MonkeyPox Feelings - Uma análise de sentimentos de tweets sobre a MonkeyPox

Fillipe Dornelas[1], Eliel Silva[1], Angelo Samir M. Miguel[2]

[1]Programa de Pós de Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

[2]Departamento Tecnologia Farmacéutica
Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil

{elielsilva, fdornelas}@ufrj.br, asmmiguel@gmail.com

Abstract. *This work presents an exploratory analysis of data about MonkeyPox from the collection of Tweets using the official Twitter channel, as well as the application of Machine Learning techniques to present the most relevant terms commented and the feelings predominant. It was found that the predominant feeling was anger for general tweets, when specific topics such as "gay couple" were addressed, the feeling of anger remains and appears to be correlated with presidential speeches in the period.*

Resumo. *Este trabalho tem o objetivo de apresentar uma análise exploratória de dados a respeito de Varíola dos Macacos (MonkeyPox) a partir da coleta de Tweets utlizando o canal oficial do Twitter, como também a aplicação de técnicas de Machine Learning para apresentar os termos mais relavantes comentados e os sentimentos predominantes. Verificou-se que o sentimento predominante foi o de Raiva para tweets gerais, quando abordados temas específicos como "casal gay" o sentimento de raiva se mantém e aparenta correlação com discursos presidenciais no período.*

1. Introdução

A *MonkeyPox* - ou varíola dos macacos em tradução livre - é uma zoonose causada pelo vírus monkeypox, do gênero *Orthopoxvirus*, pertencente à família Poxviridae. Por ser uma zoonose, casos em seres humanos eram registrados em uma frequência muito baixa, tendo só o primeiro caso identificado em 1970, em uma criança, na República Democrática do Congo [WHO]. No entanto, em 2022 o mundo assistiu o surgimento de um surto que teve seu nascimento em Londres, em 5 de maio, aonde o paciente zero desenvolveu lesões cutâneas ao voltar de uma viagem à Nigéria. No dia 23 de Julho, a Organização Mundial da Saúde declarou que a monkeypox constitui uma Emergência de Saúde Pública de importância mundial, levando a vários governos a oficialirazem políticas de monitoramento e controle a nível de saúde pública. Somado a esse cenário, há também a grande quantidade de informações que circula diariamente nas redes sociais. Uma vez que, após o cenário de pandêmico houve o aumento do número de horas que um brasileiro gasta conectado à internet, compreender a natureza das informações contidas em cada publicação pode auxiliar a entender como que eventos relacionados à *Monkeypox* impactam as opiniões de usuários finais nas Redes Sociais.

O Twitter é uma das maiores redes sociais no mundo em termos de usuários ativos - em 2021 a rede contava com 211 milhões de usuários ativos. Essa rede permite compartilhar variados tipos de conteúdo, seja através de publicações textuais conhecidas como *tweets*, ou do compartilhamento de um *tweet* de um usuário. Esse tipo de publicação é conhecido como *retweet*. Devido a essa característica, o Twitter pode ser considerado um ambiente rico para entender as distintas opiniões, sentimentos dos usuários sobre assuntos que estão em alta no mundo. Por último, mas não menos importante, a característica textual das publicações dessa rede permite entender os principais tópicos e palavras relacionados à um tema conhecer quais palavras e tópicos foram as mais importantes.

Diante desse contexto, a proposta desse artigo consiste em utilizar técnicas de aprendizado de máquina para entender quais foram as palavras chaves e sentimentos dos usuários no Twitter eme relação ao aumento no número de casos de *Monkeypox* em humanos no ano de 2022. Nesse sentido, alguma Perguntas de Pesquisa (PP) foram construídas para delimitar e orientar o escopo do trabalho:

- **PP1**: Como montar uma base de dados com *tweets* em pt-Br relacionados à *Monkeypox*?
- **PP2**: Como descobrir os termos chaves contidos nos *tweets* em português *retweets* relacionados à *Monkeypox*?
- **PP3**: Como estão distribuídos os sentimentos na base de dados coletada?

