida com os dados colhidos normalizados e padronizados é feito o processo de classificação de polaridade utilizando os dicionários léxicos SentiLex e SentiLex modificado (que, de forma paralela também é feita toda melhoria desse dicionário para posteriormente realizar a comparação dos resultados), como mostra a Figura 12. A proposta do trabalho é procurar entender a opinião de usuários do Twitter em relação à vacinação da Covid-19, utilizando-se dos princípios do campo de pesquisa de mineração de opinião.

Figura 12 – Fluxograma do dispositivo



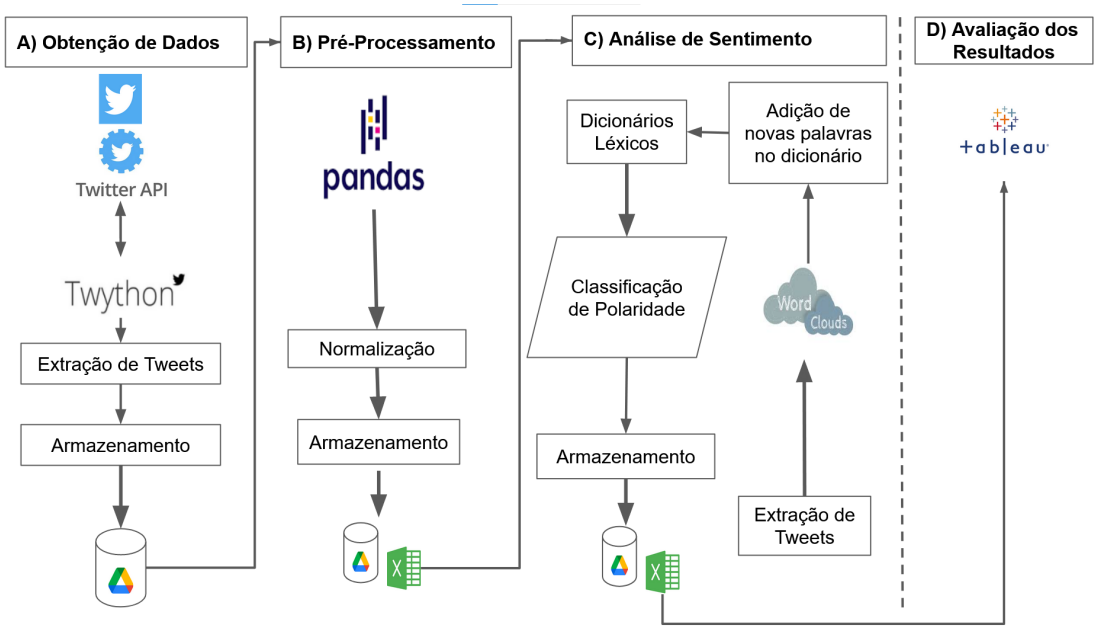
Fonte: O autor

3.2 Arquitetura do sistema

A arquitetura geral do sistema consiste na captação de dados gerados pela interação dos usuários do twitter, onde publicam pequenos textos, muitas das vezes acompanhado de imagens, vídeos e links. Porém como a mineração de texto é feita com base em dados de característica textual é necessário realizar a etapa de pré-processamento que é onde é realizado a normalização e padronização para uma análise efetiva e assertiva dos sentimentos.

Logo após o pré processamento os dados normalizados são armazenados em um banco de dados que serão utilizados pelo protótipo na fase de classificação, que é onde será separado o texto em unidades menores ou seja, separar palavra por palavra contida na sentença e verificar a polaridade que cada palavra representa dentro do dicionário léxico, em seguida o método é capaz de inferir qual é a polaridade do texto de entrada. E por último é realizada avaliação de desempenho com o objetivo de validar se o método proposto, está gerando os resultados esperados, como pode ser visto na Figura 13.

Figura 13 – Arquitetura do sistema



Fonte: O autor

(A) **Obtenção dos dados**

Nesta fase é coletado um conjunto de dados contidos em posts públicos realizados dentro da plataforma twitter, usando a API do twitter e bibliotecas do python para consultar, coletar e manipular esses dados. As pesquisas dos tweets vão ser restritas a posts que contenham o nome de pelo menos uma das vacinas comercializadas no brasil e o idioma foi definido em apenas conteúdos em português. No final da coleta, todos os dados são reunidos em um grande conjunto de dados localizado no Google Drive.

(B) **Pré-processamento**

O objetivo dessa pesquisa é fazer uma análise textual dos sentimentos de posts realizados na plataforma twitter em relação às vacinas da covid-19. Porém, muitas das vezes o conteúdo dos posts publicados no twitter são compostos também de imagens, URL, vídeos e hashtags. Sendo assim, torna-se importante realizar a limpeza desses dados brutos. Logo, serão utilizados os módulos e bibliotecas em Python, com a proposta de deixar o texto em uma forma mais apropriada para o processo de análise de sentimentos.

(C) **Classificação de polaridade**

Com a etapa de pré-processamento realizada, os dados estão preparados para a classificação de polaridade, com o propósito de categorizar o sentimento contido no texto. Nesta etapa será realizada a categorização do sentimento dos textos contidos

nos tweets publicados pelos usuários do twitter em positivo, negativo ou neutro. Cada palavra do texto possui uma pontuação correspondente dentro do dicionário léxico: palavra positiva conta como + 1, enquanto as negativas e neutras palavras como - 1 e 0, respectivamente. O dicionário léxico escolhido como base para realização da pesquisa será o Sentilex, que de acordo com (*Carvalho and Silva*, 2015) é um léxico criado com o propósito de realizar pesquisas de mineração de opinião de textos escritos em português. Além disso, serão realizadas melhorias nesse dicionário, através da adição de palavras relacionadas ao tema covid-19 com a proposta de trazer mais assertividade.

