

**UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
INSTITUTO DE CIÊNCIA, ENGENHARIA E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA, AMBIENTE E
SOCIEDADE**

Franciele Leal Farias

**APLICAÇÃO DE MINERAÇÃO DE TEXTOS E ANÁLISE DE SENTIMENTOS A
POSTAGENS DO TWITTER ACERCA DAS VACINAS CONTRA A COVID -19**

Teófilo Otoni - MG

2022

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
INSTITUTO DE CIÊNCIA, ENGENHARIA E TECNOLOGIA

Franciele Leal Farias

APLICAÇÃO DE MINERAÇÃO DE TEXTOS E ANÁLISE DE SENTIMENTOS A
POSTAGENS DO TWITTER ACERCA DAS VACINAS CONTRA A COVID -19

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologia, Ambiente e Sociedade da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM como requisito parcial para a obtenção do título de Mestre.

Orientadora: Dra. Lorena Sophia Campos de Oliveira

Teófilo Otoni - MG

2022

Catálogo na fonte - Sisbi/UFVJM

F224a Farias, Franciele Leal
2022 APLICAÇÃO DE MINERAÇÃO DE TEXTOS E ANÁLISE DE SENTIMENTOS
A POSTAGENS DO TWITTER ACERCA DAS VACINAS CONTRA A COVID -19
[manuscrito] / Franciele Leal Farias. -- Teófilo Otoni, 2022.
68 p. : il.

Orientador: Prof. Lorena Sophia Campos de Oliveira.

Dissertação (Mestrado Profissional em Tecnologia, Ambiente e Sociedade) -- Universidade Federal dos Vales do Jequitinhonha e Mucuri, Programa de Pós-Graduação em Tecnologia, Ambiente e Sociedade, Teófilo Otoni, 2022.

1. Inteligência Artificial. 2. Mineração de Textos. 3. Análise de sentimentos. 4. Pandemia Covid 19. I. Oliveira, Lorena Sophia Campos de. II. Universidade Federal dos Vales do Jequitinhonha e Mucuri. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFVJM com os dados fornecidos pelo(a) autor(a).
Este produto é resultado do trabalho conjunto entre o bibliotecário Rodrigo Martins Cruz/CRB6-2886 e a equipe do setor Portal/Diretoria de Comunicação Social da UFVJM


FRANCIELE LEAL FARIAS

Aplicação de Mineração de Textos e Análise de Sentimentos a postagens do Twitter acerca das vacinas contra a Covid-19.


Dissertação apresentada ao
MESTRADO EM TECNOLOGIA,
AMBIENTE E SOCIEDADE, nível de
MESTRADO como parte dos requisitos
para obtenção do título de MESTRA
EM TECNOLOGIA, AMBIENTE E
SOCIEDADE

Orientador (a): Prof.^a Dr.^a Lorena
Sophia Campos De Oliveira


Data da aprovação : 26/09/2022

Documento assinado digitalmente
 LORENA SOPHIA CAMPOS DE OLIVEIRA
Data: 28/09/2022 13:24:58-0300
Verifique em <https://verificador.iti.br>

Prof.Dr.^a LORENA SOPHIA CAMPOS DE OLIVEIRA - UFVJM


Documento assinado digitalmente
 ALEXANDRE FAISSAL BRITO
Data: 03/10/2022 21:33:50-0300
Verifique em <https://verificador.iti.br>

Prof.Dr. ALEXANDRE FAISSAL BRITO - UFVJM

Documento assinado digitalmente
 RAQUEL DE SOUZA POMPERMAYER
Data: 08/10/2022 20:19:12-0300
Verifique em <https://verificador.iti.br>

Prof.Dr.^a RAQUEL DE SOUZA POMPERMAYER - UFVJM

Prof.Dr. FÁBIO SILVA DE SOUZA - UFVJM

Documento assinado digitalmente
 FÁBIO SILVA DE SOUZA
Data: 03/10/2022 20:34:50-0300
Verifique em <https://verificador.iti.br>

TEÓFILO OTONI

“Porque dele, e por Ele, e para Ele são todas as coisas; glórias, pois a Ele eternamente. Amém”

Romanos 11.36

AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me dado a oportunidade de estar neste mestrado e por ter me sustentado com suas mãos em todos esses anos.

A minha família por ter me apoiado durante este período, especialmente a minha mãe que sempre foi e será a minha fortaleza em todos os momentos. Esse mestrado não é meu e sim nosso, mamãe.

A minha orientadora Dra. Lorena Sophia, muito obrigada por confiar em mim, até nos momentos em que nem eu mesma acreditava que conseguiria.

Aos meus colegas de trabalho e de universidade por estarem sempre me motivando, incentivando e ajudando a prosseguir nesta caminhada, em especial a José Antônio que foi meu suporte e auxiliador em todos os momentos.

A Universidade Federal dos Vales do Jequitinhonha e Mucuri pela oportunidade de realizar esse sonho.

Agradeço ainda a cada pessoa que de alguma maneira possa ter contribuído para a conclusão desse processo. A todos vocês o meu muito obrigada.

RESUMO

A pandemia da Covid-19 é o maior problema sanitário do século XXI e já ceifou a vida de milhares de pessoas. A rapidez com que a doença se espalhou e modificou a vida da população mundial gerou uma grande quantidade de emoções e sentimentos nas pessoas. Desde a descoberta do novo coronavírus, iniciou-se uma corrida pelo desenvolvimento de uma vacina que fosse eficaz para o combate da doença, crescendo o anseio da população pela sua chegada. O trabalho realiza a análise dos sentimentos que a população brasileira desenvolveu em relação às vacinas criadas para o combate da Covid-19, por meio da utilização das técnicas de análise de sentimento e mineração de dados. A construção do banco de dados ocorreu através da captação de postagens públicas disponibilizadas pela API do *Twitter*. O algoritmo desenvolvido durante a pesquisa é baseado na linguagem de programação *Python* e implementado na plataforma *Jupyter Notebook*. O processo de análise de sentimentos foi realizado através da análise semântica, com uso do dicionário de léxicos para a língua portuguesa SentiLex-PT.

Palavras-chave: Pandemia, vacina, Covid-19, mineração de textos, análise de sentimentos.

ABSTRACT

The Covid-19 pandemic is the biggest health problem of the 21st century and has already claimed the lives of thousands of people. The speed with the disease spread and changed the lives of the world's population generated a lot of emotions and feelings in people. Since the discovery of the new coronavirus, a race began to develop a vaccine that would be effective to combat the disease, growing the population's desire for its arrival. The work analyzes the feelings that the Brazilian population has developed in relation to vaccines created to combat Covid-19, through the use of sentiment analysis and data mining techniques. The construction of the database took place through the capture of public posts made available by the Twitter API. The algorithm developed during the research is based on the Python programming language and implemented on the Jupyter Notebook platform. The sentiment analysis process was carried out through semantic analysis, using the lexicon dictionary for the Portuguese language SentiLex-PT.

Keywords: Pandemic, vaccine, Covid-19, text mining, sentiment analysis.

SUMÁRIO

1 INTRODUÇÃO	11
2 REVISÃO DE LITERATURA	13
2.1 ESTADO DA ARTE	13
2.2 BANCO DE DADOS	14
2.2.1 <i>Modelo Relacional</i>	15
2.2.2 <i>Modelo Não Relacional</i>	16
2.3 MINERAÇÃO DE TEXTOS.....	17
2.3.1 <i>Abordagens dos dados</i>	18
2.3.1.1 <i>Análise Semântica</i>	19
2.3.1.2 <i>Análise Estatística</i>	19
2.3.2 <i>Áreas de conhecimento em Mineração de Textos</i>	20
2.3.2.1 <i>Processamento de Linguagem Natural</i>	20
2.3.2.2 <i>Recuperação de Informação</i>	20
2.3.2.3 <i>Ciência Cognitiva</i>	21
2.3.2.4 <i>Mineração de Dados</i>	21
2.3.2.5 <i>Estatística</i>	21
2.3.2.6 <i>Aprendizado de Máquina</i>	22
2.3.3 <i>Etapas do Processo de Mineração de Textos</i>	22
2.3.3.1 <i>Seleção</i>	22
2.3.3.2 <i>Pré-Processamento</i>	23
2.3.3.2.1 <i>Tokenização</i>	23
2.3.3.2.2 <i>Remoção de Stopwords</i>	24
2.3.3.2.3 <i>Normalização</i>	24
2.3.3.3 <i>Transformação</i>	25
2.3.3.3.1 <i>Representação de documentos</i>	25
2.3.3.3.2 <i>Bag of Words</i>	26
2.3.3.3.3 <i>Consultas</i>	27
2.3.3.4 <i>Mineração</i>	27
2.3.3.4.1 <i>Naive Bayes</i>	27
2.3.3.4.2 <i>Clusterização</i>	27
2.3.3.4.3 <i>Sumarização</i>	28
2.3.3.5 <i>Avaliação dos Resultados</i>	28
2.4 A PANDEMIA DO COVID-19.....	29

2.4.1 <i>O Brasil</i>	29
2.4.2 <i>Vacinas</i>	31
3 MATERIAIS E MÉTODOS	34
3.1 METODOLOGIA DE PESQUISA	34
3.2 LINGUAGEM DE PROGRAMAÇÃO <i>PYTHON</i>	34
3.2 AMBIENTE DE DESENVOLVIMENTO INTEGRADO.....	35
3.3 O <i>TWITTER</i>	35
3.4 ANÁLISE DE SENTIMENTOS.....	37
3.4.1 <i>Granularidade da análise</i>	39
3.4.2 <i>Léxico de opiniões</i>	39
3.5 MÉTRICAS DE VALIDAÇÃO	40
4 RESULTADOS E DISCUSSÕES.....	42
4.1 ALGORITMO DE COLETA DE DADOS.....	42
4.1.1 <i>Dados de Teste</i>	42
4.1.2 <i>Dados da pesquisa</i>	43
4.2 ALGORITMO DE ANÁLISE DE SENTIMENTOS	44
5 CONSIDERAÇÕES FINAIS.....	52
REFERÊNCIAS BIBLIOGRÁFICAS	53
ANEXO A – ARTIGO PUBLICADO.....	57

1 INTRODUÇÃO

A pandemia da Covid 19 é considerada o maior desafio do século 21. Um vírus novo com uma grande taxa de transmissibilidade que não permitiu que medidas de contenção efetivas fossem descobertas antes que este se alastrasse por todos os países do mundo, fazendo milhares de vítimas fatais. De tal modo, a pandemia acarretou problemas nas esferas social, política, econômica e educacional, gerando um aglomerado de sentimentos e emoções na população. Dentro desse contexto mundial, o Brasil se encontra atualmente incluído não apenas em um ambiente hostil gerado pela doença como também tem sofrido com uma instabilidade gerada por crises políticas e governamentais (SOUZA et al., 2020).

Inseridos nesse cenário caótico, se encontram milhares de brasileiros que não possuem segurança para desempenhar suas tarefas diárias, bem como muitas vezes não confiam nas informações que lhes são passadas pela imprensa, gerando assim problemas psicológicos não apenas nas pessoas que se encontram na linha de frente do combate a Covid 19, mas também naquelas que não possuem a certeza se conseguirão manter seu emprego, se terão como sustentar sua família ou ainda se irão ter atendimento médico adequado caso seja necessário.

Durante este mesmo período de tempo, onde o caos entre as pessoas era disseminado, cientistas e estudiosos desenvolviam diversos estudos científicos em busca de uma vacina eficaz no combate ao novo coronavírus. O fato de quatro dessas pesquisas terem sido realizadas no Brasil aumentava nos brasileiros a proporção das emoções que eram produzidas e compartilhadas diariamente através das redes sociais (CASTRO, 2021).

Nos últimos anos as redes sociais tem aumentado a sua adesão, tornando-se algumas das principais plataformas de comunicação utilizadas para compartilhar informações e para que as pessoas possam se expressar (SILVA; MALHEIROS, 2019). Dentro do contexto da pandemia e do período de isolamento social durante o lockdown realizado em quase todos os países, o volume de dados produzidos através das redes sociais aumentou significativamente e as tornaram preciosas fontes de dados a serem analisados.

Ao se considerar a grande quantidade de emoções produzidas na população, tem se a análise de sentimentos, uma área da ciência da computação, que se tornou de grande popularidade nos trabalhos em que se desejava encontrar opiniões e sentimentos que poderiam ser associados a produtos, personagens ou lugares, e que fossem expostos em mídias sociais. Porém, atualmente, esta área pode ser aplicada a trabalhos de âmbito social e que beneficiem a população (RAMOS; FREITAS, 2019).

Para realizar tal trabalho, a análise de sentimentos está sendo associada a mineração de textos, um conjunto de métodos multidisciplinares utilizados para extrair regularidades, encontrar tendências ou padrões em dados textuais, identificando assim informações úteis que dificilmente seriam descobertas utilizando métodos tradicionais de consulta (MORAIS; AMBROSIO, 2007).

Dentro de tais contextos apresentados, o presente trabalho tem por objetivo realizar a análise dos sentimentos que a população brasileira desenvolveu em relação às vacinas criadas para o combate da Covid-19, por meio da utilização das técnicas de análise de sentimento e mineração de dados. Para realizar tal análise, a construção do banco de dados ocorreu através da captação de postagens públicas disponibilizadas pela API do Twitter durante o período que abrange 03 de novembro de 2020 a 20 de fevereiro de 2021.

É importante destacar que foi construído um banco de dados que represente uma amostra aleatória das postagens realizadas neste período de tempo, não compreendendo a totalidade das postagens realizadas. Dentro deste período se encontram a data de aprovação e aplicação da primeira vacina contra Covid-19 no mundo e no Brasil.

2 REVISÃO DE LITERATURA

Primeiramente serão apresentados alguns trabalhos que apresentam abordagens semelhantes a realizada nesta pesquisa e que obtiveram resultados satisfatórios. Na sequência serão apresentadas as revisões bibliográficas referentes as técnicas utilizadas: bancos de dados, mineração de textos e análise de sentimentos. Ainda será expressa uma contextualização quanto a pandemia do Covid-19.

2.1 Estado da Arte

Silva e Malheiros (2019) apresentaram um trabalho que possui por objetivo a coleta e classificação de postagens compartilhadas publicamente no *Twitter*, gerando um conjunto de dados para a análise de sentimentos. Para atingir tal objetivo foi desenvolvido um coletor de mensagem na linguagem *Python*, utilizando a *Application Programming Interface* (API) do *Twitter*. Na sequência foi construída uma aplicação web em Java Script para que pessoas voluntárias ao trabalho pudessem realizar a classificação das mensagens coletadas em: positivo, negativo e neutras.

A pesquisa obteve um banco de dados final com 2.787 tweets sendo que 888 foram classificados como positivos, 881 como negativos e 1.018 como neutros. Realizaram-se ainda testes com classificadores de aprendizagem de máquina treinados com este banco de dados onde o melhor resultado foi 0,4503 de acurácia média encontrada pelo *LogisticRegression* (SILVA; MALHEIROS, 2019).

Figueiredo et al (2018) efetuou uma revisão sistemática que tem por objetivo identificar e classificar as principais técnicas que eram utilizadas em outros trabalhos para realizar a análise de sentimentos, ou seja, expressar e analisar emoções contidas em comentários de redes sociais. A construção do banco de dados foi feita através de uma pesquisa em cinco bases científicas que retornaram 628 trabalhos relacionados.

Após a coleta inicial, os trabalhos foram classificados de acordo com os critérios de seleção pré-definidos pelos autores, como: idioma em que o texto foi escrito, se o trabalho utilizava técnicas de mineração de texto aplicadas a redes sociais, se o trabalho não havia sido publicado pelo mesmo autor em mais de uma base sendo assim contado mais de uma vez, entre outros critérios. Ao concluir esta análise 53 trabalhos foram considerados adequados para o estudo e 575 não adequados. Como resultado dessa revisão obteve-se que os principais

classificadores utilizados na aplicação da análise de sentimentos neste conjunto de trabalhos são o *Support Vector Machines* (SVM), o Naive Bayes e o Máxima Entropia. (FIGUEIREDO et al., 2018).

Junqueira e Fernandes (2018) apresentaram um trabalho onde foi realizada uma comparação entre alguns algoritmos de aprendizagem de máquina e os dicionários de léxicos mais utilizados na literatura, utilizando como base de dados as postagens em português realizadas no Twitter acerca das Olimpíadas de 2016.

Tais autores apresentam em seu trabalho a dificuldade de análise dos textos escritos e português devido à riqueza da linguagem e a complexidade gramatical que esta apresenta. Nesta pesquisa o algoritmo que apresentou melhor desempenho foi o SVM atingindo 89,5% de acurácia. (JUNQUEIRA; FERNANDES, 2018).