Para responder as questões de pesquisa definidas, esse trabalho está estruturado da seguinte maneira: No capítulo 2 está descrito os trabalhos relacionados com o estudo feito nesse trabalho; no capítulo 3 foi descrito a metodologia utilizada durante o estudo para a construção do dataset e execução dos modelos; no capítulo 4, falamos sobre os resultados obtidos; no capítulo 5, são feitas as considerações finais sobre o trabalho.

2. Trabalhos Relacionados

[Thakur 2022] em seu trabalho, entregou o primeiro dataset público de *tweets* relacionados ao surto de *MonkeyPox*. Seu trabalho teve uma contribuição notória por ser um dos primeiros do tipo na área da ciência de dados a propor a formação de uma base dados de tweets no contexto do aumento dos casos dessa enfermidade no mundo. A diferença do presente trabalho para o de Nirmalya, é na aplicação do uso dos tweets e escopo da coleta dos mesmos. Enquanto Nirmlya focou na construção de um dataset esse trabalho busca focar na aplicação de métodos de processamento de linguagem natural para tweets coletados via a interface oficial de coleta de *tweets* chamada *Twitter Academic Research API*.

[Pinto et al. 2020] realizaram uma abordagem de mineração em postagens do *Twitter* para a identificação de sentimentos e tópicos contidos em publicações relacionadas a pandêmia da COVID-19. Em seu trabalho Pinto et al. também buscaram a construção de seus próprios datasets para a aplicação dos modelos de processamento de texto. O presente trabalho se diferencia no que diz respeito ao tema e aos métodos matemáticos utilizados para a extração de informações na massa de dados. Enquanto, o primeiro trabalho foi fundamentado em uma base de dados relacionada à pandemia da COVID-19, esse trabalho foi construído observando uma base de dados construída com *tweets* sobre o surto da *MonkeyPox*.Por último, Pinto focou na utilização de algorítmos

para modelagem de tópicos, enquanto esse trabalho focou na na extração de termos importantes relacionados utilizando o algorítmo *Bag of Words*.

[Ortiz-Martínez et al. 2022] propuseram seu trabalho uma análise qualitativa e quantitava dos *tweets* relacionados a *MonkeyPox* classificando-os em publicações úteis - postagens baseadas em fatos verídicos - e em publicações de desinformação segundo critérios da medicina. Em seu trabalho, Martinez et al. buscaram classificar: as contas que realizavam as postagens em formais e informais. Enquanto objetivo deste trabalho era quantificar a proporção de contas informais que realizavam publicações de desinformação sobre a *MonkeyPox* em relação a publicações informativas, o presente trabalho visa analisar os sentimentos dos usuários expressos nas redes sociais. Outra considerável diferença entre os estudos é quantidade de dados utilizados para a análise, no trabalho usado como referência a base de dados é composta por 100 tweets enquanto o presente estudo conta com um dataset de 6445 tweets.

3. Metodologia

Ciência de Dados é uma disciplina que se fundamenta em conhecimento de diversas áreas de conhecimento com o objetivo de solucionar problemas do mundo real. Ela Compreende uma série de diferentes áreas de especialização e habilidades, combinando-as para resolver problemas e melhorar e otimizar processos [Pinheiro 2021]. Devido a essa natureza multi-disciplinar dos projetos de ciência de dados, o uso de metodologias tem se tornado uma ferramenta eficiênte para garantir a reprodutibilidade dos experimêntos. Ness contexto, esse trabalho foi desenvolvido em observando o *Data Science Lifecycle*, em outras palavras, trata-se do fluxo de processo que compreende todos os passos para que um projeto de Ciência de dados possa executar a coleta, exploração e modelagem dos dados para localizar a resposta de uma pergunta de um problema de negócio.