(D) **Avaliação de resultados**

Após a execução de todas as etapas anteriores é realizada a etapa de avaliação de resultados onde é feita a comparação de assertividade da classificação realizada com base nos dicionários léxicos citados, com uma classificação feita manualmente. O objetivo é mostrar a validação do método proposto, como também demonstrar a melhoria do Sentilex modificado em comparação com o Sentilex.

4 Desenvolvimento

O método análise de sentimentos têm sido usados em diferentes temas, como eventos políticos, análises de produtos, resenhas de filmes, entendimento de surtos e numerosos outros assuntos, como sugere (*Sunitha et al.*, 2022). Implementa-se, neste trabalho, a realização do entendimento da opinião pública em relação às vacinas da covid-19 com base na mineração das informações trocadas por twitter. Para realização do método de mineração de opiniões é preciso elaborar as etapas de obtenção de dados, pré-processamento, análise de sentimentos e avaliação de resultados.

O desenvolvimento deste trabalho consiste na obtenção de dados extraídos do Twitter por meio de sua interface de programação de aplicativos (API), chamada API twitter. O processo de autenticação da API do twitter é efetuado utilizando Twython que é uma biblioteca do python responsável por acessar dados do Twitter, através das credenciais API key e API secret key. Essas credenciais são geradas para que seja possível identificar quem está requisitando acesso e coleta desses dados. Após o processo de autenticação, os dados brutos são coletados por meio de uma função e armazenados em um banco de dados localizado no Google Drive.

Os dados coletados são pré-processados, que segundo a definição de *Qorib et al.* (2023) é processo de tornar o dado bruto em uma forma mais digerível e preparada para o processo de classificação de polaridade, utilizando bibliotecas e comandos em Python. Nesta etapa de pré-processamento, serão realizadas duas fases: a fase de normalização que consiste na remoção de conteúdos irrelevantes nos dados do twitter utilizados na análise e posteriormente a de armazenamento desses dados, já normalizados e preparados para a fase de análise de sentimentos utilizando o método de dicionário léxico .

O dicionário léxico sentilex escolhido para calcular os valores de polaridade de tweets sobre as vacinas covid, é um dicionário léxico em portugês utilizado como base para efetuar análises automáticas de sentimento (*Carvalho and Silva*, 2015). A etapa de análise de sentimentos é realizada em três fases: aprimoramento do dicionário léxico escolhido, classificação de polaridade e armazenamento. O aprimoramento do dicionário constitui na adição de novas palavras relacionadas ao tema covid-19, já a fase de classificação de polaridade irá classificar as palavras contidas nos textos dos tweets com base no valor de sentimento atribuído no léxico. Por último, os textos classificados serão armazenados para a fase de análise.

Na análise e comparação de resultados é onde se poderá observar a acurácia dos métodos propostos neste trabalho, como também a assertividade de ambos os dicionários léxicos utilizados na pesquisa. A apresentação dos resultados é gerada pelo software ta-

bleau, uma plataforma de análise onde é possível gerenciar dados e informações e gerar visões nas quais ajudará gerar insights e validar as hipóteses.

4.1 IDE e linguagem de programação

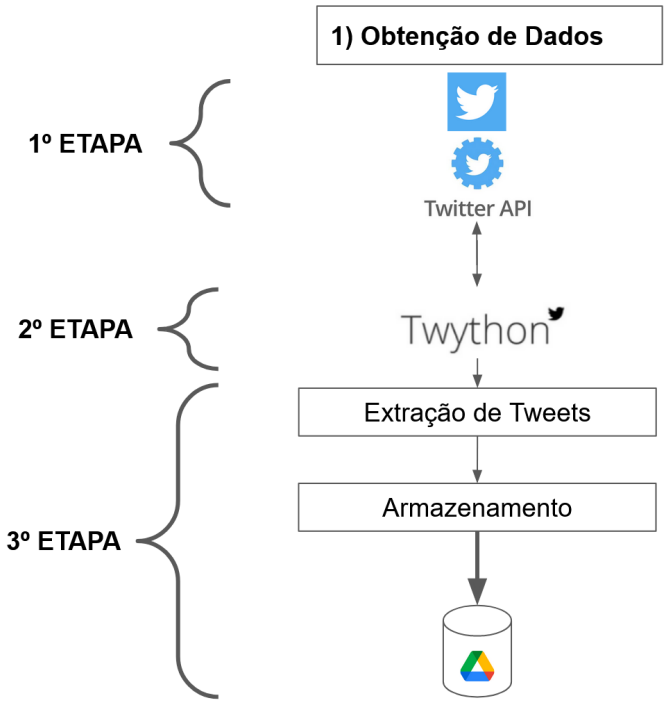
Na realização das etapas abtenção de dados, pré processamento e classificação de polaridade é utilizado como ambiente de desenvolvimento o Google Colaboratory, que é uma ferramenta que permite escrever e executar scripts em python, que é a linguagem de programação escolhida para realização desta pesquisa. A linguagem python possui diversas bibliotecas que auxiliam na manipulação de dados não estruturados, permitindo manipular arquivos de texto, com tipos de dados diversos. Neste trabalho iremos utilizar esta linguagem e suas bibliotecas dentro da plataforma em nuvem Google Colab.