Olenski et al (2020) realizaram um estudo sobre a opinião pública dos brasileiros a respeito da inclusão de medicamentos no tratamento da Covid-19, como a cloroquina e a hidroxicloroquina. Para atingir o objetivo inicial foi realizada a coleta de aproximadamente 70 mil frases utilizando palavras-chaves relacionadas ao tema como: ‘covid19’, ‘corona’, ‘cloroquina’, entre outras, durante um período de tempo pré-determinado.

Para realizar a construção do algoritmo foi realizada a construção de uma base de treino com 2000 postagens, sendo estas classificadas manualmente. O classificador de aprendizagem de máquina que se mostrou mais eficaz neste banco de dados foi o *Support Vector Classifier* (SVC) com 56,25% de acurácia média. Este trabalho possuía ainda como um objetivo secundário, demonstrar uma aplicação útil as atividades de vigilância em saúde brasileiras, extraíndo informações relevantes de maneira eficiente através de redes sociais (OLENSCKI et al., 2020).

2.2 Banco de Dados

Os bancos de dados têm se tornado ferramentas de utilização frequente nas atividades cotidianas de diversas pessoas devido a grande quantidade de dados disponíveis tanto nas organizações quanto na *Web*. Entretanto, é necessário salientar a vasta gama de modelos disponíveis no mercado, possuindo cada um deles especificações e aplicabilidades que se adequem ao local onde será implementado (SOUZA; SANTANA, 2017). Dentro destes modelos, pode-se destacar o uso do modelo relacional e do modelo não relacional.

2.2.1 Modelo Relacional

O modelo relacional (MR) de bancos de dados foi proposto pelo matemático Ted Codd, em 1970, sendo um modelo baseado nos conceitos matemáticos da teoria dos conjuntos e lógica de predicado (SOUZA; SANTANA, 2017).

A estrutura fundamental deste banco de dados é a relação. Os dados são organizados, ou estruturados, em tabelas onde as colunas representam um ou mais atributos dos dados, e as linhas representam os registros ou ocorrências do conjunto de dados (SOUZA; SANTANA, 2017). Assim sendo, nesse modelo a tabela é apresentada como uma coleção dessas relações. A Linguagem de Consulta Estruturada ou *Structured Query Language* (SQL) é considerada como linguagem padrão para Sistemas Gerenciadores de Bancos de Dados Relacionais (SGBDR).

A Figura 01 apresenta um modelo simples de tabela onde é possível verificar a estruturação de dados e são demonstradas as relações entre linhas e colunas. As colunas representam os atributos que foram coletados e serão analisados referente a um determinado indivíduo, bem como cada linha marca o registro de um indivíduo diferente.

Figura 01. Modelo de tabela

Individuo	Idade (anos)	Altura (metros)	Peso(kg)
01	23	1.75	71
02	25	1.68	73
03	30	1.83	82

Fonte: Autoria própria

A utilização do modelo relacional encontra alguns problemas devido ao crescimento acelerado do número de recursos referente aos sistemas computacionais. A principal dificuldade ocorre na necessidade de conciliar o tipo de modelo com a demanda crescente de escalabilidade, pois esta possui um limite nesse modelo (GARCIA; SOTTO, 2019).

2.2.2 Modelo Não Relacional

O modelo de banco de dados não relacional (NoSQL) permite o uso de diversos modelos de dados como gráficos, documentos e chave-valor. Esse modelo apresenta maior performance e flexibilidade para atender as características específicas de cada organização, além de possuir grande escalabilidade em seu sistema (GARCIA; SOTTO, 2019). Outras características que podem ser apresentadas é que tal modelo possui facilidade de desenvolvimento, desempenho escalável e resiliência. Este pode ainda ser dividido em quatro categorias básicas: o baseado em armazenamento chave-valor, o orientado a documentos, o orientado a coluna e o que é orientado a grafos.

Segundo Souza e Santana (2017), o modelo não relacional orientado a documentos forma uma categoria adequada para utilizar em aplicações da Web, pois estes são projetados para armazenar os dados em forma de documentos e não exigem que tais dados possuam relações ou estruturas pré-definidas. Neste modelo de banco de dados não existem tabelas e sim uma coleção de documentos, onde o programador armazena os dados em um determinado documento e no formato que lhe seja necessário. Porém, destaca-se o fato de que o programador deve estar atento para que ele próprio não se perca no meio do seu banco.

Um grande diferencial do NoSQL é o fato deste não possuir um diagrama de como o banco deve ser construído, ou quando possui tal diagrama, este é vagamente definido, o que é muito útil quando se trata de aplicações Web, onde se armazena variados tipos de conteúdo que evoluem no decorrer do tempo (GARCIA; SOTTO, 2019).

O modelo não relacional possui diversas vantagens quando comparado ao modelo relacional, contudo, antes de optar pela sua escolha é necessária uma análise mais rigorosa quanto aos dados que se deseja arquivar. Por exemplo, uma situação onde os dados estejam apresentados de forma que exista uma estruturação já definida de grande relevância para a análise a ser executada, seria mais adequada a utilização de um banco relacional (SOUZA; SANTANA, 2017).

2.3 Mineração de Textos

A Mineração de Textos (MT), ou *text mining*, é considerada pelos estudiosos como um Processo de Descoberta de Conhecimento, onde são utilizadas técnicas de análise e extração de dados a partir de documentos textuais como textos, frases ou apenas palavras. A mineração de textos é um método multidisciplinar que utiliza técnicas de Estatística, Informática, Programação e linguística para processar textos identificando informações úteis que estavam implícitas nos textos e que dificilmente seriam descobertas utilizando métodos tradicionais de pesquisa (MORAIS; AMBROSIO, 2007).

O principal objetivo da mineração de texto é encontrar termos relevantes em documentos textuais com grande volume de dados e assim estabelecer padrões e relacionamentos entre eles. Tais dados se encontram arquivados de forma estruturada, não estruturada ou semi estruturada. É importante destacar que a mineração de textos não é um mecanismo de busca, como os encontrados na internet, pois a mineração auxilia o usuário a descobrir informações que antes eram desconhecidas. Já nos mecanismos de busca, o usuário já sabe o que eles desejam procurar (PEZZINI, 2016).

Existem duas abordagens principais para o processo de mineração em dados textuais que são: a Análise Estatística, que está diretamente ligada a frequência de aparição de um determinado termo, não se atentando ao contexto em que este se encontra inserido no texto; e a Análise Semântica, que se preocupa com a funcionalidade dos termos nos textos por meio dos significados morfológico, sintático, semântico, entre outros que podem ser empregados no contexto da pesquisa realizada. Tais abordagens podem ser aplicadas sozinhas ou em associação (BRITO, 2017).

A aplicação inicialmente é realizada em um grupo teste para que o algoritmo seja treinado, e se o mesmo não alcançar um resultado satisfatório, as etapas devem ser repetidas até que o objetivo seja alcançado. Após o algoritmo se encontrar apto, este pode ser aplicado aos dados que se deseja analisar (BRITO, 2017).

A Figura 02 apresenta as etapas do processo de mineração de texto segundo o trabalho de Brito (2017), em formato de diagrama para facilitar a visualização das mesmas. É necessário salientar que cada uma das etapas possui as suas técnicas específicas para serem executadas, e serão explicadas individualmente na sequência deste trabalho.

Seleção – Se trata da extração e coleta dos dados que serão utilizados na pesquisa, criando uma base de dados textuais que represente toda a população que será analisada, ou seja, a criação do *corpus* da pesquisa.

Pré-processamento – Consiste na filtragem e limpeza dos dados, eliminando informações desnecessárias ao algoritmo.

Transformação – Se trata da classificação dos dados, separando-os de acordo com as características que se pretende observar.

Mineração – Esta etapa está diretamente ligada as técnicas de aprendizagem de máquina para obtenção de novos conhecimentos.

Avaliação – É a etapa onde será validada a eficiência do processo como um todo, analisando os dados que foram obtidos após a aplicação dos algoritmos.

Figura 02. Diagrama do Processo de Mineração de Texto



Fonte: Adaptado de Brito (2017)

2.3.1 Abordagens dos dados

De acordo com Moraes e Ambrósio (2007) são definidos dois tipos de abordagens para análise textual no contexto da mineração de textos: a Análise Semântica e a Análise Estatística.

2.3.1.1 Análise Semântica

A Análise Semântica é realizada principalmente pelo emprego de técnicas de Processamento de Linguagem Natural (PLN), onde os termos são avaliados de acordo com o contexto no qual estão inseridos para identificar a sua função. O PNL será abordado no tópico 2.3.2.1 deste texto.

Segundo Moraes e Ambrósio (2007) para realizar a análise semântica é necessário que o programador possua alguns conhecimentos específicos a análise textual, como:

- Morfológico – É o conhecimento associado a estrutura, formação e inflexão das palavras.
- Sintático – É o conhecimento acerca da estrutura de uma lista de palavras e de como estas podem ser combinadas para produzir expressões e sentenças.
- Semântico – É o conhecimento relacionado ao significado das palavras, sem análise do contexto em que estão inseridas. Pode ser utilizado em palavras independentes ou em combinações de palavras.
- Pragmático – É o conhecimento sobre a utilização das palavras em diferentes contextos e como estes podem modificar seu significado e interpretação.

Portanto, é possível concluir que as técnicas aplicadas a análise semântica têm por objetivo identificar a importância das palavras dentro da estrutura de suas sentenças ou orações. Porém, deve-se salientar que quando se utiliza um texto mais complexo, ou um documento, algumas funções são identificadas com maior grau de importância e este fato deve ser considerado no momento da análise (CORDEIRO, 2005).

2.3.1.2 Análise Estatística

A Análise Estatística está diretamente relacionada com a quantidade de vezes que determinado termo aparece no texto. Neste modelo a importância de tal termo é dada de acordo com a sua frequência de aparição. De acordo com Moraes e Ambrósio (2007) o processo de aprendizado estatístico pode conter as seguintes etapas:

- Codificação de Dados – A codificação inicial dos dados é realizada de acordo com os critérios que representem as propriedades dos dados que são importantes para atender aos objetivos da análise.

- Estimativa dos Dados – A estimativa está relacionada a busca por um modelo que esteja adequado ao problema escolhido. Pode-se utilizar algoritmos de aprendizagem ou um método de estimativa.
- Modelo de Representação de Documentos – Nesta etapa, os documentos podem passar a ser expressos como um conjunto ou bolsa de palavras, onde as palavras estão arquivadas sem levar em consideração sua ordem de ocorrência no texto, pontuação ou estrutura, não utilizando a análise semântica do texto.

2.3.2 Áreas de conhecimento em Mineração de Textos

Na aplicação do processo de mineração de textos é necessário conhecer diversas técnicas e áreas do conhecimento. Assim sendo, algumas destas serão apresentadas a seguir de forma sucinta para auxiliar na compreensão do trabalho.

2.3.2.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural pode ser definido como um conjunto de métodos formais para analisar textos e gerar frases escritas em um idioma humano (JUNIOR, 2007). Alguns dos tópicos realizados nesta técnica são a identificação de estruturas das frases e a modelagem do conhecimento e do raciocínio

O PLN é ainda considerado uma área de pesquisa que realiza a associação da Computação e da Linguística para analisar e representar textos naturais (escritos totalmente por seres humanos), tendo por objetivo principal conseguir o processamento de linguagem de textos semelhante ao humano (OLENSCKI et al., 2020).

O PLN possui um papel fundamental na técnica de mineração de textos, sendo utilizado na etapa de pré-processamento, onde sua função é fornecer um nível inicial de estruturação da informação textual, além de classificar as palavras quanto a sua função sintática. (JUNIOR, 2007).

2.3.2.2 Recuperação de Informação

A Recuperação de Informação (RI) é uma área da computação que trabalha com o armazenamento e a recuperação automática de documentos, sendo estes geralmente, em

formato textual. As máquinas de busca como as utilizadas pelo Google, Yahoo, entre outros, utilizam sistemas de RI (JUNIOR, 2007).

A RI é considerada por alguns estudiosos como um importante passo dentro da etapa de transformação, pois possui modelos para aplicação em grandes coleções de textos, como os Sistemas de Recuperação de Informações Textuais (EBECKEN et al., 2003). Nestes sistemas, os usuários escolhem os termos, também denominados palavras-chave, mais adequados para caracterizar sua busca, o sistema seleciona e retorna ao usuário os documentos considerados relevantes de acordo com a quantidade de palavras semelhantes que os mesmos possuem com a consulta (MORAIS; AMBROSIO, 2007).

2.3.2.3 Ciência Cognitiva

A Ciência Cognitiva é definida por alguns estudiosos como o estudo científico da mente ou da inteligência. Segundo Junior (2007), tal ciência é amplamente utilizada devido a sua alta interdisciplinaridade, sendo utilizada em associação com as disciplinas de psicologia (psicologia cognitiva, neurociência e disciplinas de filosofia, entre outras).

2.3.2.4 Mineração de Dados

A Mineração de Dados, ou *Data Mining* (DM) é considerada por diversos estudiosos como uma das tecnologias mais promissoras da atualidade. Um dos fatores deste sucesso é o gasto de dezenas, e até mesmo centenas de milhões de reais pelas companhias na coleta dos dados e, no entanto, nenhuma informação útil era identificada. Com o uso da mineração de dados, as informações podem ser retiradas desses dados armazenados, gerando conhecimento para os gestores das empresas, por exemplo (CAMILO et al., 2009).

A Mineração de Textos pode ser considerada como um avanço da DM, pois a estrutura e separação das etapas são semelhantes em ambas as técnicas. Porém salienta-se o fato das técnicas e algoritmos de DM estarem presentes na etapa de Mineração do outro processo.

2.3.2.5 Estatística

A estatística é uma ciência que utiliza teorias probabilísticas para explicar experimentos, eventos e estudos. Esta ciência possui por objetivo organizar e analisar dados, de maneira a reconhecer as relações que eles apresentem. No processo de mineração de

textos, a estatística é utilizada em vários momentos por meio da aplicação de modelos probabilísticos (JUNIOR, 2007).

2.3.2.6 Aprendizado de Máquina

O Aprendizado de Máquina, ou *Machine Learning* (ML) é uma subárea da Inteligência Artificial que consiste na concepção de modelos computacionais que possam ‘aprender’ através da experiência (JUNIOR, 2007). Ou seja, o modelo computacional é submetido a um conjunto de dados teste para que possa retirar de tal conjunto regras ou padrões. Algumas das técnicas de ML muito utilizadas são as Máquinas de Vetor de Suporte (SVM) e o classificador *Naive Bayes*, sendo ambos utilizados nos trabalhos apresentados no tópico 2.1 deste trabalho.

2.3.3 Etapas do Processo de Mineração de Textos

A seguir serão apresentadas individualmente cada etapa do processo de mineração de textos, identificando as suas peculiaridades.

2.3.3.1 Seleção

A etapa inicial do processo de mineração consiste na construção da base de textos de interesse e na sua redução, selecionando um núcleo que represente o conteúdo dos textos em sua totalidade. O núcleo formado é conhecido na literatura como *Corpus* ou *Corpora*. Esta etapa promove não apenas a redução dimensional do texto, como também tenta identificar similaridades em função do significado dos termos ou da sua morfologia (EBECKEN et al., 2003).

De acordo com Junior (2007) pode-se definir três ambientes gerais para a localização das fontes de dados: pastas de arquivos que se encontrem armazenadas no disco rígido do computador do usuário, tabelas de diversos bancos de dados e a *Web*. A quantidade de dados disponíveis na Web tem crescido de maneira avassaladora, entretanto, esses dados estão disponíveis em diversas formas diferentes como *blogs*, *sites* de notícias, anúncios, planilhas e nas redes sociais, o que se torna um desafio para o processo de mineração.

Após concluir a etapa de seleção dos dados é iniciado pré-processamento.

2.3.3.2 Pré-Processamento

O Pré-processamento é a etapa onde se realiza a limpeza e a estruturação inicial dos dados. Em muitos trabalhos, pode ser considerado o processo mais oneroso da MT, visto que não existe uma única técnica que seja responsável por realizar toda a estruturação necessária. Assim sendo, para obter dados com a qualidade adequada podem ser utilizadas várias técnicas associadas (JUNIOR,2007).

A seguir serão apresentadas algumas técnicas utilizadas nesta etapa.