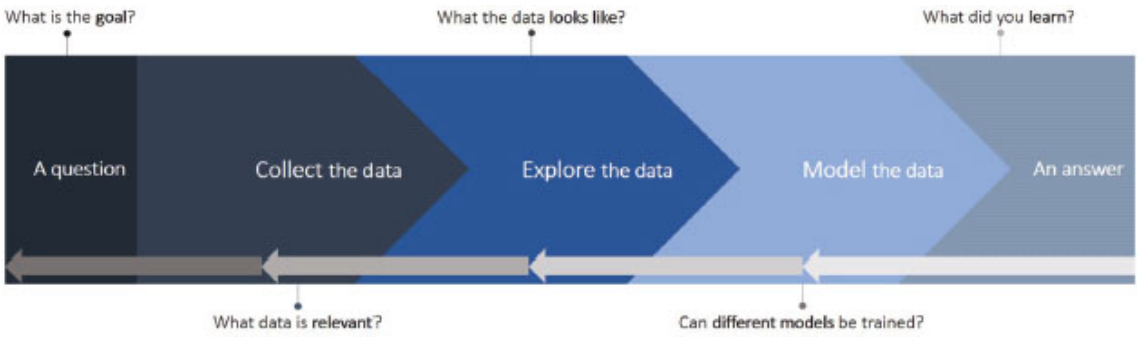

Figure 1. Ciclo de vida dos projetos de ciência de dados - Adaptado de *Introduction to Data Science*

3.1. Entendimento da(s) questão(ões)

O objetivo dessa etapa em um projeto de ciência de dados é dar aos membros da equipe que trabalharão no projeto a entender as perguntas problemas a ser resovido e quais serão os possíveis modelos de aprendizado de máquina a serem utilizados para atacar o problema. Em relação ao **PP1** embora já existam alguns trabalhos com datasets com tweets relacionados a *MonkeyPox*[Thakur 2022], a abordagem escolhida foi a construção de um novo dataset, pois se entende que para atender o principal requesito da **PP1** - ter tweets

em português - a construção de um novo dataset permitiria a construção da base dados observando as particularidades do idioma.

Referente as perguntas **PP2** e **PP3** através da análise dos trabalhos relacionados entende-se que o uma das possíveis métricas para entender **quais são os termos chaves** na base de dados é a frequência que termo ocorre na coleção de tweets. Do mesma forma, para a parte de entender a classificação do sentimento foi definido que os tweets seriam classificados de acordo com as seguintes classes: 'Neutro', 'triste', 'feliz', 'medo', 'raiva', 'nojo'.

3.2. Coleta de Dados

A segunda etapa em no fluxo de processo de ciência de dados é a coleta de dados. [Pinheiro 2021] explica que o objetivo principal dessa etapa é responder questões como: Quais dados são relevantes? Quantas fontes de dados estão envolvidas? Aonde os dados residem? O acesso aos dados já está disponível para a equipe? Há pontos a serem considerados sobre privacidade? Por fim, a coleta de dados trará como resultado uma coleção de objetos de estudo que permitirá a execução do experimento.

Conforme discutido anteriormente, o caminho a ser seguido foi o da construção de um novo dataset, pois os dados considerados relevantes para o contexto do trabalho são os tweets em português relacionados a temática *Monkeypox*. Já para a realização do acesso aos dados foi utilizada a *Twitter API for Academic Research*[1]. Trata-se de uma Interface de Programação de Aplicação anunciada em 2021 pelo Twitter focada nas necessidades da comunidade acadêmica. Essa API permite pessoas que estejam conduzindo pesquisas a acessar dados históricos e em tempo real do Twitter com recursos e funcionalidades adicionais que suportam a coleta de conjuntos de dados mais precisos, completos e imparciais. Por fim, do ponto de vista de privacidade o experimento foi executado em conformidade com as políticas de privacidade referente a *Twitter API for academic Research*.

3.3. Exploração dos dados

A terceira etapa no fluxo de processo de projetos envolvendo ciência de dados tem como objetivos medir e avaliar a qualidade e a adequação dos dados no contexto a ser trabalhado[Pinheiro 2021]. Para alcançar esses objetivos foi necessário responder as seguintes perguntas:

1. Que anomalias ou padrões são perceptíveis nos conjuntos de dados?
2. Há demasiadas variáveis para criar o modelo?
3. Há muito poucas variáveis para criar o modelo?
4. São necessárias transformações para ajustar os dados de entrada para o treinamento do modelo, como imputação, substituição, transformação, e assim por diante?

Para a realização do entendimento de padrões e anomalias, foi primeiro realizado o entendimento de um *tweet* como um objeto de análise. Tweets são basicamente a únidade básica que compõem os demais elementos do Twitter como plataforma. Um *tweet* é composto por diversos atributos como *id, created_at* e *text*. Como o escopo do trabalho

[1]https://developer.twitter.com/en/products/twitter-api/academic-research

esta limitado à somente a análise textual do contéudo do *tweet* foi definido que somente os atributos *text* e *lang* eram necessários para alimentarem os modelos.