O Google Colaboratory tem como suas principais vantagens a possibilidade de manipular notebooks hospedados do Jupyter, sem necessitar configurações, além de bibliotecas pré-instaladas, facilidades no compartilhamento de código e nenhum custo adicional para utilizá-lo (*Lourenco and Lontra* , 2020). Todo acesso à plataforma pode ser feito pelo próprio navegador e para ingressar no ambiente o usuário precisa ter uma conta google, pois o acesso a ferramenta é feito pelo Google Drive. Nela é possível organizar os códigos em Python em um conjunto de células, os chamados notebooks. O que gera uma melhor organização e escalabilidade.

4.2 Obtenção dos dados

A primeira fase deste trabalho referente ao método de análise de sentimento é o de obtenção dos dados encontrados na plataforma Twitter. Fase que é constituída de 3 etapas principais. 1) Acesso às credenciais da API do Twitter; 2) Instalar e instanciar o Twython; 3) Coleta e armazenamento dos dados consultados e consumidos pelas API 's. Como mostra a Figura 14.

Figura 14 – Etapas da fase de obtenção dos dados



Fonte: O autor

1. **Acesso às credenciais da API do Twitter -** Para ter acesso às credenciais do twitter é necessário que o usuário acesse a página de desenvolvedor da plataforma com uma conta de usuário comum. Logo após é realizado todo processo do registro e configuração do app que é onde serão geradas as credenciais de acesso à API do Twitter. Todo esse processo é necessário para que a plataforma saiba através das credenciais, quem está requisitando acesso e coleta dos dados.

2. **Instalar e instanciar o Twython -** Com as credenciais geradas é necessário instalar e instanciar o Twython que é uma biblioteca do python responsável por acessar dados do Twitter. Como ilustrado na Figura 16 as credenciais API key e API secret key salvas em um arquivo json, são passadas em forma de parâmetro e instanciando o objeto.

Figura 15 – Instalação e instanciamento da biblioteca Twython

```
#Importando Twython
from twython    import Twython

#Importando biblioteca Pandas
import pandas as pd


#Carregando credenciais do arquivo json
with open('/content/drive/MyDrive//twitter_credenciais.json', 'r') as file:
creds = json.load(file)


#Instanciando objeto
python_tweets   = Twython(creds['CONSUMER_KEY'], creds['CONSUMER_SECRET'])
```

Fonte: O autor

3. **Coleta e armazenamento dos dados consultados e consumidos -** Como última etapa dentro da fase de obtenção dos dados, temos a busca de tweets que contêm palavras chaves pré-estabelecidas definidas em uma lista. A Figura 14 ilustra a função chamada buscar tweets que requisita a API descrita anteriormente. A coleta do tweet só é realizada se o texto da postagem contém pelo menos uma das palavras mencionadas na lista.

Figura 16 – Função de busca de tweets

```
vacinas      = ['Pfizer', 'Coronavac', 'Janssen', 'Astrazeneca']
print('Serão extraídos tweets de ' + str(len(vacinas)) + ' vacinas.')

for vacina in vacinas:
  print('Iniciando processamento de ' + vacina)

  buscar_tweets(vacina , "", 50)
  df = pd.DataFrame(dict_)
  df.to_csv('/content/drive/MyDrive/Projeto Monografia/AmostraVacinas.csv',
index=False)

  print('Final do processamento de ' + vacina)
  randomico   = random.randrange(10, 30)
  print('Vamos aguardar ' + str(randomico) +' segundos.')
  time.sleep(randomico)
```

Fonte: O autor

Como esse trabalho tem como objetivo realizar a análise de sentimento dos textos gerados pelos usuários do twitter em relação a vacinação da covid-19, a Figura 17

mostra as palavras escolhidas para compor a lista de restrição de coleta são os nomes das vacinas comercializadas no Brasil, com o propósito de trazer mais assertividade nas buscas.

Figura 17 – Palavras-chave relacionadas ao tema utilizadas

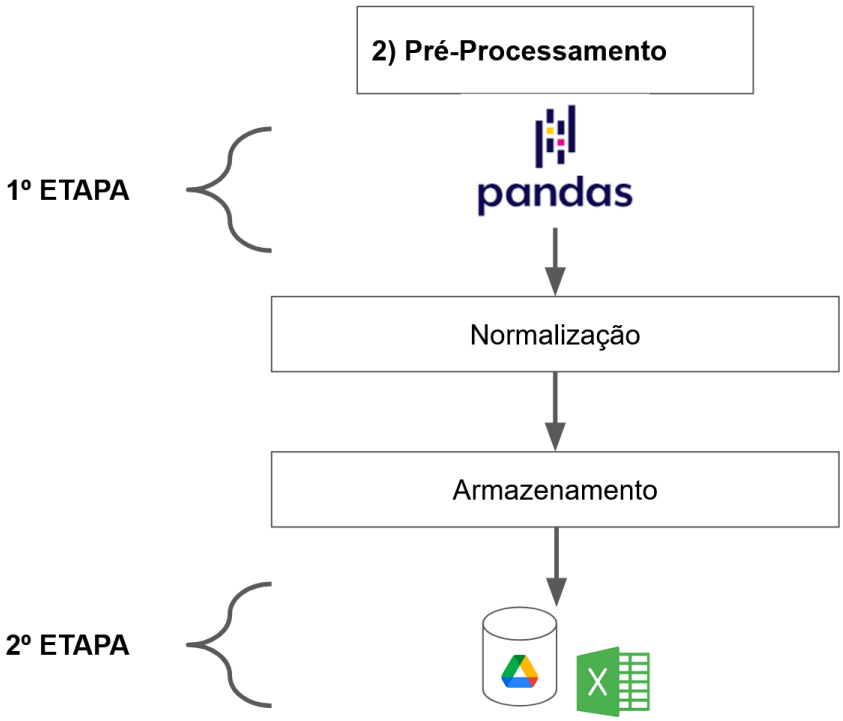
Tema	**Termos presentes na lista**
Vacinas	Pfizer;Coronavac;Janssen;Astrazeneca