2.3.3.2.1 Tokenização

A Tokenização é uma técnica que possui por objetivo a extração de unidades mínimas de texto a partir de um texto livre. Nessa técnica, cada unidade é denominada token e, na maioria dos casos, corresponde a uma palavra, entretanto pode representar também um símbolo, um caractere especial ou pontuação. Existem diversas maneiras de se obter o token em um texto, sendo a mais utilizada os caracteres delimitadores e o ‘espaço’ (JUNIOR,2007). A Figura 03 apresenta o modelo de tokenização de uma frase.

Figura 03. Modelo de Tokenização

só quero acordar desse pesadelo chamado pandemia! [só] [quero] [acordar] [desse] [pesadelo] [chamado] [pandemia] [!]

Fonte: Autoria Própria

Conforme é possível visualizar na Figura 03, durante a execução desse processo o ‘espaço’ é removido do texto. A Figura 04 demonstra alguns dos caracteres especiais considerados nessa tarefa. É importante destacar que a identificação de tokens é uma tarefa muito fácil para o ser humano, porém para o algoritmo pode se apresentar como uma tarefa onerosa devido a quantidade de funções que os caracteres delimitadores podem assumir. Por exemplo, o caractere de pontuação ‘.’ pode representar o final de uma oração, como pode ser utilizado em números ou abreviações (MORAIS; AMBROSIO, 2007).

Figura 04. Caracteres Especiais

() < . > ; / ? " ' ' " { } [] * ! @ #

Fonte: Autoria Própria

2.3.3.2.2 Remoção de *Stopwords*

Em uma base de dados textual existe diversos elementos que não possuem valor semântico relevante, sendo pertinentes exclusivamente para o entendimento e a compreensão do texto. Tais elementos são classificadas como *Stopwords*, sendo estes geralmente correspondentes aos artigos, preposições, conjunções, pronome e pontuação do texto (JUNIOR,2007).

A lista composta pelo conjunto de *stopwords* é denominada *stoplist*. Alguns programas possuem uma *stoplist* já definida, entretanto o programador pode adicionar palavras a essa lista se achar necessário (MORAIS; AMBROSIO, 2007). A Figura 05 apresenta o exemplo de uma *stoplist* para ilustrar a definição e demonstrar o conteúdo das mesmas.

Figura 05. Exemplo de *Stoplist*

de	em	os	dos	tem	muito	só
a	um	no	como	à	há	pelo
o	para	se	mas	seu	nos	pela
que	é	na	foi	sua	já	até
e	com	por	ao	ou	está	isso
do	não	mais	ele	ser	eu	ela
da	uma	as	das	quando	também	entre

Fonte: Adaptado de NLTK Python (2022)

2.3.3.2.3 Normalização

A Normalização pode ser considerada uma técnica de Redução do Léxico que possui por finalidade agrupar os tokens que compartilham um padrão em comum. Para realizar esse

processo, as abordagens disponíveis podem variar desde a identificação morfológica do token até o reconhecimento de conceitos similares, os sinônimos (JUNIOR, 2007).

A aplicação dessa técnica é extrema relevância para a aplicação da MT, pois por meio dela é realizada uma redução na massa textual ao executar o agrupamento das palavras com o mesmo padrão. A abordagem mais executada nessa técnica é o *Stemming*, onde serão removidos os sufixos, os prefixos e variações morfológicas, como flexão de gênero, número e grau (MORAIS; AMBROSIO, 2007).

Segundo Ebecken et al. (2003) existem três métodos principais de *Stemming*:

- Método do *Stemmer S* – Consiste na eliminação de alguns finais de palavras, usualmente os sufixos que determinam o plural, como na língua inglesa *ies*, *es* e *s*.
- Método de Porter – Consiste na identificação de diferentes inflexões e realiza a substituição por um radical comum. Executa a remoção de cerca de 60 sufixos diferentes.
- Método de Lovins – Consiste na remoção de um sufixo por palavra, entretanto seu algoritmo possui cerca de 250 sufixos para a língua inglesa.

2.3.3.3 Transformação

A etapa de transformação dos dados textuais consiste no reconhecimento das características do documento, sua classificação e na construção de um modelo onde possa ocorrer a recuperação destas informações (MORAIS; AMBROSIO, 2007).

A seguir serão apresentados alguns pontos importantes no processo de transformação.

2.3.3.3.1 Representação de documentos

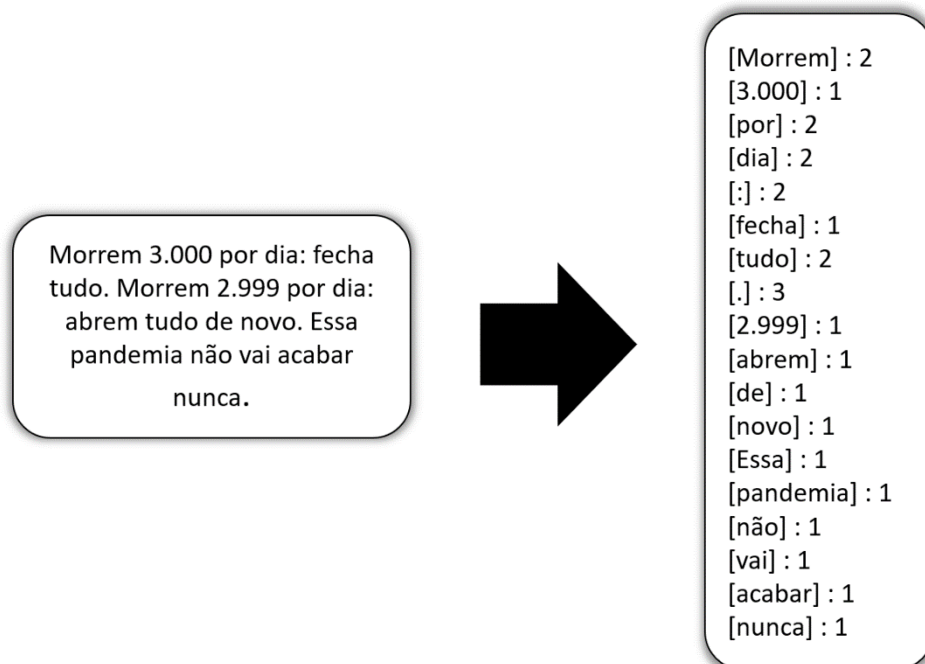
Na execução do processo de MT é imprescindível que os documentos sejam codificados de maneira que facilite sua manipulação e auxilie numa correta mensuração dos termos. Segundo Junior (2007), o Modelo de Espaço Vetorial, ou *Vector Space Model*, é um modelo de grande utilização no meio científico. Neste modelo os documentos são representados utilizando uma abstração geométrica.

De acordo com Ebecken et al. (2003) o Modelo de Espaço Vetorial possui grande utilização devido ao fato da maioria dos algoritmos da etapa de Mineração estarem aptos para lidar com este tipo de codificação. Contudo, de acordo com a quantidade de documentos que compõe o *Corpus* da pesquisa, o conjunto atingirá facilmente a marca de milhares de *tokens*.

2.3.3.3.2 *Bag of Words*

O modelo de representação de ‘saco de palavras’, ou *Bag of Words* (BoW) é amplamente utilizado em diversas aplicações de MT e aprendizado de máquina (SILVA, 2016). Neste modelo, uma coleção de documentos é apresentada como uma caixa ou saco de palavras, onde a estrutura, a ordem ou a ligação entre elas não obtêm nenhum valor para o sistema. A cada palavra é atribuído um valor de acordo com a quantidade de vezes que ela aparece no texto (JUNIOR, 2007). A Figura 06 retrata a criação de uma BoW a partir de uma frase, sendo que cada palavra recebeu o valor referente a quantidade de ocorrências. Destaca-se que neste exemplo em específico, o ‘ponto’ presente nos números não foi quantificado juntamente com o ponto que determina o final das orações.

Figura 06. Exemplo de BoW



2.3.3.3 Consultas

Uma consulta pode ser considerada uma necessidade específica apresentada por um usuário, que ao identificar que precisa de determinada informação acessa o sistema e descreve sua busca por meio de palavras-chave. A consulta utiliza das técnicas de Recuperação de Informação para retornar o resultado esperado (JUNIOR, 2007).

O poder de recuperação atribuído a um sistema se encontra exatamente na sua capacidade de atender a necessidade do usuário e retornar o resultado adequado. Para isto é importante que as consultas sejam elaboradas de forma flexível, permitindo que o usuário se expresse de maneira mais abrangente (JUNIOR, 2007).

Para considerar que uma consulta foi adequadamente executada, utiliza-se dois critérios principais: o tempo de resposta do algoritmo e a qualidade dos resultados obtidos. Ou seja, um sistema ideal retornaria os documentos mais relevantes no menor tempo possível. Entretanto, existem ainda outros fatores que podem afetar essa equação, com por exemplo a estrutura de hardware disponível para a algoritmo (JUNIOR, 2007).

2.3.3.4 Mineração

A etapa de Mineração envolve o processo de decisão acerca de quais algoritmos devem ser aplicados sobre o banco de dados. Para realizar a correta escolha do algoritmo é importante conhecer a base de dados e o qual o tipo de informação se deseja obter. A seguir serão apresentados alguns algoritmos utilizados na Mineração de Textos.

2.3.3.4.1 *Naive Bayes*

O Naive Bayes é um algoritmo baseado no Teorema de Bayes, e atualmente considerado um dos mais eficientes quando considerado o tempo de execução e a precisão em rotular novas amostras. Neste algoritmo as informações de um evento não possuem nenhuma relação com as dos demais eventos, fato que contribui para a ampla utilização, visto que este realiza uma classificação simples dos eventos (MORAIS; AMBROSIO, 2007).

2.3.3.4.2 *Clusterização*

A *Clusterização* ou agrupamento é uma tarefa que determina a organização da base de dados em grupos, denominados clusters. Para ocorrer essa divisão, os documentos são

agrupados utilizando alguma métrica que possibilite encontrar similaridades entre eles. A clusterização pode ser considerada uma técnica complexa, principalmente devido a grande quantidade de parâmetros exigidos pelos algoritmos que realizam esse processo. Esta técnica é utilizada em MT para reduzir a quantidade de tokens em uma base de dados. Deste modo, não se agrupam mais documentos e sim as palavras que o constituem (JUNIOR, 2007).

2.3.3.4.3 Sumarização

Segundo Ebecken et al. (2003), a sumarização de documentos é uma tarefa onde ocorre a redução da massa textual de maneira que se obtenha ganhos significativos no desempenho do algoritmo, contudo sem que haja perda de informações úteis ou significados-chave do texto.

De acordo com Junior (2007) a sumarização pode ser classificada como:

- Sumarização por Abstração - Se baseia na criação automática de resumos de maneira similar a realizada pelo ser humano, ou seja, considerando o entendimento do leitor.
- Sumarização por Extração – Se baseia na criação de resumos por meio da seleção de sentenças principais ou parágrafos importantes que serão inteiramente copiados do texto original. Para execução dessa técnica, outras heurísticas são utilizadas para auxiliar na detecção das sentenças importantes.

2.3.3.5 Avaliação dos Resultados

A etapa de Avaliação dos Resultados é considerada o Pós-Processamento dos Dados, sendo responsável por verificar a eficiência da aplicação dos algoritmos da etapa de mineração, avaliando assim se o objetivo foi concluído da melhor forma possível. Tal avaliação pode ocorrer de forma quantitativa, através do uso de métricas, ou qualitativa, por meio do conhecimento de especialistas no assunto. Pode-se ainda salientar que o uso de elementos gráficos, por meio de ferramentas de visualização, é considerado uma forma adequada de se analisar uma mineração de textos (JUNIOR, 2007). No tópico 3.5 serão apresentadas as métricas utilizadas neste trabalho.

2.4 A Pandemia do Covid-19

Em dezembro de 2019 uma nova espécie do coronavírus, denominada SARS-CoV-2, foi identificada pela primeira vez em Wuhan, uma cidade localizada na província de Hubei, na China. Tal espécie de vírus era responsável por desencadear uma doença respiratória agressiva, denominada Covid-9, que rapidamente se espalhou por todo o planeta tornando-se o maior problema sanitário do século XXI. O processo de transmissão do vírus ocorre de pessoa para pessoa, por meio do contato com gotículas de saliva ou superfícies contaminadas. O período de incubação do vírus pode variar de 2 a 14 dias (CARDOSO et al., 2021).

Segundo a Organização Mundial de Saúde (OMS) a doença pode se manifestar de diversas formas e níveis, sendo os principais: pessoas assintomáticas, quadros gripais leves e síndrome respiratória aguda. Deve-se ainda salientar que existe maior suscetibilidade de contágio e quadros graves da doença em pessoas idosas, acima de 60 anos de idade, e pessoas com comorbidades (OMS, 2020).

De acordo com trabalho realizado por Pimentel e Silva (2020) foram inúmeros os problemas causados pela pandemia: surtos de síndromes respiratórias graves e agudas, necessitando de tratamentos mais complexos como o uso de Unidade de Terapia Intensiva (UTI) por várias pessoas ao mesmo tempo; pouco conhecimento sobre os padrões de letalidade, infectividade, mortalidade e transmissibilidade do vírus; problemas psicológicos gerados pelo período de isolamento e quarentena; além da falta de medicamentos adequados para o tratamento da doença e demora na aprovação e aplicação das vacinas. As primeiras informações consideradas consenso entre os estudiosos foram a presença de grupos de riscos, pessoas mais propensas a desenvolverem casos graves da doença, e comportamentos que aumentavam o risco de transmissão como o contato pessoal indiscriminado e aglomerações.

2.4.1 O Brasil

Desde o aparecimento do novo coronavírus, as redes mundiais de comunicação iniciaram uma cobertura massiva sobre as informações obtidas acerca da doença divulgando cada nova descoberta e como esta poderia auxiliar no combate a expansão da pandemia. O Brasil apresentou um comportamento semelhante ao restante das redes mundiais, inclusive no que abrange a divulgação de publicações, notícias e reportagens que foram recebidas inicialmente pela população com certo tom de incredibilidade, não acreditando na gravidade

da doença. Com o decorrer do tempo e o avanço desordenado da doença, a percepção da realidade por parte da população foi alterada, favorecendo a adoção de medidas de segurança e o surgimento de novas emoções e sentimentos em relação a pandemia (PIMENTEL; SILVA, 2020).

Após a chegada do vírus ao Brasil, toda a população se encontrou obrigada a seguir uma série de medidas de prevenção, bem como o governo em todos os âmbitos, Federal, Estadual, Distrital e Municipal, passou a adotar protocolos para o enfrentamento da Covid-19. As principais ações preventivas foram instituídas internacionalmente e divulgadas pela OMS (2020) visando ocasionar a diminuição da curva epidêmica e o controle epidemiológico, sendo tais medidas:

- Cuidados sistemáticos a serem adotados com a higiene pessoal e residencial;
- Boas práticas de alimentação e condições gerais do corpo, como atividades físicas, entre outros;
- A adoção do isolamento social em determinadas condições de saúde, a quarentena em viajantes e o afastamento interpessoal;
- A utilização de máscaras de proteção individual.

De acordo com o avanço da doença e sua expansão no cenário global, as populações foram submetidas a um período de intenso temor, acarretando pensamentos irracionais e atitudes antes consideradas inimagináveis. A população brasileira além do temor pelo desconhecido e de angústia por medo da morte, enfrentava ainda um agravante causado pelas inúmeras fake News espalhadas pelos meios de comunicação e pelas redes sociais. Assim, sentimentos como ansiedade, estresse, pânico, medo, raiva, entre outros, se tornaram extremamente comuns no cotidiano da população (PIMENTEL; SILVA, 2020).

Os problemas gerados pela falta de informações precisas no Brasil se tornaram tão amplos que diversos meios de comunicação desenvolveram seus próprios painéis para acompanhar o avanço da pandemia, visto que o Ministério da Saúde não divulgava as informações dentro dos prazos que eram necessários. O Conselho Nacional de Secretários de Saúde desenvolveu o Painel CONASS de acompanhamento ao avanço da Covid 19 no Brasil, sendo que em 30 de março de 2022 já eram contabilizados quase 30 milhões de brasileiros infectados pelo vírus e 659.504 vidas perdidas para a doença (CONASS, 2022).