No que cerne a necessidade de transformações e ajuste de dados, por se tratar de um trabalho de processamento de linguagem natural foi identificada a necessidade de realizar o pré-processamento de cada texto de cada Tweet. Para isso, a ferramenta *Natural Language Toolkit* (NLKT)[2] foi utilizada para executar as seguintes atividades:

1. Transformação de todas as palavras do tweet para caixa baixa;
2. Remoção de *stopwords*;
3. Remoção de caracteres especiais;

3.4. Aplicar modelos nos dados.

Essa etapa tem como objetivo, ajustar os modelos em uma parte dos dados e avaliar o desempenho do modelo em outra parte dos dados [Pinheiro 2021]. A primeira parte dos dados é conhecida como conjunto de treinamento e a segunda parte como conjunto de validação. O modelo utilizado nesse trabalho foi previamente treinado com um dataset referente à pandemia causada pelo vírus SARS-COV-2. Desse modo o objetivo nessa etapa foi validar esse modelo em uma nova base de dados relacionada à *MonkeyPox*. As perguntas respondidas nessa etapa são:

1. Validar todos o modelo previamente treinado com base em um conjunto de dados diferente.
2. Avaliar os resultados do modelo e avalia-los com base nos objetivos das Perguntas Problemas definidas na primeira parte desse trabalho.

Nas próximas subseções serão discutidas as técnicas utilizadas nesse trabalho para análise e construção das respostas de cada Pergunta Problema.

3.4.1. Word Cloud

Uma nuvem de termos ou palavras (*word cloud* ou *weighted list*) é uma técnica utilizada normalmente para descrever as palavras chaves de documentos textuais como *websites* [Murthy and Scholar 2020]. Os termos descritos nessa forma de visualização são normalmente palavras, e a importância de cada termo é descrita de acordo com a fonte e a cor utilizada para descrever um termo. Essa escolha para visualização de dados é util para que o interessado na análise textual tenha uma rápida percepção sobre os termos mais proeminentes em uma coleção de documentos.

No contexto desse trabalho, a escolha do uso desse algorítmo se justifica pela necessidade de achar uma resposta para a **PP1**. Em outras palavras, a escolha desse algorítmo se justifica pela necessidade de visualizar os principais termos que compõem as postagens de tweets na língua portuguesa. Além disso, a implementação do Word Cloud escolhida para esse trabalho foi a feita utilizando a biblioteca feita por Andreas Mueller [Mueller 2016].

[2]https://www.nltk.org/

3.4.2. Análise de sentimentos

Análise de setimentos ou mineração de opiniões é o campo de estudos que tem o objetivo de realizar o estudo computacional das opiniões, atitudes e emoções de pessoas através de uma entidade. Uma entidade pode representar indivíduos, eventos ou tópicos [**?**]. Análise de sentimentos também pode ser considerado um processo de classificação, aonde há três níveis de análise de sentimentos: nível de documento, nível de sentença e nível de aspecto. No contexto desse trabalho a análise de sentimento realizada foi feita a nível de documento, os seja o objetivo é entender a polaridade de cada documento (*tweet*).

Do ponto de vista ferramental, foi utilizada uma técnica conhecida como *Support Vector Machine*. Basicamente, SVM são modelos de aprendizado supervisionado que estão associados com algorítmos que analizam dados para classifiação e análise de regressão. No caso desse trabalho o algorítmo utilizado foi a implementação da biblioteca Scikit learn *Support Vector Machine*[3]

Para a realização da Análise dos sentimentos, foi utilizado um modelo para classificação de sentenças de texto em 6 classes de sentimentos ('Neutro', 'triste', 'feliz', 'medo', 'raiva', 'nojo'). Assim como ilustrado na Figura 1, o modelo é composto por 2 módulos o módulo de vetorização de do documentos e o *Support Vector Machine Classifier*. O módulo de vetorização de documentos faz uso da biblioteca **SpaCy** para gerar vetores que são a representação de cada *tweet* em um espaço vetorial aonde o SVC será executado para achar a tarefa de classificação.