Fonte: O autor

4.3 Pré-processamento

A Segunda fase é o pré-processamento, que é a etapa onde é realizada todo tratamento dos tweets captados anteriormente. Nela é feito os processos de limpeza, normalização e padronização. Como alguns tweets vem acompanhados de imagens, vídeos e links se torna necessário um tratamento para melhorar a qualidade dos dados e ter uma representação mais estruturada, tendo em vista que a mineração de opinião realizada neste trabalho é textual. Como ilustrado na Figura 18, nesta fase são realizadas as seguintes etapas: 1) Normalização; 2) Armazenamento.

Figura 18 – Etapas da fase de pré-processamento



Fonte: O autor

1. **Normalização -** No processo de normalização são desempenhadas atividades com a finalidade de preparar os dados para execução da tarefa de classificação de polarização, onde os textos são separados em unidades menores, ou seja, separar palavra por palavra contida na sentença e posteriormente realizar a verificação da polaridade que a palavra representa dentro do dicionário, para isso é necessário que o texto esteja o mais acurado possível. Desse modo, essa é uma das partes mais importantes para o bom desempenho das análises. Dentro dessa etapa são realizadas as seguintes atividades:

   - Remover linhas duplicadas;
   - Remoção de símbolos e caracteres especiais;
   - Remoção de links;
   - Remoção de imagens, vídeos ou emoticons;
   - Remoção de Stopwords (preposições, pronomes, artigos);
   - Padronizar todo o texto para letras minúsculas.

   A Figura 19 mostra o processo de normalização que é realizado por duas funções: a processamento1, no qual é responsável pela normalização dos dados e a processamento2, que tem como função fazer a limpeza desses dados.

Figura 19 – Funções de Normalização

```
def processamento1(limpeza):
  limpeza        = limpeza.lower()
  frase = unicodedata.normalize('NFD', limpeza)


  return   frase.encode('ascii', 'ignore').decode('utf8')
```

```
def processamento2(texto_retirar):
  texto_retirar     = re.sub(r'(@[A-Za-z0-9áéíóúÁÉÍÓÚâêîôÂÊÎÔãõÃÕçÇ
$@-_.&+]+)|(https?://[A-Za-z0-9áéíóúÁÉÍÓÚâêîôÂÊÎÔãõÃÕçÇ$@-_.&+]+)|(https?)', " ",
texto_retirar)
  return texto_retirar
```

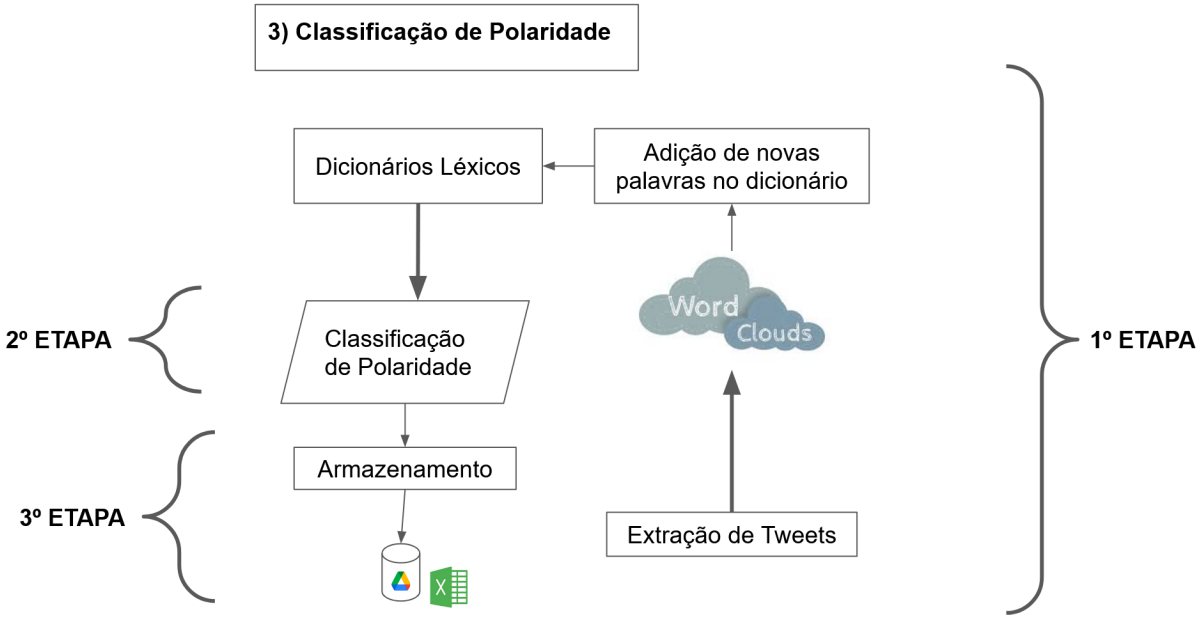
Fonte: O autor

2. **Armazenamento-** Terminada a etapa de normalização dos dados tratados, eles serão salvos em um banco de dados localizado no google drive, para que em seguida sejam usados na fase análise de sentimentos.

4.4 Classificação de polaridade

Na classificação de polaridade é feita a categorização dos tweets que foram coletados e pré-processados. O objetivo dessa fase é expressar os sentimentos presentes nos textos dos tweets. Essa fase consiste das seguintes etapas: 1) Aprimoramento do dicionário léxico escolhido; 2) Classificação de polaridade; 3) Armazenamento. Como ilustrado na Figura 20.