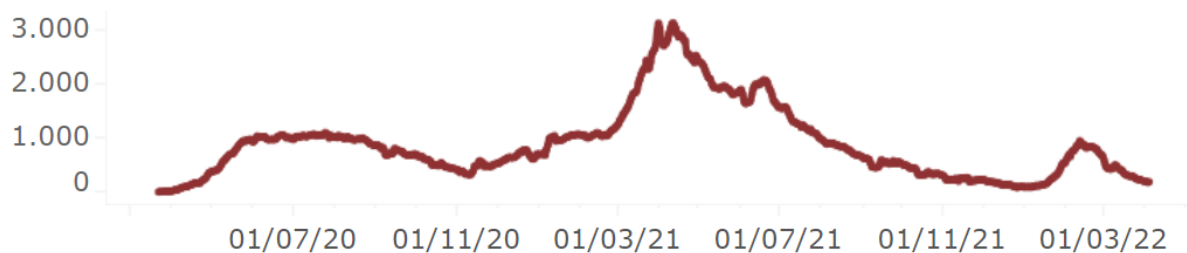
A Figura 07 apresenta a curva de avanço de casos do Covid-19 disponibilizada pelo painel CONASS desde o início da contabilização até o dia 30 de março de 2022. Já a Figura 08 apresenta a curva de óbitos contabilizados neste mesmo período.

Figura 07. Curva dos casos da Covid-19 no Brasil



Fonte: CONASS (2022)

Figura 08. Curva de mortes por Covid-19 no Brasil



Fonte: CONASS (2022)

2.4.2 Vacinas

Desde a descoberta do novo coronavírus, iniciou-se uma corrida pelo desenvolvimento de uma vacina que fosse eficaz para o combate da doença. Foram registrados junto a OMS cerca de 200 projetos de vacinas para a Covid-19. Dentre estes, quatro foram efetuados no

Brasil, fazendo com que a população estivesse familiarizada com as etapas do processo de desenvolvimento (CASTRO, 2020).

A existência dessa grande quantidade só foi possível devido ao gigantesco investimento realizado pelos governos dos países desenvolvidos, de algumas organizações não governamentais e das empresas farmacêuticas, que por muitas vezes trabalharam entre si para obter uma vacina no menor período de tempo possível (DOMINGUES, 2021).

Uma medida inédita foi adotada em relação a vacina da Covid 19: uma ação liderada pela OMS denominada *Covax Facility*, que possuía o objetivo de acelerar o desenvolvimento e a fabricação de vacinas contra a Covid-19 para garantir que todos os países tivessem acesso a elas toda a população mundial fosse vacinada (DOMINGUES, 2021).

Ainda no ano de 2020, as primeiras vacinas obtiveram autorização para uso emergencial em alguns países europeus e nos Estados Unidos. Contudo, no Brasil o uso emergencial das vacinas só foi autorizado pela Agência Nacional de Vigilância Sanitária, a ANVISA, em 17 de janeiro de 2021. Após alguns minutos a enfermeira Mônica Calazans foi a primeira brasileira a ser vacinada em território nacional (CASTRO, 2020).

Observando a vacinação quanto ao contexto internacional, é possível verificar a desigualdade no acesso e distribuição das vacinas, pois antes mesmo da primeira vacina ter o seu uso aprovado, certos países, geralmente do norte global, já haviam antecipados contratos com grupos farmacêuticos para garantir o acesso de doses de vacina para toda a sua população. Em contrapartida, neste mesmo momento os países menos desenvolvidos não possuíam contratos que garantissem doses para vacinar sequer os grupos prioritários. (CASTRO, 2020).

A processo de vacinação no Brasil encontrou diversos problemas desde seu início, como a ausência de um planejamento nacional adequado para aplicação das vacinas, falta de equipamentos como seringas e agulhas, problemas diplomáticos com lideranças de países que produziam as vacinas e os insumos necessários para a produção nacional, rivalidade política entre os políticos de esquerda e direita, além de inúmeras *Fake News* que dispersavam notícias falsas sobre as vacinas (CASTRO, 2020).

Diante de tudo que foi até aqui exposto, analisar o contexto e as emoções da população brasileira associadas a vacinação foi considerado um fato que possui potencial para produzir

conhecimento sobre este período caótico enfrentado. Para que esta análise seja executada de forma adequada, a pesquisa foi dividida em três momentos específicos, sendo eles:

- Primeiro período (03 de novembro de 2020 – 07 de dezembro de 2020) – este período compreende o momento entre a descoberta oficial da primeira vacina no mundo e um dia antes da primeira pessoa ser vacinada contra a Covid-19.
- Segundo período (08 de dezembro de 2020 – 16 de janeiro de 2021) – neste ponto é importante verificar a expectativa dos brasileiros, pois já existia uma vacina sendo aplicada pelo mundo, porém não havia a liberação da ANVISA para que o processo de vacinação fosse liberado no Brasil.
- Terceiro período (17 de janeiro de 2021 – 20 de fevereiro de 2021) – a vacinação é liberada pela ANVISA e é a primeira pessoa é vacinada no Brasil contra a Covid-19. Os momentos após essa liberação são de suma importância para serem visualizados devido as ramificações dos problemas que foram ocorrendo dia a dia neste processo.

3 MATERIAIS E MÉTODOS

3.1 Metodologia de Pesquisa

Uma pesquisa científica pode receber quatro classificações distintas de acordo com o trabalho de Gil (2017), sendo classificadas segundo sua:

- Natureza – básica ou aplicada;
- Delineamento – qualitativa ou quantitativa;
- Propósito – exploratória, descritiva, explicativa ou normativa;
- Procedimentos técnicos - Pesquisa bibliográfica, pesquisa documental, estudo de caso, levantamento, pesquisa-ação, pesquisa experimental, pesquisa participante, estudo de campo e estudo de coorte.

A pesquisa é classificada quanto a sua natureza como uma pesquisa básica pura onde será gerado conhecimento direcionado para uma determinada área, porém este conhecimento não gera aplicação imediata. Quanto ao delineamento da pesquisa, esta pode ser classificada como pesquisa de método misto devido a utilização de procedimentos quantitativos e qualitativos (GIL, 2019).

De acordo com Gil (2017), a pesquisa é classificada quanto ao seu propósito como uma pesquisa descritiva, pois possui por objetivo a descrição dos sentimentos de uma determinada população, o que pode ser considerado uma característica do grupo.

Quanto aos procedimentos técnicos, é classificada como um estudo de caso por se tratar do estudo de um fenômeno contemporâneo dentro do seu contexto real, constituindo a busca por determinado conhecimento específico através de observações de tal fenômeno (GIL, 2017). Entretanto, destaca-se ainda o fato de que foi realizada uma pesquisa bibliográfica extensa durante a execução desta pesquisa.

3.2 Linguagem de Programação *Python*

O algoritmo desenvolvido durante a pesquisa foi construído na linguagem de programação *Python*. O *Python* foi criado em 1990 por Guido van Rossum, no Instituto Nacional de Pesquisa para Matemática e Ciência da Computação da Holanda (CWI), sendo

concebido a partir de outra linguagem existente na época em questão, a ABC (BORGES, 2014).

A linguagem *Python* é simples, clara e objetiva, contudo é também muito poderosa e forte, possibilitando o seu uso em diversas aplicações. A utilização desta vem crescendo em áreas da computação como inteligência artificial, banco de dados, jogos, aplicativos para celulares, entre outros (MENEZES, 2010).

Python é software livre, ou seja, é uma linguagem disponibilizada gratuitamente e mantida pelo trabalho da *Python Software Foundation* e pela colaboração de inúmeros programadores que realizam constantes modificações (BORGES, 2014).

Ao considerar o processo de análise de dados, o *Python* pode atrair comparações com outras linguagens de computação, como o R, entretanto esta tem avançado constantemente e se transformou em uma opção popular para tais análises (MCKINNEY, 2019).

O crescimento da sua popularidade pode estar relacionado ao fato de *Python* ser uma linguagem completa contando com diversas bibliotecas para auxiliarem nas mais variadas tarefas, aumentando assim a produtividade de um programador que a utilize (MENEZES, 2010).

3.2 Ambiente de Desenvolvimento Integrado

As linguagens de programação necessitam de um Ambiente de Desenvolvimento Integrado, ou *Integrated Development Environment* (IDE), para a implementação dos códigos de desenvolvimento de um programa (FRANCO; ADANIYA, 2018).

A IDE utilizada nesta pesquisa é o *Jupyter Notebook*, uma ferramenta que possui códigos iterativos baseados na *Web*, os denominados *notebooks*, que oferecem suporte para dezenas de linguagens de programação, incluindo o Python. Os notebooks são ferramentas extremamente úteis para exploração, visualização e análise de dados (MCKINNEY, 2019). O Jupyter é disponibilizado em conjunto com outros aplicativos no Repositório Anaconda.

3.3 O Twitter

As redes sociais se tornaram extremamente populares nos últimos anos, sendo parte do cotidiano das pessoas. Alguns estudiosos atribuem grande parte dessa adesão devido a sua

capacidade de compartilhar informações entre usuários de qualquer parte do mundo instantaneamente. Se a quantidade de usuários avançou em larga escala, a quantidade de dados produzidos é ainda maior, criando assim um desafio para organizar e extrair informações úteis a partir de tais dados (SILVA, 2016).

O *Twitter* é uma plataforma popular nos dias atuais e se destaca pelas interações praticamente imediatas. Neste, cada usuário possui uma espécie de microblog onde pode escrever mensagens de até 140 caracteres, os ‘*tweets*’ (FRANCO; ADANIYA, 2018). Atualmente a plataforma do *Twitter* já disponibiliza opções de contas onde o usuário possui a capacidade de escrever mensagens com uma quantidade maior de caracteres.

O *Twitter* tem se tornado uma plataforma interessante para o uso em pesquisas devido a sua aceitação pela população, como também pela possibilidade que as pessoas tem de se expressar não apenas por meio de palavras, mas também através do uso de imagens, vídeos e emoticons, tornando-a cada vez mais popular (RODAS et al., 2022).

Ao se considerar os *tweets* isoladamente, estes podem parecer aleatórios e pouco representativos quanto a opinião geral de uma população, entretanto quando são observados dentro de um contexto maior, eles permitem a análise e identificação dos assuntos que são mais discutidos na Web em determinado momento, bem como possibilita a análise das opiniões sobre um determinado tema específico (FRANCO; ADANIYA, 2018).

A seguir serão apresentados alguns termos específicos da plataforma, de acordo com trabalho de Silva (2016):

- *Tweet* – mensagem ou post realizado na plataforma.
- *Username* ou *User* – registro que o usuário necessita fazer para utilizar uma conta no *Twitter*. A interação entre usuários pode ser feita através da menção do *user* de um usuário no post de outro usuário.
- *Hashtag* – ao acrescentar o símbolo # a uma palavra ou expressão, o usuário pode atribuir tags a seus tweets, por exemplo #VacinaSim.
- *Follower* – um usuário no *Twitter* se conecta a outro usuário através dessa opção, assim quando um postar um novo tweet, seus *followers* serão notificados desta ação. Contudo, diferentemente de outras redes sociais, a relação de seguir alguém é unilateral, ou seja, o usuário pode seguir uma pessoa e esta não o seguir de volta.

- *Retweet* – a função de *retweet* (RT) é usada para propagar a mensagem de outro usuário para seus seguidores. Nos casos de RT o *tweet* não pode ser modificado.

A Figura 09 mostra o exemplo de um tweet publicado pelo perfil do Ministério da Saúde cujo *user* é @minsaude. Na imagem é possível verificar alguns dos termos supracitados como o uso de *hashtags*, #PARTIUVACINAR, e os *retweets* que são visualizados no segundo ícone da parte inferior da mesma, ou seja, esse *tweet* foi *retweetado* 18 vezes

Figura 09. Exemplo de tweet



Fonte: Twitter (2022)

3.4 Análise de Sentimentos

A análise de sentimentos, também denominada pelos estudiosos como Mineração de Opinião, é um método utilizado para identificar emoções e opiniões em textos, a partir do uso das técnicas de Mineração de Textos, Informática, Linguística, entre outros (BRITO, 2017). A análise de sentimentos é considerada pelos estudiosos como um subcampo do Processamento de Linguagem Natural, uma área definida como conjunto de técnicas computacionais criadas

para analisar e representar textos naturais, de forma a obter um processamento de linguagem o mais próximo possível do humano (OLENSCKI et al., 2020).

Sentimento é definido por Liu (2012) como uma opinião ou uma avaliação sobre algum aspecto ou objeto. Já a análise do sentimento, envolve o PLN, para que extrair as emoções expressas principalmente em textos online, onde esta área tem sido bastante utilizada para verificar a popularidade de pessoas, objetos, lugares e situações (BRITO, 2017).

A classificação dos sentimentos identifica a polaridade do texto analisado, mesmo que este seja um documento, uma frase ou apenas uma palavra, classificando-o em positivo, negativo ou neutro. Para realização de uma classificação utilizando sentimentos mais amplos, é necessário realizar alterações nos algoritmos para que estes sejam capazes de determinar tais emoções (LIU, 2012).

Segundo Silva (2016) a análise de sentimentos aplicada a redes sociais pode ser caracterizada em quatro tipos, sendo eles:

- Supervisionada – Esta técnica é realizada por meio do uso de algum algoritmo de aprendizagem de máquina e exige a existência de uma base de dados para treinamento que seja previamente rotulada por um especialista.
- Guiada pelo uso de léxico – Neste modelo é utilizada uma lista de termos classificados, por alguém especializado, em positivos e negativos que direcionará o processo de identificação da polaridade da base de dados.
- Supervisionada Híbrida - Nesta abordagem são aplicados simultaneamente um algoritmo de aprendizagem de máquina e uma lista de léxicos.
- Baseada em Grafos – Esta técnica utiliza as características específicas das redes sociais, como os relacionamentos e interações entre os usuários para realizar a análise.

A análise de sentimentos é um método que utiliza a mineração de textos em sua execução, além de apresentar também suas técnicas específicas de acordo com a aplicação desejada (BRITO, 2017).

3.4.1 Granularidade da análise

De acordo com o trabalho apresentado por Franco e Adaniya (2018) a análise de sentimentos ocorre em três níveis distintos dependendo do grau de granularidade da análise, sendo eles:

- A nível de documento – Neste nível o documento como um todo é classificado em uma única entidade. Sua maior utilidade está na classificação de textos de grande especificidade, como na avaliação e feedbacks de produtos e serviços.
- A nível de sentença – Neste nível o documento pode conter diversas entidades com classificações independentes, sendo as sentenças do documento o menor nível a se considerar.
- A nível de aspecto – Nesta tarefa a análise é muito específica de maneira que se identifique a opinião apresentada sobre uma determinada entidade e não no contexto da sentença ou do documento. Deste modo, uma sentença pode ser classificada na sua totalidade como negativa, mas a classificação de uma entidade específica dentro da sentença pode se apresentar como positiva.

3.4.2 Léxico de opiniões

Na literatura são apresentadas palavras, expressões ou até mesmo frases que são importantes indicadores de sentimentos, sendo estas classificadas de acordo com a sua polaridade em positivas ou negativas. A lista composta por palavras individuais de sentimentos e expressões idiomáticas é denominada léxicos de sentimentos ou léxicos de opinião. Os léxicos de opinião podem ser criados manualmente ou automaticamente por meio de palavras pré-definidas que são expandidas de acordo com a análise semântica (SILVA, 2016).

As palavras e frase de sentimentos são muito importantes para a análise de sentimentos, entretanto segundo Silva (2016), estas podem apresentar alguns problemas como:

- De acordo com o contexto da frase as palavras podem assumir polaridades diferentes.
- Uma frase que contenha um léxico de opinião pode não expressar nenhum tipo de sentimento, sendo este fato muito comum em sentenças interrogativas.

- As sentenças que contêm sarcasmo são extremamente difíceis de lidar, pois o programa não consegue reconhecer este tipo de situação.
- Frases que não contêm palavras de sentimentos podem, ainda assim, expressar uma opinião.

Diversos estudiosos produziram em seus trabalhos léxicos de opiniões e dicionários que auxiliam o desenvolvimento de pesquisa posteriores. Deste modo, destacam-se:

- Liu (2002) criou em seu trabalho um léxico conhecido como Bing Liu Lexicon e realizou sua aplicação em diversas pesquisas. Este léxico é para a língua inglesa e contém 2006 palavras classificadas como positivas e 4683 como negativas.
- Mohammed et al. (2013) criou o Sentiment140 Lexicon em sua pesquisa. Tal léxico utiliza emoticons classificados como positivos e negativos e por meio desses emoticons, as frases são rotuladas.
- Silva et al. (2012) desenvolveu em seu trabalho o SentiLex-PT, um léxico de sentimentos para a língua portuguesa que possui 7014 lemas e 82347 formas flexionadas.

Neste trabalho será empregado o SentiLex-PT para realizar a análise de sentimentos na base de dados construída.