O módulo *Support Vector Machine Classifier* consiste de um classificador que utiliza o algorítmo *Support Vector Classifier*, que processará como entrada um vetor com 300 elementos que representam um elemento em um hiperplano. Por fim o classificador retorna a classe do texto de acordo com a posição do elemento.

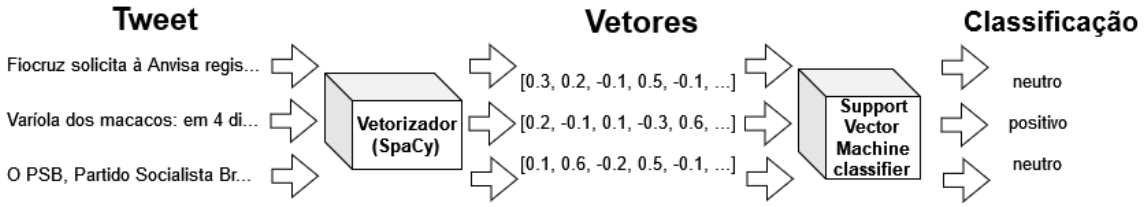

Figure 2. Modelo para classificação de sentimentos, adaptado de Pinto et al.

4. Experimento

Foi realizado um experimento para atestar a qualidade do modelo de classificação de sentimentos e obter também os termos mais relevantes contidos na coleção de tweets em português. O código fonte, utilizado na experimentação se encontra no repositório desse projeto no Github[4].As próximas subseções detalham, respectivamente, as configurações do experimento, e os resultados obtidos.

[3]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[4]github.com/

4.1. Configuração

Para a realização do experimento, um dataset com 6445 tweets entre os dias 11 de Agosto de 2022 até 25 de Setembro de 2022 foi construído através da API - o mesmo não será utilizado no escopo desse trabalho por razões legais impostas pelos termos de uso da *Twitter API for Academic Research*.

Os termos utilizados para a realização das buscas foram palavras relacionadas ao campo semântico da *MonkeyPox* na língua portuguesa (varíola dos macacos, variola dos macacos, #varioladosmacacos). Para a realização das requisições é condição necessária possuir uma conta de desenvolvedor no Twitter[5] com direitos de acesso sobre a *Academic Research API*.

Em relação ao ambiente utilizado para desenvolvimento, foi utilizado o Python versão 3.7 no *Google ColabResearch*[6]. As bibliotecas utilizadas foram NLTK, wordcloud, MATPLOTLib, SciKit-learn.

4.2. Resultados

4.2.1. Wordcloud

Com o objetivo de responder a **PP2**, foi utilizado o algorítimo para processamento e visualização de texto Wordcloud. Conforme exibido na figura 3, os principais termos obtidos foram as palavras como 'varíola macaco', 'adverso vacina', 'COVID chamada', 'casos confirmados', 'casal gay'. Através dessa nuvem de palavras é possível notar algumas características importantes na base de *tweets*. Dois exemplos são: a grande quantidade de postagens relacionadas ao número de casos crescente no Brasil durante a coleta dos tweets; indícios de uma preconceituosa crença reforçada por uma fala do presidente do Brasil em entrevista no dia 9 Agosto de 2022[Leite 2022] de que a doença fosse exclusiva à membros da comunidade LGBTQIA+.

Figure 3. Nuvem de palavras contendo os termos mais importantes na base de *tweets*

[5]https://developer.twitter.com
[6]https://colab.research.google.com/

4.2.2. Análise de Sentimentos

Conforme descrito no gráfico de barras da imagem 4, o modelo utilizado para Análise de Sentimentos classificou 3478 postagens como Neutra, 575 associadas ao sentimento de 'tristeza', 303 relacionadas a 'felicidade', 274 à 'medo', 1710 relacionadas à raiva e 105 publicações à 'nojo'. Já a figura 5 demonstra a mesma análise de sentimento feita em um recorte para tweets que contém as palavras 'casal Gay', através da mesma pode-se notar que os sentimentos mais recontes nesses *tweets* foram 'nojo' e 'raiva'. Essas características também são indícios da propagação de uma crença preconceituosa. Diante do exposto esse trabalho apresenta a resposta para a PP2.