Figura 20 – Etapas da fase de classificação de Polaridade



Fonte: O autor

1. **Aprimoramento do dicionário léxico -** A dinâmica de melhoria do dicionário sentilex consistiu da seguinte forma: Durante todo o mês de agosto de 2022 foram coletados tweets que na sua composição tinham termos relacionais as vacinas do coronavírus, no total foram catalogados cerca de 122 novos termos. A ideia era incorporar ao dicionário sentilex original novas palavras que surgiram ao longo da pandêmica da covid-19 e que não estavam presentes no dicionário, tendo como objetivo principal trazer mais acuracidade e assertividade no processo de classificação e identificação das palavras. O processo de descobrimento das novas palavras foi realizado através das chamadas nuvens de palavras, que é uma representação visual de palavras mais presentes em um conjunto de dados. A Figura 20 mostra uma das wordcloud realizada, na semana do dia 05/08/2022.

Figura 21 – Nuvem de Palavras

Fonte: O autor

2. **Classificação de polaridade -** Dentro do processo de classificação de polaridade temos dois movimentos: o de tokenização e o de classificação. O processo de tokenização consiste na ação de separar os textos em unidades menores, ou seja, palavra por palavra. Para que após, seja realizada a etapa de classificação, que consiste no processo de consultar cada token (palavra presente no texto), nos dicionários utilizados na pesquisa e logo em seguida ocorre o retorno da polaridade correspondente da palavra presente no dicionário, com esses valores de retorno podendo ser 1, -1 ou 0. Logo em seguida é realizado um somatório de todas palavras classificadas presentes na sentença, resultando na classificação final do texto (positiva, negativa ou neutra). Como demonstra a Figura 22

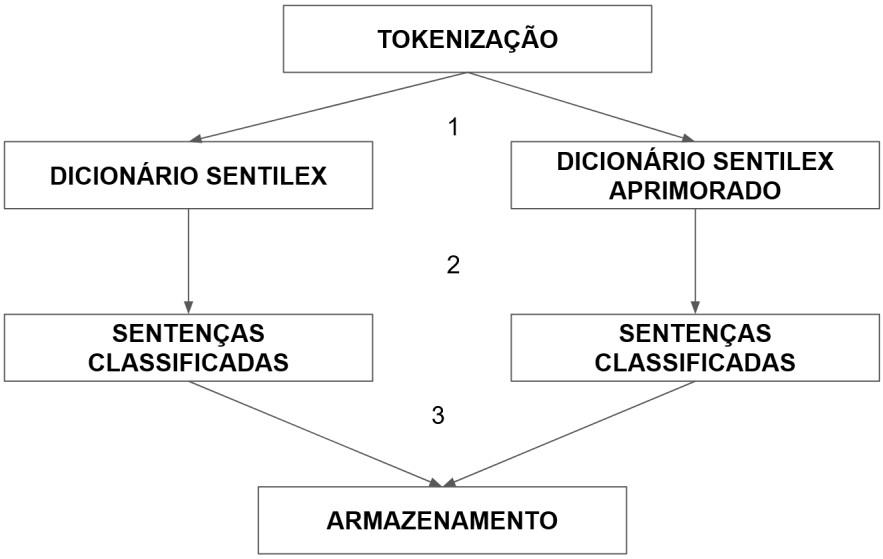
Figura 22 – Função de Classificação de polaridade

```
sentimento      = pd.DataFrame()

teste = df2['text'].apply(processamento1)
teste = teste.apply(processamento2)

for Frase in teste:
   tokens_covid2         = Frase.split(' ')

   l_sentimento = []
   for p in tokens_covid2:
      l_sentimento.append(int(dic_palavra.get(p, 0)))
      score = sum(l_sentimento)
      if score > 0:
        senti_numero  = 'Positivo'
      elif score == 0:
       senti_numero  = 'Neutro'
      else:
        senti_numero  = 'Negativo'
   total_sentimento_rodada   = senti_numero
   sentimento                = sentimento.append({'sentimento':
total_sentimento_rodada}, ignore_index=True)
   total_sentimento_rodada   = 0
   senti_numero              = 0

sentimento
```

Fonte: O autor

3. **Armazenamento-** Com a etapa de classificação finalizada, as sentenças classificadas serão salvas em um banco de dados localizado no google drive, para que posteriormente sejam feitas as análises comparativas em relação aos dois dicionários léxicos utilizados neste trabalho. Como ilustra a Figura 23 .