3.5 Métricas de validação

Para avaliar a performance de um algoritmo de mineração de textos, a literatura possui algumas métricas a serem observadas. Segundo Morais e Ambrósio (2007) e Junior (2007), as pesquisas dentro desta área do conhecimento, devem valer-se de tais métricas:

- Precisão – Mede a habilidade de um sistema em recuperar apenas os documentos relevantes a consulta considerando determinada classe.

$$\text{precisão } (A) = \frac{\text{total de exemplos corretamente classificados da classe } A}{\text{total de exemplos corretamente classificados}}$$

- Recall (Eficiência) – Mede a habilidade de um sistema em recuperar os documentos mais relevantes para o usuário.

$$\text{recall } (A) = \frac{\text{total de exemplos corretamente classificados da classe } A}{\text{total de exemplos da classe } A}$$

- Acurácia – Mede a proporção total de classificações corretas.

$$\text{acurácia} = \frac{\text{total de amostras classificadas corretamente (todas as classes)}}{\text{total de exemplos do conjunto teste}}$$

A ‘Classe A’ supracitada nas equações de recall e precisão é uma classe hipotética a ser observada no experimento, podendo esta ser qualquer uma das classificações realizadas. No caso deste experimento em específico, seria a representação de uma classe utilizada para classificação, sendo estas: positivo, negativo e neutro.

4 RESULTADOS E DISCUSÕES

Neste capítulo serão relatados os resultados obtidos com a pesquisa, iniciando com os dois algoritmos criados, um para a coleta de dados e outro para análise de sentimentos, e nas informações expressas nas bases de dados.

4.1 Algoritmo de Coleta de Dados

O primeiro algoritmo produzido é responsável por realizar a coleta de dados, através da aplicação da biblioteca para linguagem *Python*, *Tweepy*. É importante salientar que neste momento do trabalho, na execução prática, alguns dos termos supracitados serão empregados.

Para iniciar o processo é necessário possuir uma conta no *Twitter* para desenvolvedores. Após a aprovação da conta, o usuário pode criar um aplicativo para ser vinculado a IDE escolhida pelo mesmo. O *Twitter* criará automaticamente credenciais para liberar o acesso do usuário a uma quantidade pré-determinada de tweets, de acordo com o modelo de conta solicitado. As credenciais fornecidas pela plataforma são individuais e não podem ser compartilhadas com outras pessoas.

Na sequência do processo é realizada a autenticação por meio da inserção das credenciais no algoritmo construído, estabelecendo uma conexão entre a IDE e o *Twitter*. O *Tweepy* já possui alguns módulos que permitem a coleta de dados, sendo assim necessária a definição dos parâmetros para efetuar a busca.

4.1.1 Dados de Teste

A construção da base de dados para testes foi realizada no período de primeiro de abril de 2021 a 07 de abril de 2021. A coleta foi realizada através da busca por *endpoints*, por meio do *Tweepy*. Nesse momento é importante destacar que a API do *Twitter* passou por atualizações recentes e atualmente o retorno para cada pesquisa realizada é de no máximo 100 tweets, selecionados de acordo com perfis de interação próprios da plataforma.

Os *endpoints* utilizados para esta busca foram: *pandemia* e *Covid19*. Não foram consideradas as *hashtags* nesse momento. Como resultado desta busca inicial, obteve-se 700 *tweets*, nos quais foram realizados os tratamentos iniciais de limpeza, onde foram excluídas as publicações que não possuíam classificação de sentimentos, como os informativos publicados

por diversos perfis de canais de comunicação. Esse procedimento será explicado detalhadamente, no tópico 4.2.

Após a exclusão restaram 248 tweets que foram manualmente classificados, dos quais 150 são negativos, 52 positivos e 46 nulos. A presença de um número de publicações negativas muito superior as demais pode ser explicada pelo fato dos sentimentos que a pandemia produzia na população eram em sua essência de medo, tristeza, angústia, entre outros já apresentados neste trabalho. A Tabela 01 traz exemplos dos *tweets* classificados. Os erros de ortografia e gramática que as publicações possuíam foram mantidos.

Tabela 01. Exemplo de Tweets da Base de Testes

Tweets	Polaridade
só quero acordar desse pesadelo chamado pandemia	Negativo
milhões de pessoas em situação de imigração se veem desumanizadas e desprotegidas, cenário ainda mais grave no meio de uma pandemia que hierarquiza vidas	Negativo
Nessa pandemia to aprendendo valorizar cada vez mais o abraço, as risadas, as alegrias as tristezas a oportunidade de congregar e de conviver. Quando tudo isso passar não quero perder meu tempo com coisas fúteis, quero investir meu tempo em pessoas!	Positivo
E em meio a essa pandemia, só tenho a agradecer e ficar feliz por poder estar com minha família e ter uma ceia feliz	Positivo
Fome atingiu 19 milhões de brasileiros durante a pandemia, diz pesquisa.	Neutro
Gilmar Mendes contraria Kassio Nunes e respalda fechamento de igrejas e templos durante pandemia	Neutro

Fonte: Autoria Própria

4.1.2 Dados da pesquisa

A construção da base de dados para a pesquisa foi realizada no período de 03 de novembro de 2020 a 20 de fevereiro de 2021. Esta segunda coleta também foi realizada através da busca por *endpoints*, por meio da biblioteca *Tweepy*. A base de dados para pesquisa foi dividida em 3 grandes grupos, onde cada grupo compreende o período de 5 semanas epidemiológicas. O primeiro período inicia 03 de novembro de 2020 e termina em 07 de dezembro de 2020, na data em que ocorreu a primeira vacinação contra Covid 19 no mundo.

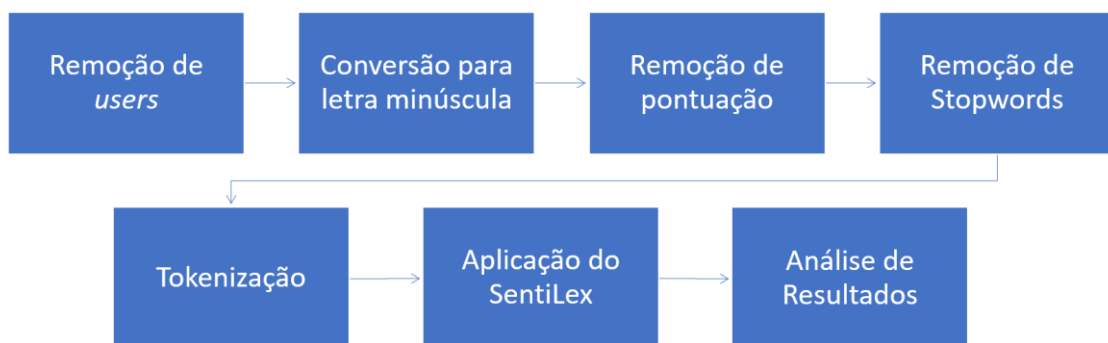
O segundo período compreende de 08 de dezembro de 2020 a 16 de janeiro de 2021, sendo este o período em que já existiam vacinas em aplicação pelo mundo, entretanto não havia iniciado o protocolo vacinal no Brasil. O terceiro período inicia em 17 de janeiro de 2021 e encerra em 20 de fevereiro de 2021, marcando os primeiros dias de vacinação contra a Covid 19 no Brasil.

Os *endpoints* utilizados para esta busca foram: vacina, coronavírus e Covid19. Nesta busca também não foram consideradas as *hashtags*. Como resultado desta busca inicial, obteve-se 11000 *tweets*, nos quais foram realizados os tratamentos abaixo especificados. Após o tratamento inicial, a base de dados para a pesquisa era constituída de 5041 *tweets*.

4.2 Algoritmo de Análise de Sentimentos

A Figura 10 apresenta a estrutura de execução do algoritmo construído nesta pesquisa. O algoritmo foi aplicado no conjunto de dados obtidos na etapa de coleta de dados, classificado como Banco de Dados não Relacional, pois não existe uma relação entre pré-estabelecida entre os dados (tópico 2.2.2 deste trabalho).

Figura 10. Estrutura do Algoritmo



Fonte: Autoria Própria

O algoritmo é iniciado com a limpeza dos *tweets* mantendo apenas aqueles que possam verdadeiramente fornecer uma informação quanto a opinião dos usuários, para isso

foram excluídas as publicações que possuíam conteúdos informativos, como as realizadas pelas páginas de meios de comunicação. A Figura 11 apresenta um exemplo de *tweet* que foi excluído por não apresentar informações relevantes quanto a opiniões.

Figura 11. Exemplo de *Tweet* a ser excluído



Fonte: *Twitter* (2022)

Para possibilitar esta etapa de exclusão foram identificados os *users* das páginas de comunicação e os *tweets* retornados por tais usuários eram excluídos, até que restassem apenas as publicações que continham opiniões. A Figura 12 apresenta alguns dos *users* excluídos.

Para realizar o Processamento de Linguagem Natural foi utilizada a biblioteca NLTK. Esta biblioteca está disponível para a linguagem de programação Python e contém por volta de 35 módulos. Cada modulo possui seus diversos submódulos que realizam múltiplas tarefas de PLN, como tokenização, remoção de *stopwords*, e possui inclusive um módulo para análise de sentimentos, entre outros módulos (ARAÚJO,2017). Neste trabalho não foi utilizado o módulo de análise de sentimentos.

Figura 12. Alguns *Users* excluídos

@CNNBrasil @UOL @exame @RadioBandNewsBH @UOLNoticias
 @o_antagonista @Estadao @folha @otempo @opovo
 @g1saopaulo @JornalOGlobo @ESPNBrasil

Fonte: Autoria Própria

Os módulos aplicados neste algoritmo, na etapa de Pré-Processamento são: a remoção de caracteres especiais e pontuação, a remoção de *stopwords* (conforme consta no tópico 2.3.3.2.2 deste trabalho), a tokenização (conforme tópico 2.3.3.2.1) e a conversão para letras minúsculas.

Após a etapa de pré-processamento é realizada a aplicação do dicionário de léxicos SentiLex-PT aos *tweets*. A aplicação deste consiste na busca de cada palavra que compõe o tweet dentro da estrutura do dicionário. Neste modelo de avaliação cada tweet é entendido como um vetor, onde as palavras são suas estruturas. Cada palavra recebe um valor e o resultado obtido é o somatório do valor das palavras. A classificação e o valor atribuído as palavras estão expressos na Tabela 02. Entretanto, é necessário destacar que o dicionário já apresenta palavras pré-definidas com seus devidos valores.

Tabela 02. Valor atribuído as palavras de acordo com o dicionário

Tipo de palavra	Valor
Palavras que expressam sentimentos negativos	-1
Palavras que expressam sentimentos positivos	+1
Palavras que não expressam sentimentos	0

Fonte: Adaptado de Silva (2016).

Na sequência das etapas da MT, é realizada a avaliação dos resultados obtidos. Tal avaliação será realizada através da utilização da métrica acurácia apresentada no tópico 3.5 deste trabalho. A acurácia encontrada no algoritmo aplicado é 68,5 %. O valor pode ser

considerado pequeno quando comparado aos algoritmos que utilizam técnicas de aprendizagem de máquina, entretanto, é considerado adequado ao se tratar da aplicação de dicionários de léxicos.

A Figura 13 apresenta os resultados obtidos para a análise de sentimento aplicada no banco de dados da pesquisa, conforme a divisão de períodos supracitada. Para justificar certos comportamentos encontrados, como a presença majoritária de tweets negativos em ambos os períodos, é importante salientar que essa negatividade pode não estar diretamente associada a vacina, ou seja, a pessoa pode não estar defendendo a não aplicação da vacina e sim estar discorrendo sobre algum acontecimento acerca da vacinação que causa sentimentos negativos na pessoa que realizou a publicação. A Tabela 03 apresenta a quantidade de tweets observados para cada período

Tabela 03. Valores obtidos pelo algoritmo

Período	Positivo	Negativo	Neutro	Total
Primeiro	435	747	343	1525
Segundo	562	851	336	1749
Terceiro	643	865	259	1767

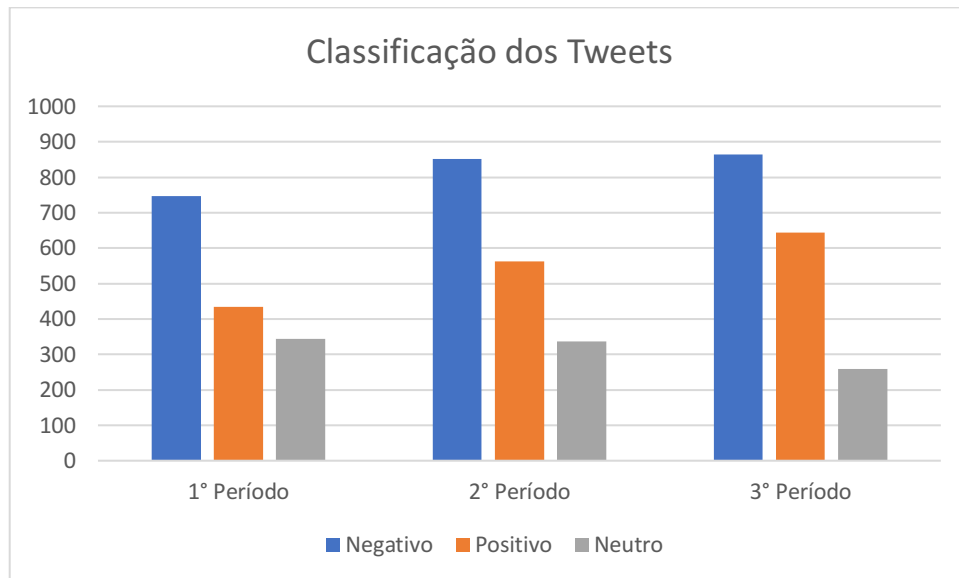
Fonte: Autoria Própria.

A seguir serão apresentados alguns temas e notícias presentes no período de coleta de dados e que possivelmente seriam responsáveis por algumas dessas variações:

- A ocorrência de eleições municipais no Brasil.
- A interrupção que a Anvisa realizou nos estudos das vacinas realizados no território brasileiro.
- Caos instaurado em Manaus devido a falta de oxigênio e de leitos para atendimento dos pacientes.
- Declarações diversas realizadas pelo presidente Jair Bolsonaro, que causaram grande alvoroço nas redes sociais, como a aplicação da vacina que supostamente transformaria as pessoas em ‘jacaré’.
- Com o início da vacinação, surgem novos problemas como as pessoas que burlavam as filas de pessoas prioritárias para receberem as vacinas antes da sua convocação oficial.

- Profissionais da saúde que foram gravados no momento da vacinação e não injetaram nenhum medicamento nos pacientes.

Figura 13. Resultado da pesquisa



Fonte: Autoria Própria

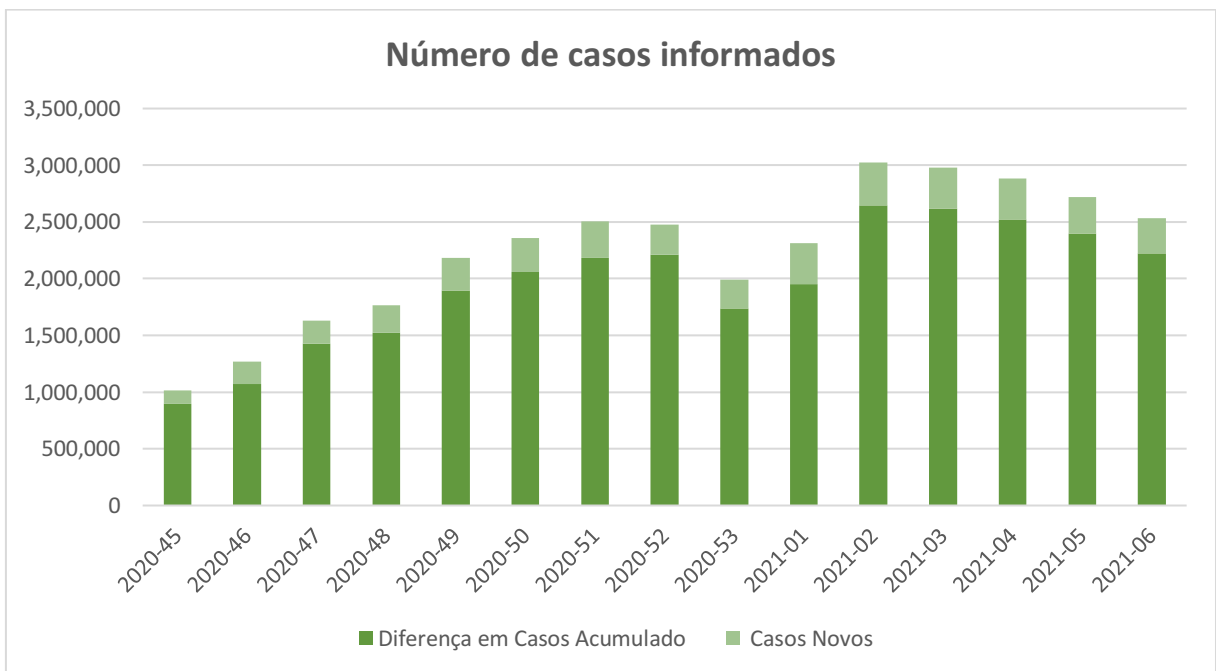
Para proporcionar uma visão mais clara do cenário em que se encontrava o Brasil no momento em que tais dados foram coletados, a Figura 14 apresenta a quantidade de óbitos contabilizados nas semanas epidemiológicas em que se realizou a pesquisa e a Figura 15 expõe a quantidade crescente de casos de pessoas infectadas. A vacina teve seu uso emergencial liberado pela ANVISA no momento que o Brasil iniciava a curva ascendente de casos da segunda onda de contaminação. É importante ainda destacar que o ápice de óbitos no país ocorreu após a liberação da vacina, entretanto não havia imunizante o suficiente para que toda a população fosse vacinada em um curto período de tempo.

Figura 14. Quantidade de óbitos por semana epidemiológica



Fonte: Adaptado CONASS (2022)

Figura 15. Quantidade de casos novos por semana epidemiológica




Fonte: Adaptado CONASS (2022)

Algumas discussões são necessárias para esclarecer determinados pontos, sendo um deles a escolha do Twitter como plataforma para a coleta de dados. A quantidade de informações compartilhadas pelos usuários sobre a pandemia, chegou em certo momento a ser considerado por alguns estudiosos como uma infodemia, pois o volume de notícias e informações imprecisas compartilhadas estavam se espalhando pelo mundo em maior velocidade do que o próprio vírus (RODAS et al., 2022).

A rede social Twitter é conhecida por desde seu lançamento prezar pela disseminação de fatos e informações verídicas, realizando a verificação e rápida remoção de muitas Fake News. É importante salientar que tal plataforma não é totalmente eficiente, mas procura realizar uma análise mais crítica do que as demais plataformas.

Visando conter o compartilhamento de informações enganosas por falta de conhecimento do usuário do assunto e apresentar fontes cada vez mais confiáveis, o Twitter divulgou que a partir de 16 de dezembro de 2020 adotaria uma nova abordagem em relação a informações sobre a vacina, incluindo avisos e alertas nos tweets que possuíssem informações duvidosas. A Figura 16 apresenta um indicador dos avisos (RODAS et al., 2022). A Figura 17 foi retirada do próprio Twitter e mostra outra medida adotada pela plataforma para garantir que os usuários tenham acesso a informações verídicas.

Figura 16. Medidas adotadas pelo Twitter



Informações enganosas	Aviso	Remoção
Afirmações questionáveis	Aviso	Alerta
Afirmações não confirmadas	Sem medida	Sem medida*
	Moderada	Grave
Propensão a dano		

Fonte: Rodas et al. (2022).

Figura 17. Exemplo do Twitter em mostrar informações reais

Conheça os fatos

Para garantir que você tenha as melhores informações sobre o novo coronavírus, recursos da Organização Pan-Americana da Saúde (OPAS) estão disponíveis.

Saiba mais

Vacinas contra a COVID-19

Fonte: Twitter (2022).

Outra questão que deve ser abordada é a dificuldade na execução de análises em postagens realizadas em qualquer rede social devido não apenas a complexidade gramatical da língua portuguesa, mas também aos erros de ortografia presentes em diversas publicações. Tais erros fazem com que as palavras não sejam corretamente percebidas pelo algoritmo, apesar de que para o entendimento humano estes não causem danos a compreensão do que se deseja expor, para o computador estes erros comprometem o desempenho do algoritmo como um todo. A Figura 18 apresenta um exemplo dos erros de ortografia presentes nas publicações realizadas.

Figura 18. Exemplo de tweet com erro de ortografia



Fonte: Twitter (2022)

5 CONSIDERAÇÕES FINAIS

A pandemia do Covid 19 já é considerada por vários estudiosos como o maior problema sanitário do XXI. A sua rápida expansão no cenário global gerou um estado de caos na população mundial, que de repente se encontra isolada, privada de sua liberdade e do contato social. Tais situações contribuíram para criar um misto de sentimentos nas pessoas, gerado principalmente pelo temor do desconhecido, visto que não se conhecia muito sobre o vírus, o seu comportamento e tratamentos que fossem eficazes.

Neste contexto é iniciada uma corrida mundial pelo desenvolvimento de uma vacina que pudesse evitar o contágio do vírus ou pelo menos, diminuir a proporção de casos graves que acarretavam em óbitos. Quando considerado apenas o cenário pandêmico brasileiro, a situação era agravada pela falta de investimentos nas pesquisas sobre as vacinas, o número crescente de *Fake News* que atrapalhavam não só o processo de desenvolvimento como posteriormente veio a dificultar a aceitação da população em se vacinar, crises políticas, entre outras dificuldades já apresentadas.

Assim sendo, tal pesquisa foi desenvolvida com o objetivo de realizar um estudo sobre a percepção dos brasileiros acerca da vacina contra a Covid-19, baseado em uma amostra aleatória das postagens realizadas neste período de tempo, não compreendendo a totalidade das postagens realizadas pela população no Twitter, no período marcado pela aprovação das primeiras vacinas no Brasil e no mundo. O algoritmo desenvolvido apresentou acurácia de 68,5%, sendo considerado um valor adequado.

Os resultados obtidos nesta pesquisa foram publicados em um artigo na *Revista Research, Society and Development*, conforme pode ser visualizado no Anexo A deste trabalho. Para dar continuidade a pesquisa, seria viável o desenvolvimento de um algoritmo baseado na aprendizagem de máquina para que sejam realizadas comparações entre seus respectivos resultados. Tal desenvolvimento necessita de uma base de dados para treinamento do algoritmo que possua uma quantidade elevada de publicações já classificadas de acordo com a polaridade possibilitando uma análise adequada do modelo.

REFERÊNCIAS BIBLIOGRÁFICAS

ARANHA, C.; PASSOS, E. **A tecnologia de mineração de textos**. Revista Eletrônica de Sistemas de Informação, v. 5, n. 2, 2006.

ARAÚJO, L. G. de A. **SENTIMENTALL VERSÃO 2: Desenvolvimento de Análise de Sentimentos em Python**. Trabalho de Conclusão de Curso, Centro Universitário Luterano de Palmas, Palmas, Tocantins, 2017.

BORGES, L. E. **Python para desenvolvedores: aborda Python 3.3**. Novatec Editora, 2014.

BRITO, E. M. N. **Mineração de Textos: detecção automática de sentimentos em comentários nas mídias sociais**. Dissertação: Programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade Fundação Mineira de Educação e Cultura — FUMEC. Belo Horizonte, 2017.

CAMILO, C. O.; SILVA, J. C. da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFG), p. 1-29, 2009.

CARDOSO, R. F. et al. **COVID-19: An epidemiological challenge**. Research, Society and Development, [S. l.], v. 10, n. 7, p. e32110716313, 2021. DOI: 10.33448/rsd-v10i7.16313.

CASTRO, R. **Vacinas contra a Covid-19: o fim da pandemia?** Physis: Revista de Saúde Coletiva, v. 31, p. e310100, 2021.

CONASS - Conselho Nacional de Secretários de Saúde (2022). **Painel Covid-19**. Disponível em: <<https://www.conass.org.br/painelconasscovid19/>>. Acesso em 17 de julho de 2022.

CORDEIRO, A. D. et al. **Gerador inteligente de sistemas com auto-aprendizagem para gestão de informações e conhecimento**. Tese – PhD, Universidade Federal de Santa Catarina, Departamento de Engenharia da Produção, 2005.

DOMINGUES, C. M. A. S. **Desafios para a realização da campanha de vacinação contra a COVID-19 no Brasil**. Cadernos de Saúde Pública, v. 37, 2021.

EBECKEN, N; LOPES, M; COSTA, M. **Mineração de Textos**, chapter 13, p. 337–370. Barueri, São Paulo: Manole, 2003.

FIGUEIREDO, E. B.; CATINI, R. C.; MENDES, L. M. **Mineração de Textos: Análise de Sentimento em Redes Sociais-Revisão Sistemática**. Anais do WCF, v. 5, p. 24-29, 2018.

FRANCO, R. B.; ADANIYA, M. H. A. da C. **Sistemas de análise de sentimentos usando dados do Twitter**. Revista Terra & Cultura: Cadernos de Ensino e Pesquisa, v. 34, n. esp., p. 111-118, 2018.

GARCIA, V. S.; SOTTO, E. C. S. **Comparativo Entre os Modelos de Banco de Dados Relacional e Não-Relacional**. Revista Interface Tecnológica, v. 16, n. 2, p. 12-24, 2019.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 7 ed. Rio de Janeiro: Atlas, 2019.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6 ed. Rio de Janeiro: Atlas, 2017.

JUNIOR, J. R. C. **Desenvolvimento de uma Metodologia para Mineração de Textos**. Pontifícia Universidade Católica de Rio de Janeiro: Rio de Janeiro, Brasil, 2007.

JUNQUEIRA, K. T. C.; FERNANDES, A. M. da R. **Análise de sentimento em redes sociais no idioma português com base em mensagens do twitter**. Anais do Computer on the Beach, p. 681-690, 2018.

LADEIA, D. N. et al. **Análise da saúde mental na população geral durante a pandemia de Covid-19**. Revista Eletrônica Acervo Saúde, n. 46, p. e3925-e3925, 2020.

LIU, B. **Sentiment analysis and opinion mining**. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MCKINNEY, W. **Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython**. Novatec Editora, 2019.

MENEZES, N. N. C. **Introdução a programação com Python**. São Paulo: Novatec, 2010.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de textos**. Relatório Técnico–Instituto de Informática (UFG), 2007.

Mohammad, S.M.; Kirichenko, S.; Zhu, X. **Building the state-of-the-art in sentiment analysis of tweets. Proceedings of the seventh international workshop on Semantic Evaluation Exercises.** Atlanta, Georgia, Estados Unidos, 2013.

OLENSCKI, J. et al. **Aplicação de análise de sentimentos no Twitter para avaliação da percepção pública quanto a cloroquina.** Em: Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde. SBC, 2020. p. 500-505.

OMS - Organização Mundial da Saúde (2020). **Coronavirus disease 2019 (COVID-19): Situation Report –51.** Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid19.pdf?sfvrsn=1ba62e57_10> Acesso em 15 de julho de 2022.

PEZZINI, A. **Mineração de textos: Conceito, processo e aplicações.** REAVI-Revista Eletrônica do Alto Vale do Itajaí, v. 5, n. 8, p. 58-61, 2016.

PIMENTEL, A. do S. G.; SILVA, M. de N. R. M. de O. **Psychic Health in Times of Corona Virus.** Research, Society and Development, [S. l.], v. 9, n. 7, p. e11973602, 2020. DOI: 10.33448/rsd-v9i7.3602.

RAMOS, B.; FREITAS, C. **“Sentimento de quê?”: uma lista de sentimentos para a análise de sentimentos.** STIL, p. 15-18, 2019.

RODAS, C. M. et al. **Análise de sentimentos sobre as vacinas contra Covid-19: um estudo com algoritmo de machine learning em postagens no Twitter.** Revista Saúde Digital. Fortaleza, 2022.

SILVA, E. P. da; MALHEIROS, Y. **Um conjunto de dados extraído do Twitter para análise de sentimentos na língua portuguesa.** Trabalho de Conclusão de Curso, Universidade Federal da Paraíba, 2019.

SILVA, N. F. F. da. **Análise de sentimentos em textos curtos provenientes de redes sociais.** Tese de Doutorado, Universidade de São Paulo, 2016.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. **Building a Sentiment Lexicon for Social Judgement Mining. In Lecture Notes in Computer Science (LNCS) / Lecture Notes in**

Artificial Intelligence (LNAI), International Conference on Computational Processing of Portuguese (PROPOR), 17-20 April, 2012, Coimbra.

SOUSA, A. R. de et al. **Sentimento e emoções de homens no enquadramento da doença Covid-19**. *Ciência & Saúde Coletiva*, v. 25, p. 3481-3491, 2020.

SOUZA, F. J. de; SANTANA, P. H. A. de. **ESTUDO DE CASO: ANÁLISE ENTRE BANCO DE DADOS RELACIONAL E NÃO RELACIONAL**. Trabalho de Conclusão de Curso, Universidade Evangélica de Goiás, 2017.

TWITTER. Página Inicial. Disponível em < <https://twitter.com>>. Acesso em 10 de março de 2022.

ANEXO A – ARTIGO PUBLICADO

Research, Society and Development, v. 11, n. 13, e364111335490, 2022
 (CC BY 4.0) | ISSN 2525-3409 | DOI: <http://dx.doi.org/10.33448/rsd-v11i13.35490>

Mineração de textos e análise de sentimentos aplicados a postagens do Twitter acerca das vacinas contra a Covid-19

Text mining and sentiment analysis applied to Twitter posts about Covid-19 vaccines

Minería de texto y análisis de sentimiento aplicado a publicaciones de Twitter sobre vacunas Covid-19

Recebido: 19/09/2022 | Revisado: 29/09/2022 | Aceitado: 03/10/2022 | Publicado: 09/10/2022

Franciele Leal Farias

ORCID: <https://orcid.org/0000-0002-4380-4987>
 Universidade Federal dos Vales do Jequitinhonha e Mucuri, Brasil
 E-mail: franleal09@hotmail.com

Lorena Sophia Campos de Oliveira

ORCID: <https://orcid.org/0000-0003-0704-1828>
 Universidade Federal dos Vales do Jequitinhonha e Mucuri, Brasil
 E-mail: lorena.sco@gmail.com

Resumo

A pandemia da Covid-19 já é considerada por muitos estudiosos o maior problema sanitário do século XXI e ceifando a vida de milhares de pessoas. A rapidez com que a doença se espalhou e modificou a vida da população mundial gerou uma grande quantidade de emoções e sentimentos nas pessoas. Desde a descoberta do novo coronavírus, iniciou-se uma corrida pelo desenvolvimento de uma vacina que fosse eficaz para o combate da doença, crescendo o anseio da população pela sua chegada. O trabalho realiza a análise dos sentimentos que a população brasileira desenvolveu em relação às vacinas criadas para o combate da Covid-19, por meio da utilização das técnicas de análise de sentimento e mineração de dados. A construção do banco de dados ocorreu através da captação de postagens públicas disponibilizadas pela API do Twitter. O algoritmo desenvolvido durante a pesquisa é baseado na linguagem de programação Python e implementado na plataforma Jupyter Notebook. O processo de análise de sentimentos foi realizado através da análise semântica, com uso do dicionário de léxicos para a língua portuguesa SentiLex-PT.

Palavras-chave: Pandemia; Vacina; Covid-19; Mineração de textos; Análise de sentimentos.

Abstract

The Covid-19 pandemic is already considered by many scholars to be the biggest health problem of the 21st century and claiming the lives of thousands of people. The speed with the disease spread and changed the lives of the world's population generated a lot of emotions and feelings in people. Since the discovery of the new coronavirus, a race began to develop a vaccine that would be effective to combat the disease, growing the population's desire for its arrival. The work analyzes the feelings that the Brazilian population has developed in relation to vaccines created to combat Covid-19, through the use of sentiment analysis and data mining techniques. The construction of the database took place through the capture of public posts made available by the Twitter API. The algorithm developed during the research is based on the Python programming language and implemented on the Jupyter Notebook platform. The sentiment analysis process was carried out through semantic analysis, using the lexicon dictionary for the Portuguese language SentiLex-PT.

Keywords: Pandemic; Vaccine; Covid-19; Text mining; Sentiment analysis.

Resumen

Muchos académicos ya consideran que la pandemia de Covid-19 es el mayor problema de salud del siglo XXI y se cobra la vida de miles de personas. La velocidad con la que la enfermedad se propagó y cambió la vida de la población mundial generó muchas emociones y sentimientos en las personas. Desde el descubrimiento del nuevo coronavirus se inició una carrera por desarrollar una vacuna que fuera eficaz para combatir la enfermedad, y ha crecido el deseo de la población por su llegada. El trabajo analiza los sentimientos que la población brasileña ha desarrollado en relación a las vacunas creadas para combatir el Covid-19, mediante el uso de técnicas de análisis de sentimiento y minería de datos. La construcción de la base de datos se realizó a través de la captura de publicaciones públicas disponibles a través de la API de Twitter. El algoritmo desarrollado durante la investigación está basado en el lenguaje de programación Python e implementado en la plataforma Jupyter Notebook. El proceso de análisis de sentimiento se llevó a cabo a través del análisis semántico, utilizando el diccionario de léxico para la lengua portuguesa SentiLex-PT.