Figura 23 – Armazenamento da Classificação de polaridade

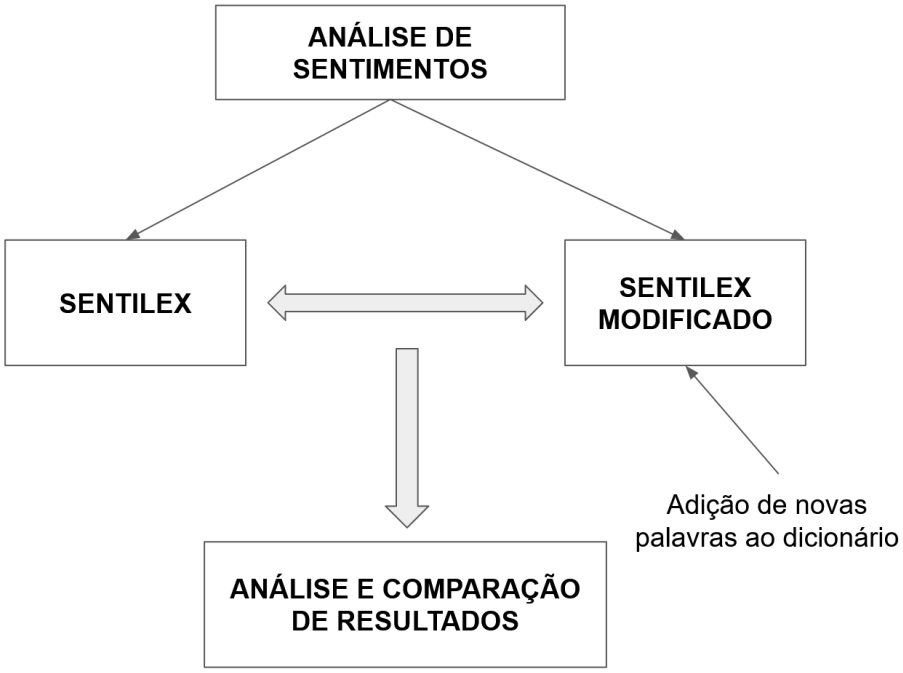


Fonte: O autor

4.5 Avaliação dos resultados

Como última fase temos a de avaliação de resultados, onde será feita a comparação das classificações de polaridade realizadas com base nos dicionários Sentilex original e o Sentilex modificado com as alterações de melhorias feitas com objetivo de tornar a classificação mais assertiva. A dinâmica de comparação ilustrada na Figura 24 será realizada da seguinte forma: Uma amostra de dados relacionada ao tema de vacinação da covid-19 será captada do twitter, será classificada de forma manual e posteriormente será feita a comparação das classificações feitas com base nos dicionários, identificando qual dos dicionários chega mais próximo da classificação feita por humanos.

Figura 24 – Esquema da análise de resultados



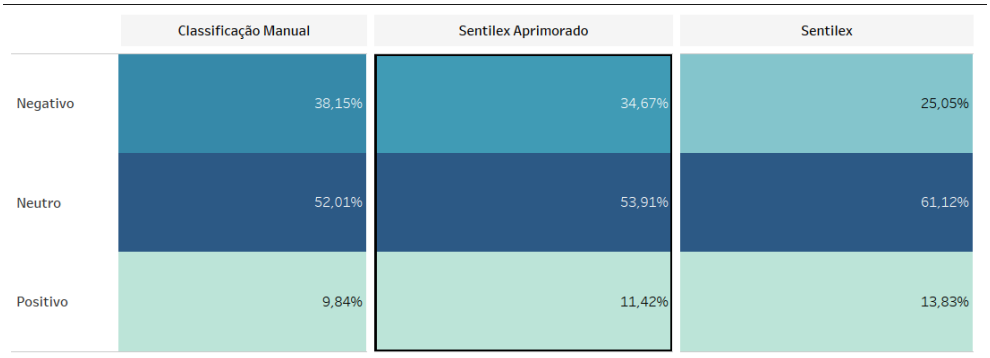
Fonte: O autor

5 Resultados

Nesta seção serão discutidos os resultados relacionados a avaliação da análise de sentimentos de uma amostra com 500 tweets, que foram captados, pré-processados e classificados. Utilizando como base para a classificação, os dicionários léxicos Sentilex original e o Sentilex modificado com as alterações de melhorias feitas com objetivo de tornar a classificação mais assertiva. Como descrito no capítulo anterior, a dinâmica de validação foi realizada fazendo uma comparação dos dicionários e da análise manual, com o propósito de analisar qual dicionário mais se aproxima de uma avaliação humana.

Considerando esse cenário, o software escolhido para reproduzir a apresentação dos resultados foi o tableau desktop, que é uma ferramenta utilizada para realizar projeções visuais de métricas e criação de Dashboards. Através dela é possível criar campos calculados utilizando-se de lógica de programação. Os resultados serão mostrados em forma de tabela de destaque, onde sua representação é definida pela a linha onde a cor mais intensa está sinalizando maior representatividade dentro do resultado. Sendo assim, temos o panorama:

Figura 25 – Resultados dos dicionários léxicos



Fonte: O autor

Como ilustrado na Figura 25, na análise manual, dos 500 tweets classificados cerca 190 (38,15%) foram classificados como negativos, 259 (52,01%) foram classificados como neutros e 49 (9,84%) como positivos. Já na análise utilizando como base o Sentilex Modificado, dos 500 tweets classificados cerca 173 (34,67%) foram classificados como negativos, 269 (53,91%) foram classificados como neutros e 57 (11,42%) como positivos. Por último temos os resultados das análises utilizando como base o Sentilex original, onde 125 (25,05%) tweets foram classificados como negativos, 305 (61,12%) foram classificados como neutros e 69 (11,42%) como positivos.

5.1 Discussões

Podemos perceber que mesmo diante dos três recortes distintos (Sentilex Original, Sentilex Modificado e Classificação manual), a amostra dos 500 tweets recolhida na semana 07/11, demonstra que parte representativa dos tweets é considerada como Neutra. Podemos perceber também que temos uma maior semelhança de comportamento entre os cenários Sentilex Modificado e Classificação manual, demonstrando assim que as melhorias realizadas nessa pesquisa no Dicionário Sentilex original geraram valor e impacto positivo, quando fazemos classificação de tweets relacionados a vacinação da covid-19. Como ilustrado na figura 25.