Palabras clave: Pandemia; Vacuna; Covid-19; Extracción de textos; Análisis de los sentimientos.

1. Introdução

Um vírus novo foi responsável por desencadear uma doença respiratória agressiva, com grande taxa de transmissibilidade que rapidamente se espalhou por todo o planeta, não permitindo que medidas de contenção efetivas fossem descobertas antes que este se alastrasse por todos os países do mundo, fazendo milhares de vítimas fatais. O processo de transmissão do vírus ocorre de pessoa para pessoa, por meio do contato com gotículas de saliva ou superfícies contaminadas. O período de incubação do vírus pode variar de 2 a 14 dias (Cardoso, et al., 2021). De tal modo, a pandemia acarretou problemas nas esferas social, política, econômica e educacional, gerando um aglomerado de sentimentos e emoções na população. Dentro desse contexto mundial, o Brasil se encontra atualmente incluído não apenas em um ambiente hostil gerado pela doença como também tem sofrido com uma instabilidade gerada por crises políticas e governamentais (Souza, et al., 2020).

Desde o aparecimento do novo coronavírus, as redes mundiais de comunicação iniciaram uma cobertura massiva sobre as informações obtidas acerca da doença divulgando cada nova descoberta e como esta poderia auxiliar no combate a expansão da pandemia. O Brasil apresentou um comportamento semelhante ao restante das redes mundiais, contudo a divulgação de publicações, notícias e reportagens foram recebidas inicialmente pela população com certo tom de incredibilidade, não acreditando na gravidade da doença. Com o decorrer do tempo e o avanço desordenado da doença, a percepção da realidade por parte da população foi alterada, favorecendo a adoção de medidas de segurança e o surgimento de novas emoções e sentimentos em relação a pandemia (Pimentel & Silva, 2020).

Durante este mesmo período de tempo, onde o caos entre as pessoas era disseminado, cientistas e estudiosos desenvolviam diversos estudos científicos em busca de uma vacina eficaz no combate ao novo coronavírus. O fato de quatro dessas pesquisas terem sido realizadas no Brasil aumentava nos brasileiros a proporção das emoções que eram produzidas e compartilhadas diariamente através das redes sociais (Castro, 2021).

Nos últimos anos as redes sociais tem aumentado a sua adesão, tornando-se algumas das principais plataformas de comunicação utilizadas para compartilhar informações e para que as pessoas possam se expressar (Silva & Malheiros, 2019). Dentro do contexto da pandemia e do período de isolamento social durante o *lockdown* realizado em quase todos os países, o volume de dados produzidos através das redes sociais aumentou significativamente e as tornaram preciosas fontes de dados a serem analisados.

Ao se considerar a grande quantidade de emoções produzidas na população, tem se a análise de sentimentos, uma área da ciência da computação, que se tornou de grande popularidade nos trabalhos em que se desejava encontrar opiniões e sentimentos que poderiam ser associados a produtos, personagens ou lugares, e que fossem expostos em mídias sociais. Porém, atualmente, esta área pode ser aplicada a trabalhos de âmbito social e que beneficiem a população (Ramos & Freitas, 2019).

Para realizar tal trabalho, a análise de sentimentos está sendo associada a mineração de textos, um conjunto de métodos multidisciplinares utilizados para extrair regularidades, encontrar tendências ou padrões em dados textuais, identificando assim informações úteis que dificilmente seriam descobertas utilizando métodos tradicionais de consulta (Morais & Ambrósio, 2007).

Dentro de tais contextos o presente trabalho realiza a análise dos sentimentos que a população brasileira desenvolveu em relação às vacinas criadas para o combate da Covid-19, por meio da utilização das técnicas de análise de sentimento e mineração de dados. A construção do banco de dados ocorreu através da captação de postagens públicas disponibilizadas pela API do Twitter durante o período que abrange 03 de novembro de 2020 a 20 de fevereiro de 2021.

É importante destacar que foi construído um banco de dados que represente uma amostra aleatória das postagens realizadas neste período de tempo, não compreendendo a totalidade das postagens realizadas. Dentro deste período se encontram a data de aprovação e aplicação da primeira vacina contra Covid-19 no mundo e no Brasil.

2. Metodologia

O algoritmo desenvolvido durante a pesquisa foi construído na linguagem de programação Python. A linguagem Python é simples, clara e objetiva, contudo é também muito poderosa e forte, possibilitando o seu uso em diversas aplicações. A utilização desta vem crescendo em áreas da computação como inteligência artificial, banco de dados, jogos, aplicativos para celulares, entre outros (Menezes, 2010). As linguagens de programação necessitam de um Ambiente de Desenvolvimento Integrado, ou *Integrated Development Environment* (IDE), sendo utilizada nesta pesquisa o *Jupyter Notebook*, uma ferramenta que possui códigos iterativos baseados na *Web*, os denominados notebooks, que oferecem suporte para dezenas de linguagens de programação, incluindo o Python (Mckinney, 2019).

Para a construção do banco de dados a ser analisado, o Twitter foi escolhido como a rede social a ser utilizada devido ao fato de suas publicações serem de acesso público e possuir diversas formas de se obter tais publicações, como por exemplo, através da sua API. O Twitter é uma plataforma popular nos dias atuais e se destaca pelas interações praticamente imediatas. Neste, cada usuário possui uma espécie de microblog onde pode escrever mensagens de até 140 caracteres, os *'tweets'* (Franco & Adaniya, 2018). Atualmente a plataforma do Twitter já disponibiliza opções de contas onde o usuário possui a capacidade de escrever mensagens com uma quantidade maior de caracteres. Serão apresentados nos tópicos a seguir uma breve contextualização da pandemia juntamente com as técnicas de mineração de textos e análise de sentimentos que foram utilizadas na pesquisa.

2.1 A Pandemia da Covid-19

Em dezembro de 2019 uma nova espécie do coronavírus, denominada SARS-CoV-2, foi identificada pela primeira vez em Wuhan, uma cidade localizada na província de Hubei, na China. Tal espécie de vírus era responsável por desencadear uma doença respiratória agressiva, denominada Covid-19, que rapidamente se espalhou por todo o planeta tornando-se o maior problema sanitário do século XXI (Cardoso, et al., 2021). Segundo a Organização Mundial de Saúde (OMS) a doença pode se manifestar de diversas formas e níveis, sendo os principais: pessoas assintomáticas, quadros gripais leves e síndrome respiratória aguda. Deve-se ainda salientar que existe maior suscetibilidade de contágio e quadros graves da doença em pessoas idosas, acima de 60 anos de idade, e pessoas com comorbidades (OMS, 2020).

Desde o aparecimento do novo coronavírus, as redes mundiais de comunicação iniciaram uma cobertura massiva sobre as informações obtidas acerca da doença divulgando cada nova descoberta e como esta poderia auxiliar no combate a expansão da pandemia. O Brasil apresentou um comportamento semelhante ao restante das redes mundiais, inclusive no que abrange a divulgação de publicações, notícias e reportagens que foram recebidas inicialmente pela população com certo tom de incredibilidade, não acreditando na gravidade da doença. Com o decorrer do tempo e o avanço desordenado da doença, a percepção da realidade por parte da população foi alterada, favorecendo a adoção de medidas de segurança e o surgimento de novas emoções e sentimentos em relação a pandemia (Pimentel & Silva, 2020).

De acordo com o avanço da doença e sua expansão no cenário global, as populações foram submetidas a um período de intenso temor, acarretando pensamentos irracionais e atitudes antes consideradas inimagináveis. A população brasileira além do temor pelo desconhecido e de angústia por medo da morte, enfrentava ainda um agravante causado pelas inúmeras fake News espalhadas pelos meios de comunicação e pelas redes sociais. Assim, sentimentos como ansiedade, estresse, pânico, medo, raiva, entre outros, se tornaram extremamente comuns no cotidiano da população (Pimentel & Silva, 2020).

Os problemas gerados pela falta de informações precisas no Brasil se tornaram tão amplos que diversos meios de comunicação desenvolveram seus próprios painéis para acompanhar o avanço da pandemia, visto que o Ministério da Saúde não divulgava as informações dentro dos prazos que eram necessários. O Conselho Nacional de Secretários de Saúde

desenvolveu o Painel CONASS de acompanhamento ao avanço da Covid 19 no Brasil, sendo que em 30 de março de 2022 já eram contabilizados quase 30 milhões de brasileiros infectados pelo vírus e 659.504 vidas perdidas para a doença (CONASS, 2022).

Desde a descoberta do novo coronavírus, iniciou-se uma corrida pelo desenvolvimento de uma vacina que fosse eficaz para o combate da doença. Foram registrados junto a OMS cerca de 200 projetos de vacinas para a Covid-19. Dentre estes, quatro foram efetuados no Brasil, fazendo com que a população estivesse familiarizada com as etapas do processo de desenvolvimento (Castro, 2020).

A existência dessa grande quantidade só foi possível devido ao gigantesco investimento realizado pelos governos dos países desenvolvidos, de algumas organizações não governamentais e das empresas farmacêuticas, que por muitas vezes trabalharam entre si para obter uma vacina no menor período de tempo possível. Uma medida inédita foi adotada em relação a vacina da Covid 19: uma ação liderada pela OMS denominada *Covax Facility*, que possuía o objetivo de acelerar o desenvolvimento e a fabricação de vacinas contra a Covid-19 para garantir que todos os países tivessem acesso a elas toda a população mundial fosse vacinada (Domingues, 2021).

Ainda no ano de 2020, as primeiras vacinas obtiveram autorização para uso emergencial em alguns países europeus e nos Estados Unidos. Contudo, no Brasil o uso emergencial das vacinas só foi autorizado pela Agência Nacional de Vigilância Sanitária, a ANVISA, em 17 de janeiro de 2021. Após alguns minutos a enfermeira Mônica Calazans foi a primeira brasileira a ser vacinada em território nacional (Castro, 2020).

2.2 Mineração de Textos

A Mineração de Textos (MT), ou *text mining*, é considerada pelos estudiosos como um Processo de Descoberta de Conhecimento, onde são utilizadas técnicas de análise e extração de dados a partir de documentos textuais como textos, frases ou apenas palavras. A mineração de textos é um método multidisciplinar que utiliza técnicas de Estatística, Informática, Programação e linguística para processar textos identificando informações úteis que estavam implícitas nos textos e que dificilmente seriam descobertas utilizando métodos tradicionais de pesquisa (Morais & Ambrósio, 2007).

O principal objetivo da mineração de texto é encontrar termos relevantes em documentos textuais com grande volume de dados e assim estabelecer padrões e relacionamentos entre eles. Tais dados se encontram arquivados de forma estruturada, não estruturada ou semi estruturada. É importante destacar que a mineração de textos não é um mecanismo de busca, como os encontrados na internet, pois a mineração auxilia o usuário a descobrir informações que antes eram desconhecidas. Já nos mecanismos de busca, o usuário já sabe o que eles desejam procurar (Pezzini, 2016).

Existem duas abordagens principais para o processo de mineração em dados textuais que são: a Análise Estatística, que está diretamente ligada a frequência de aparição de um determinado termo, não se atentando ao contexto em que este se encontra inserido no texto; e a Análise Semântica, que se preocupa com a funcionalidade dos termos nos textos por meio dos significados morfológico, sintático, semântico, entre outros que podem ser empregados no contexto da pesquisa realizada. Tais abordagens podem ser aplicadas sozinhas ou em associação (Brito, 2017).

A aplicação inicialmente é realizada em um grupo teste para que o algoritmo seja treinado, e se o mesmo não alcançar um resultado satisfatório, as etapas devem ser repetidas até que o objetivo seja alcançado. Após o algoritmo se encontrar apto, este pode ser aplicado aos dados que se deseja analisar (Brito, 2017).

A Figura 1 apresenta as etapas do processo de mineração de texto segundo o trabalho de Brito (2017), em formato de diagrama para facilitar a visualização das mesmas.

Figura 1. Diagrama do Processo de Mineração de Texto



Fonte: Adaptado de Brito (2017).

Cada uma das etapas apresentadas na Figura 1 tem grande importância dentro do processo e possui as suas peculiaridades. As etapas podem ser definidas como:

Seleção – Se trata da extração e coleta dos dados que serão utilizados na pesquisa, criando uma base de dados textuais que represente toda a população que será analisada, ou seja, a criação do corpus da pesquisa.

Pré-processamento – Consiste na filtragem e limpeza dos dados, eliminando informações desnecessárias ao algoritmo.

Transformação – Se trata da classificação dos dados, separando-os de acordo com as características que se pretende observar.

Mineração – Esta etapa está diretamente ligada às técnicas de aprendizagem de máquina para obtenção de novos conhecimentos.

Avaliação – É a etapa onde será validada a eficiência do processo como um todo, analisando os dados que foram obtidos após a aplicação dos algoritmos.

2.3 Análise de Sentimentos

A análise de sentimentos, também denominada pelos estudiosos como Mineração de Opinião, é um método utilizado para identificar emoções e opiniões em textos, a partir do uso das técnicas de Mineração de Textos, Informática, Linguística, entre outros (Brito, 2017). A análise de sentimentos é considerada pelos estudiosos como um subcampo do Processamento de Linguagem Natural, uma área definida como conjunto de técnicas computacionais criadas para analisar e representar textos naturais, de forma a obter um processamento de linguagem o mais próximo possível do humano (Olenski, et al., 2020).

Sentimento é definido por Liu (2012) como uma opinião ou uma avaliação sobre algum aspecto ou objeto. Já a análise do sentimento, envolve o PLN, para que extrair as emoções expressas principalmente em textos online, onde tal área tem sido bastante utilizada para verificar a popularidade de pessoas, objetos, lugares e situações (Brito, 2017).

A classificação dos sentimentos identifica a polaridade do texto analisado, mesmo que este seja um documento, uma frase ou apenas uma palavra, classificando-o em positivo, negativo ou neutro. Para realização de uma classificação utilizando sentimentos mais amplos, é necessário realizar alterações nos algoritmos para que estes sejam capazes de determinar tais emoções (Liu, 2012).

Segundo Silva (2016) a análise de sentimentos aplicada a redes sociais pode ser caracterizada em quatro tipos:

- **Supervisionada** – Esta técnica é realizada por meio do uso de algum algoritmo de aprendizagem de máquina e exige a existência de uma base de dados para treinamento que seja previamente rotulada por um especialista.

- Guiada pelo uso de léxico – Neste modelo é utilizada uma lista de termos classificados, por alguém especializado, em positivos e negativos que direcionará o processo de identificação da polaridade da base de dados.
- Supervisionada Híbrida - Nesta abordagem são aplicados simultaneamente um algoritmo de aprendizagem de máquina e uma lista de léxicos.
- Baseada em Grafos – Esta técnica utiliza as características específicas das redes sociais, como os relacionamentos e interações entre os usuários para realizar a análise.

A análise de sentimentos é um método que utiliza a mineração de textos em sua execução, além de apresentar também suas técnicas específicas de acordo com a aplicação desejada (Brito, 2017).

3. Resultados e Discussão

Inicialmente foi realizada a construção do primeiro algoritmo, sendo este responsável por realizar a coleta de dados, através da aplicação da biblioteca para linguagem Python, *Tweepy*. Para realizar o processo de coleta é necessário possuir uma conta no Twitter para desenvolvedores. Após a aprovação da conta, o usuário pode criar um aplicativo para ser vinculado a IDE escolhida pelo mesmo. O Twitter criará automaticamente credenciais para liberar o acesso do usuário a uma quantidade pré-determinada de tweets, de acordo com o modelo de conta solicitado. As credenciais fornecidas pela plataforma são individuais e não podem ser compartilhadas com outras pessoas.

A construção da base de dados para testes foi realizada no período de primeiro de abril de 2021 a 07 de abril de 2021. A coleta foi realizada através da busca por *endpoints*, por meio do *Tweepy*. Nesse momento é importante destacar que a API do *Twitter* passou por atualizações recentes e atualmente o retorno para cada pesquisa realizada é de no máximo 100 *tweets*, selecionados de acordo com perfis de interação próprios da plataforma.

Os *endpoints* utilizados para esta busca foram: 'pandemia' e 'Covid19'. Não foram consideradas as *hashtags* nesse momento. Como resultado desta busca inicial, obteve-se 700 *tweets*, nos quais foram realizados os tratamentos iniciais de limpeza, onde foram excluídas as publicações que não possuíam classificação de sentimentos, como os informativos publicados por diversos perfis de canais de comunicação.

Após a exclusão restaram 248 *tweets* que foram manualmente classificados, dos quais 150 são negativos, 52 positivos e 46 nulos. A presença de um número de publicações negativas muito superior as demais pode ser explicada pelo fato dos sentimentos que a pandemia produzia na população eram em sua essência de medo, tristeza, angústia, entre outros já apresentados neste trabalho. A Tabela 1 traz exemplos dos tweets classificados na base de dados para testes.

A construção da base de dados para a pesquisa foi realizada no período de 03 de novembro de 2020 a 20 de fevereiro de 2021. Esta segunda coleta também foi realizada através da busca por *endpoints*, por meio da biblioteca *Tweepy*. A base de dados para pesquisa foi dividida em 3 grandes grupos, onde cada grupo compreende o período de 5 semanas epidemiológicas que abrangem em três momentos específicos, sendo eles:

- Primeiro período (03 de novembro de 2020 – 07 de dezembro de 2020) – este período compreende o momento entre a descoberta oficial da primeira vacina no mundo e um dia antes da primeira pessoa ser vacinada contra a Covid-19.
- Segundo período (08 de dezembro de 2020 – 16 de janeiro de 2021) – neste ponto é importante verificar a expectativa dos brasileiros, pois já existia uma vacina sendo aplicada pelo mundo, porém não havia a liberação da ANVISA para que o processo de vacinação fosse liberado no Brasil.
- Terceiro período (17 de janeiro de 2021 – 20 de fevereiro de 2021) – a vacinação é liberada pela ANVISA e é a

primeira pessoa é vacinada no Brasil contra a Covid-19. Os momentos após essa liberação são de suma importância para serem visualizados devido as ramificações dos problemas que foram ocorrendo dia a dia neste processo.

Tabela 1. Exemplo de *tweets* da Base de Testes.

Tweets	Polaridade
só quero acordar desse pesadelo chamado pandemia	Negativo
milhões de pessoas em situação de imigração se veem desumanizadas e desprotegidas, cenário ainda mais grave no meio de uma pandemia que hierarquiza vidas	Negativo
Nessa pandemia to aprendendo valorizar cada vez mais o abraço, as risadas, as alegrias as tristezas a oportunidade de congregar e de conviver. Quando tudo isso passar não quero perder meu tempo com coisas fúteis, quero investir meu tempo em pessoas!	Positivo
E em meio a essa pandemia, só tenho a agradecer e ficar feliz por poder estar com minha família e ter uma ceia feliz	Positivo
Fome atingiu 19 milhões de brasileiros durante a pandemia, diz pesquisa.	Neutro
Gilmar Mendes contraria Kassio Nunes e respalda fechamento de igrejas e templos durante pandemia	Neutro

Fonte: Autoria Própria (2022).

Ao se analisar a Tabela 1 é importante salientar que durante a pesquisa, bem como na composição do artigo, os erros de ortografia e gramática que as publicações possuíam foram mantidos.

Os *endpoints* utilizados para a construção do banco de dados da pesquisa foram: ‘vacina’, ‘coronavírus’ e ‘Covid19’. Nesta busca também não foram consideradas as hashtags. Como resultado desta busca inicial, obteve-se 11000 *tweets*, nos quais foram realizados os tratamentos abaixo especificados. Após o tratamento inicial, a base de dados para a pesquisa era constituída de 5041 *tweets*.

3.1 Algoritmo de Análise de Sentimentos

A Figura 2 apresenta a estrutura de execução do algoritmo construído nesta pesquisa. O algoritmo foi aplicado no conjunto de dados obtidos na etapa de coleta de dados, classificado como Banco de Dados não Relacional, pois não existe uma relação entre pré-estabelecida entre os dados.

O algoritmo é iniciado com a limpeza dos tweets mantendo apenas aqueles que possam verdadeiramente fornecer uma informação quanto a opinião dos usuários, e para isso foram excluídas as publicações que possuíam conteúdos informativos, como as realizadas pelas páginas de meios de comunicação.

Figura 2. Estrutura do Algoritmo



Fonte: Autoria Própria (2022)

A Figura 2 apresenta sobre o modelo de um fluxograma as etapas a serem executadas no algoritmo. É importante ressaltar que tal fluxograma mostra uma versão do fluxograma apresentado na Figura 1, porém com as técnicas aplicadas no algoritmo criado nesta pesquisa.

Figura 3. Exemplo de *tweet* a ser excluído.



Fonte: Twitter (2022).

A Figura 3 representa um exemplo de *tweet* a ser excluído, pois apesar de ele possuir uma informação relevante, este não apresenta a opinião ou sentimento de nenhum usuário, não sendo assim considerado apto para constituir o banco de dados da pesquisa.

Para possibilitar esta etapa de exclusão foram identificados os *users* das páginas de comunicação e os *tweets*

retornados por tais usuários eram excluídos, até que restassem apenas as publicações que continham opiniões. A Figura 4 apresenta alguns dos *users* excluídos.

Figura 4. *Users* excluídos.

@CNNBrasil @UOL @exame @RadioBandNewsBH @UOLNoticias
 @o_antagonista @Estadao @folha @otempo @opovo
 @g1saopaulo @JornalOGlobo @ESPNBrasil

Fonte: Autoria Própria (2022).

Para realizar o Processamento de Linguagem Natural foi utilizada a biblioteca NLTK. Esta biblioteca está disponível para a linguagem de programação Python e contém por volta de 35 módulos. Cada módulo possui seus diversos submódulos que realizam múltiplas tarefas de PLN, como *tokenização*, remoção de *stopwords*, e possui inclusive um módulo para análise de sentimentos, entre outros módulos (Araújo, 2017).

Os módulos aplicados neste algoritmo, na etapa de Pré-Processamento são: a remoção de caracteres especiais e pontuação, a remoção de *stopwords*, a *tokenização* e a conversão para letras minúsculas.

Após a etapa de pré-processamento é realizada a aplicação do dicionário de léxicos SentiLex-PT aos *tweets*. A aplicação deste consiste na busca de cada palavra que compõe o *tweet* dentro da estrutura do dicionário. Neste modelo de avaliação cada *tweet* é entendido como um vetor, onde as palavras são suas estruturas. Cada palavra recebe um valor e o resultado obtido é o somatório do valor das palavras. A classificação e o valor atribuído as palavras estão expressos na Tabela 2. Entretanto, é necessário destacar que o dicionário já apresenta as palavras com seus devidos valores definidos.

Tabela 2. Valor atribuído as palavras de acordo com o dicionário.

Tipo de palavra	Valor
Palavras que expressam sentimentos negativos	-1
Palavras que expressam sentimentos positivos	+1
Palavras que não expressam sentimentos	0

Fonte: Adaptado de Silva (2016).

A Tabela 2 informa os valores que são concedidos a cada palavra de acordo com o dicionário SentiLex-PT. Desta maneira um *tweet* vai ter sua classificação expressa pelo somatório dos valores individuais de cada uma das suas palavras

Na sequência das etapas da MT, é realizada a avaliação dos resultados obtidos. Tal avaliação será realizada através da utilização da métrica acurácia. A acurácia encontrada no algoritmo aplicado é 68,5 %. O valor pode ser considerado pequeno quando comparado aos algoritmos que utilizam técnicas de aprendizagem de máquina, entretanto, é considerado adequado ao se tratar da aplicação de dicionários de léxicos.

A Tabela 3 e a Figura 5 apresentam os resultados obtidos para a análise de sentimento aplicada no banco de dados da

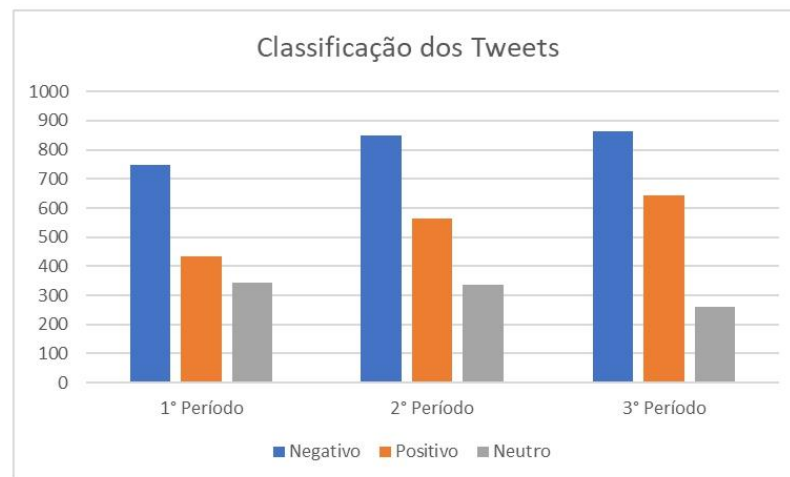
pesquisa, conforme a divisão de períodos supracitada. Para justificar certos comportamentos encontrados, como a presença majoritária de tweets negativos em ambos os períodos, é importante salientar que essa negatividade pode não estar diretamente associada a vacina, ou seja, a pessoa pode não estar defendendo a não aplicação da vacina e sim estar discorrendo sobre algum acontecimento acerca da vacinação que causa sentimentos negativos na pessoa que realizou a publicação.

Tabela 3. Valores obtidos pelo algoritmo.

Período	Positivo	Negativo	Neutro	Total
Primeiro	435	747	343	1525
Segundo	562	851	336	1749
Terceiro	643	865	259	1767

Fonte: Autoria Própria (2022)

Figura 5. Resultado da pesquisa.



Fonte: Autoria Própria (2022).

Verificando os resultados apresentados na Tabela 3 e na Figura 5, pode-se perceber a tendência predominante dos *tweets* em serem negativos, porém ao se analisar os *tweets* positivos, é visível um aumento considerável na quantidade de ocorrência dos mesmos. Tal aumento pode representar que a população se encontrava otimista quanto ao fato de vacinas terem tido o seu uso autorizado, bem como o início da vacinação no Brasil.

A seguir serão apresentados alguns temas e notícias presentes no período de coleta de dados e que possivelmente seriam responsáveis por algumas das variações que geraram essa negatividade predominante:

- A ocorrência de eleições municipais no Brasil.
- A interrupção que a ANVISA realizou nos estudos das vacinas realizados no território brasileiro.
- Caos instaurado em Manaus devido à falta de oxigênio e de leitos para atendimento dos pacientes.
- Declarações diversas realizadas pelo presidente Jair Bolsonaro, que causaram grande alvoroço nas redes sociais, como a aplicação da vacina que supostamente transformaria as pessoas em 'jacaré'.
- Com o início da vacinação, surgem novos problemas como as pessoas que burlavam as filas de pessoas

prioritárias para receberem as vacinas antes da sua convocação oficial.

- Profissionais da saúde que foram gravados no momento da vacinação e não injetaram nenhum medicamento nos pacientes.

Figura 6. Quantidade de óbitos por semana epidemiológica.



Fonte: Adaptado CONASS (2022).

Figura 7. Quantidade de casos novos por semana epidemiológica.



Fonte: Adaptado CONASS (2022).

Para proporcionar uma visão mais clara do cenário em que se encontrava o Brasil no momento em que tais dados foram coletados, a Figura 6 apresenta a quantidade de óbitos contabilizados nas semanas epidemiológicas em que se realizou a pesquisa e a Figura 7 expõe a quantidade crescente de casos de pessoas infectadas. A vacina teve seu uso emergencial liberado pela ANVISA no momento que o Brasil iniciava a curva ascendente de casos da segunda onda de contaminação. É importante ainda destacar que o ápice de óbitos no país ocorreu após a liberação da vacina, entretanto não havia imunizante o suficiente para que toda a população fosse vacinada em um curto período de tempo.

4. Conclusão

A pandemia do Covid-19 já é considerada por vários estudiosos como o maior problema sanitário do XXI. A sua rápida expansão no cenário global gerou um estado de caos na população mundial, que de repente se encontra isolada, privada de sua liberdade e do contato social. Tais situações contribuíram para criar um misto de sentimentos nas pessoas, gerado principalmente pelo temor do desconhecido, visto que não se conhecia muito sobre o vírus, o seu comportamento e tratamentos que fossem eficazes.

Assim sendo, tal pesquisa foi desenvolvida com o objetivo de realizar um estudo sobre a percepção dos brasileiros acerca da vacina contra a Covid-19, baseado nas postagens realizadas pela população no Twitter, no período marcado pela aprovação das primeiras vacinas no Brasil e no mundo. O algoritmo desenvolvido apresentou acurácia de 68,5%, sendo considerado um valor adequado para este modelo de pesquisa.

Para dar continuidade ao trabalho, em uma pesquisa futura seria viável o desenvolvimento de um algoritmo baseado na aprendizagem de máquina para que sejam realizadas comparações entre seus respectivos resultados. Tal desenvolvimento necessita de uma base de dados para treinamento do algoritmo que possua uma quantidade elevada de publicações já classificadas de acordo com a polaridade possibilitando uma análise adequada do modelo.

Referências

- Araújo, L. G. de A. (2017). *Sentimental! Versão 2: Desenvolvimento de Análise de Sentimentos em Python*. Trabalho de Conclusão de Curso em Ciência da Computação do Centro Universitário Luterano de Palmas. Palmas.
- Brito, E. M. N. (2017). *Mineração de Textos: detecção automática de sentimentos em comentários nas mídias sociais*. Dissertação: Programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Universidade Fundação Mineira de Educação e Cultura — FUMEC. Belo Horizonte.
- Cardoso, R. F. et al. (2021). COVID-19: An epidemiological challenge. *Research, Society and Development*, 10(7), e32110716313. DOI: 10.33448/rsd-v10i7.16313.
- Castro, R. (2021). Vacinas contra a Covid-19: o fim da pandemia? *Physis: Revista de Saúde Coletiva*, 31, e310100.
- Conselho Nacional de Secretários de Saúde - CONASS. (2022). *Painel Covid-19*. Acesso em 17 de julho de 2022 em <https://www.conass.org.br/painelconasscovid19>.
- Domingues, C. M. A. S. (2021). Desafios para a realização da campanha de vacinação contra a COVID-19 no Brasil. *Cadernos de Saúde Pública*, 37.
- Franco, R. B., & da Costa Adaniya, M. H. A. (2018). Sistemas de análise de sentimentos usando dados do Twitter. *Revista Terra & Cultura: Cadernos de Ensino e Pesquisa*, 34(esp.), 111-118.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.
- Mckinney, W. (2019). *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora.
- Menezes, N. N. C. (2010). *Introdução a programação com Python*. São Paulo: Novatec.
- Morais, E. A. M., & Ambrósio, A. P. L. (2007). *Mineração de textos*. Relatório Técnico: Instituto de Informática da Universidade Federal de Goiás.
- Olenscki, J., Xavier, F., Acosta, A., Saraiva, A., & Sallum, M. (2020). Aplicação de análise de sentimentos no Twitter para avaliação da percepção pública quanto a cloroquina. Em *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde* (pp. 500-505). SBC.
- Organização Mundial da Saúde - OMS. (2020). *Coronavirus disease 2019 (COVID-19): Situation Report –51*. Acesso em 15 de julho de 2022 em https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid19.pdf?sfvrsn=1ba62e57_10.
- Pezzi, A. (2017). Mineração de textos: conceito, processo e aplicações. *Revista Brasileira De Contabilidade E Gestão*, 5(8), 58-61.
- Pimentel, A. do S. G., & Silva, M. de N. R. M. de O. (2020). Saúde psíquica em tempos de Corona vírus. *Research, Society and Development*, 9(7), e11973602. DOI: 10.33448/rsd-v9i7.3602.
- Ramos, B., & Freitas, C. (2019). “Sentimento de quê?” uma lista de sentimentos para a Análise de Sentimentos. *STIL*, 15-18.
- Silva, E. P. da & Malheiros, Y. (2019). *Um conjunto de dados extraído do Twitter para análise de sentimentos na língua portuguesa*. Trabalho de Conclusão de Curso em Sistemas de Informação da Universidade Federal da Paraíba.
- Silva, N. F. F. (2016). *Análise de sentimentos em textos curtos provenientes de redes sociais*. Tese de Doutorado: Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo – ICMC-USP. São Carlos.
- Sousa, A. R. D., Carvalho, E. S. D. S., Santana, T. D. S., Sousa, Á. F. L., Figueiredo, T. F. G., Escobar, O. J. V., Mota, T. N. & Pereira, Á. (2020). Sentimento e emoções de homens no enquadramento da doença Covid-19. *Ciência & Saúde Coletiva*, 25, 3481-3491.
- Twitter. (2022). *Página Inicial*. Acesso em 15 de julho de 2022 em <https://twitter.com>.



UFVJM