Portanto, com esses resultados também podemos concluir que:

- Em Tweets Classificados Neutros:

Figura 26 – Tabela de resultados de assertividade de em Tweets Classificados Neutros

Tipo de Dicionário	% Assertividade com a classificação humana
Sentilex Original	82,23%
Sentilex Modificado	96,13%

Fonte: O autor

- Em Tweets Classificados Positivos:

Figura 27 – Tabela de resultados de assertividade de em Tweets Classificados Positivos

Tipo de Dicionário	% Assertividade com a classificação humana
Sentilex Original	59,18%
Sentilex Modificado	83,67%

Fonte: O autor

- Em Tweets Classificados Negativos:

Figura 28 – Tabela de resultados de assertividade de em Tweets Classificados Negativos

Tipo de Dicionário	% Assertividade com a classificação humana
Sentilex Original	65,78%
Sentilex Modificado	91,05%

Fonte: O autor

Conclusão

As redes sociais têm se tornado um dos principais meios de discussões de ideias e acesso a informações de forma dinâmica e ágil. Desde de o início da pandemia da covid-19 esse debate nas redes tem se tornado cada vez mais intenso. Com isso, analisar o conteúdo gerado nessas redes pode nos dar uma percepção da opinião pública em relação a esse tema tão importante para sociedade.

Este trabalho desenvolveu uma pesquisa utilizando o método de análise de sentimentos em dados retirados do twitter relacionados aos imunizantes da doença da covid-19 comercializados no Brasil, com o objetivo de identificar padrões de tweets (positivos, negativos ou neutros), além da proposta de realizar melhorias no dicionário léxico escolhido, com finalidade de trazer uma maior assertividade na classificação de polaridade. A coleta dos dados utilizados na melhora do dicionário léxico sentilex, foi realizada em todo mês de agosto de 2022, Já a amostra de 500 tweets coletados, utilizados para validação e comparação da assertividade dos dicionários (Sentilex e Sentilex Modificado), foi realizada na semana 07/11/2022.

A partir desse estudo podemos perceber que a maior parte dos tweets presentes na amostra de validação e comparação, apresentou sentimentos neutros sobre os imunizantes da covid-19 no brasil, em todos os recortes utilizados na classificação de polaridade (Sentilex original, Sentilex modificado e Classificação manual). A comparação entre os dicionários utilizados como base de análise de sentimento, demonstrou uma maior semelhança de comportamento entre os cenários Sentilex Modificado e Classificação manual, trazendo um cenário de impacto positivo nas alterações realizadas no dicionário em questão.

Trabalhos futuros podem mesclar os tipos de abordagem, dicionários léxicos e aprendizagem de máquina, utilizando de diversos algoritmos explorar com maior profundidade os tweets classificados como negativos, separando por temas, por exemplo. Como o trabalho de (*Jenhani et al.*, 2016) que realizou uma abordagem híbrida utilizando tanto regras linguísticas quanto técnicas de aprendizagem de máquina. No seu trabalho, o grupo utilizou um gramatical chamado ODIN como um mecanismo linguístico e consumiu o material classificado pelas regras linguísticas para construir o material de treinamento usado no processo de aprendizagem de máquina.

Já *Yousefinaghania et al.* (2021) onde seu trabalho também teve como objetivo a utilização do método de análise de sentimento como mecanismo de entendimento das opiniões públicas de diversas nacionalidades em relação às vacinas COVID-19, tendo como base de dados o conteúdo do Twitter. Na sua pesquisa o grupo utilizou como tipo

de abordagem para classificação de polaridade um léxico em Python chamado Valence Aware Dictionary e Sentiment Reasoner (VADER) e posteriormente reclassificou os tweets polarizados em novos grupos (antivacina, hesitantes e pró-vacina e tweets neutros) randomicamente.

A compreensão da opinião pública sobre a vacinação usando mineração de opinião de posts realizados na plataforma twitter pode ajudar agências de saúde a aumentar mensagens positivas e eliminar mensagens negativas relacionadas aos imunizantes através de material informativo, a fim de melhorar a captação de vacinas.

Referências

Ansari, M. Z., Aziz, O. Siddiqui, H. Mehra, and K. P. Singh, Analysis of political sentiment orientations on twitter, p. 1821–1828, 2020. Citado na página 33.

Barbosa, J. A., A aplicabilidade da tecnologia na pandemia do novo coronavírus (covid-19), pp. 48–52, 2020. Citado na página 32.

BBC, Coronavírus: o que é distanciamento social e como ele pode reduzir (e muito) o número de infectados, *Tech. rep.*, evolucional, 2020. Citado 2 vezes nas páginas 30 e 31.

Becker, and Tumitan, Introdução à mineração de opiniões: Conceitos, aplicações e desafios., 2013. Citado 3 vezes nas páginas 14, 19 e 22.

Benevenuto, F., F. Ribeiro, M. Araujo, ssfsfsfs, and ojdadojad, Métodos para análise de sentimentos em mídias sociais, 2018. Citado 5 vezes nas páginas 17, 21, 22, 23 e 24.

Braun, D., Brasil tem a quarta maior base de usuários do twitter no mundo., *Tech. rep.*, Valor Investe, 2022. Citado na página 14.

Carvalho, P., and M. J. Silva, Empresas orientadas a dados e análises: a tecnologia está a serviço da tomada de decisão?, 2015. Citado 2 vezes nas páginas 39 e 40.

Dhein, W., O. Neto, C. Carvalho, and oandoa, Pandemia e o consumo de notícias nas redes sociais, *Tech. rep.*, Gente Globo, 2022. Citado na página 25.

Duarte, P. M., Covid-19: Origem do novo coronavírus, pp. 3585–3590, 2020. Citado na página 28.

Fiocruz, Observatório covid-19, *Tech. rep.*, Fiocruz, 2021. Citado na página 32.

Freaza, V., Data-driven education: Entenda as práticas, *Tech. rep.*, evolucional, 2018. Citado na página 19.

Gilliam, W. S., A. A. Malik, M. Shafiq, and M. Klotz, Transmissão covid-19 em programas de cuidados infantis nos eua, 2020. Citado na página 28.

Guimarães, P., and C. Rodrigues, 4 em cada 10 brasileiros afirmam receber fake news diariamente, *Tech. rep.*, CNN, 2022. Citado na página 13.

Henrique, C., R. Teixeira, anihadi, and jidfjdi, EvoluÇÃo das comunicaÇÕes atÉ a internet das coisas: A passagem para uma nova era da comunicaÇÃo human, 2021. Citado 2 vezes nas páginas 25 e 26.

Huerta, D. T., J. B. Hawkins, J. S. Brownstein, iadjnfojaf, and oNANDJO, Exploring discussions of health and risk and public sentiment in massachusetts during covid-19 pandemic mandate implementation: A twitter analysis, p. 100851, 2021. Citado na página 35.

Islam, M. F., J. Cotler, L. A. Jason, and sfnjsojs, Post-viral fatigue and covid-19: lessons from past epidemics, pp. 61–69, 2020. Citado na página 29.

Jenhani, F., M. S. Gouider, and L. B. Said, A hybrid approach for drug abuse events extraction from twitter, p. Sunitha, 2016. Citado na página 53.

Logghe, H. J., M. A. Boeck, y. MPH, and S. B. Atallah, Decoding twitter: Understanding the history, instruments, and techniques for success, pp. 904–908, 2016. Citado na página 27.

Lourenco, L., and T. Lontra , Google colab: saiba o que é essa ferramenta e como usar! betrybe, *Tech. rep.*, betrybe, 2020. Citado na página 41.

Lourenco, L., and T. Lontra , O brasileiro ama redes sociais, *Tech. rep.*, Gente Globo, 2022. Citado 2 vezes nas páginas 26 e 27.

Maurosso, B., E. Marinete, A. M. Almeida, osfosf, and aooadn, Análise de sentimentos/-mineração de opinião: Uma revisão bibliográfica, pp. 80–99, 2017. Citado 2 vezes nas páginas 23 e 24.

Naseem, U., I. Razzak, M. Khushi, and P. W. E. e Jinman Kim, Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis, p. 1003–1015, 2021. Citado 2 vezes nas páginas 33 e 34.

Nezhad, Z. B., and M. A. Deihimi, Twitter sentiment analysis from iran about covid 19 vaccine, 2022. Citado 2 vezes nas páginas 16 e 17.

Petr, W. A., X. Zhang, R. Nir-Paz, and aojaosd, Doença do coronavírus 2019 (covid-19)., 2020. Citado 3 vezes nas páginas 28, 29 e 30.

Pezzini, A., MineraÇÃo de textos: Conceito, processo e aplicaÇÕes, pp. 01–13, 2016. Citado na página 19.

Qorib, M., T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on covid-19 vaccination twitter dataset, p. 118715, 2023. Citado na página 40.

Resende, I., and N. H. Alpaca, Estagnação da vacinação contra covid ameaça combate à doença, aponta fiocruz, *Tech. rep.*, CNN, 2022. Citado na página 15.

Sampaio, A. G., AnÁlise de sentimentos, 2021. Citado 5 vezes nas páginas 18, 19, 20, 21 e 22.

Sarsam, S. M., H. Al-Samarraie, A. I. Alzahrani, W. Alnumay, and A. P. Smith, A lexicon-based approach to detecting suicide-related messages on twitter, pp. 1746–8094, 2021. Citado na página 34.

Shahriar, K. T., M. N. Islam, M. M. Anwar, and I. H. Sarker, Covid-19 analytics: Towards the effect of vaccine brands through analyzing public sentiment of tweets, p. 100969, 2022. Citado na página 17.

Shereen, M. A., S. Khan, A. Kazmi, N. Bashir, and R. Siddique, Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses, pp. 91–98, 2020. Citado na página 28.

Siru, L., and L. Jialin, Public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis, pp. 5499 – 5505, 2021. Citado na página 16.

Soares, D. J., Empresas orientadas a dados e análises: a tecnologia está a serviço da tomada de decisão?, 2017. Citado 2 vezes nas páginas 18 e 19.

Sumitro, P. A., Rasiban, D. I. Mulyana, and W. Saputro, Analisis sentimen terhadapat vaksin covid-19 di indonesia pada twitter menggunakan metode lexicon based, pp. 50–56, 2021. Citado na página 35.

Sunitha, D., P. Raj, N. Babu, A. Suresh, and G. Suresh, Twitter sentiment analysis using ensemble based deep learning model towards covid-19 in india and european countries., pp. 164 – 170, 2022. Citado 3 vezes nas páginas 14, 21 e 40.

Team, E., Qualitative vs. quantitative data: what's the difference?, *Tech. rep.*, fullstory, 2022. Citado na página 18.

Viana, T., Dado, informação, conhecimento e competência, 2014. Citado na página 18.

Viteri, S. B. A., Análisis de sentimientos para twitter con vader y textblob, pp. 2697–3405, 2021. Citado na página 34.

Xu, H., R. Liu, Z. Luo, and M. Xu, Covid-19 vaccine sensing: Sentiment analysis and subject distillation from twitter dat, p. 100016, 2022. Citado na página 33.

Yousefinaghania, S., D. R, M. S., P. A., and S. S, An analysis of covid-19 vaccine sentiments and opinions on twitter, 2021. Citado 3 vezes nas páginas 13, 15 e 53.