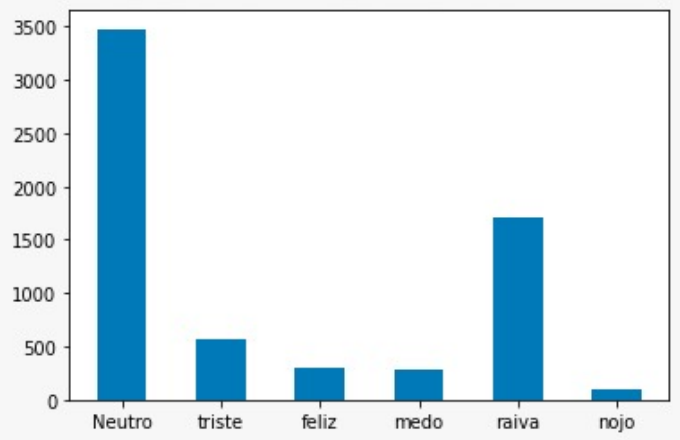

Figure 4. Gráfico de barras indicando a distribuição de Tweets pelos 6 sentimentos

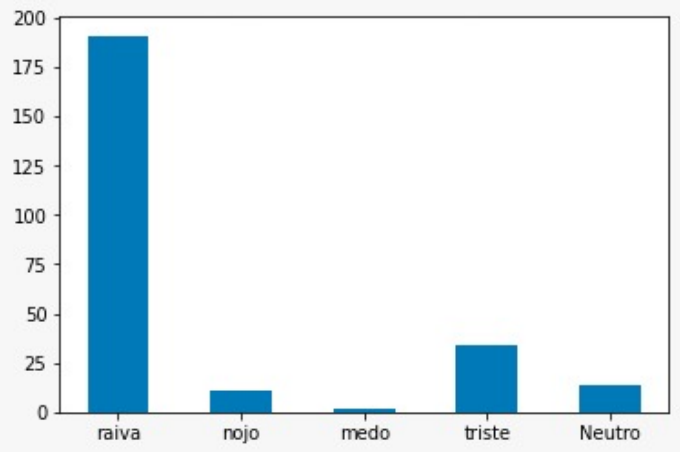

Figure 5. Gráfico de barras indicando a distribuição de tweets com os termos 'casal gay' distribuídos pelos 6 sentimentos

5. Considerações Finais e Trabalhos Futuros

Este artigo apresentou um estudo de caso sobre análises de sentimentos e termos importantes no contexto do surto de *MonkeyPox* para pessoas que possuem o português como idioma. Para alcançar esse objetivo uma busca por *tweets* em português sobre no contexto

da varíola dos macacos foi realizada dentro do período de tempo referente ao início do surto da varíola dos macacos em seres humanos. Referente a exploração do *dataset* construído, foram utilizadas técnicas de pré-processamento de texto para remoção de palavras e termos que não possuem relevância semântica. Esse pré-processamento foi condição necessária para a geração das entradas dos algorítmos de *Cloud of Words* e *Support Vector Classification*. Por fim, os modelos foram aplicados nos dados pré-processados com o objetivo de obter as percepções necessárias para responder todas as perguntas problema desse trabalho.

Após a aplicação do algorítmo do classificador de sentimentos e do algorítmo de texto para construção de nuvens de palavras, percebeu-se que os principais termos contidos na base de dados eram temos relacionados a postagens de cunho informativo como por exemplo *tweets* que informavam a quantidade de casos de varíola dos macacos em alguma região, seguidos por postagens que eram manifestação de raiva sobre algum aspecto envolvendo a *MonkeyPox*.Em conjunto com a análise de sentimentos, foi possível verificar que a grande quantidade de *tweets* informativos influenciaram a grande quantidade de *tweets* com sentimento neutro. Outro aspecto interessante do trabalho foi encontrar que alguns relações entre eventos do mundo real e palavras chaves e sentimentos na base de dados, por exemplo, a relação de uma entrevista do presidente do brasil, com os termos 'casal gay' e as classificações desses *tweets* nas classes de raiva e nojo.

Como trabalhos futuros, pode-se citar a utilização de outras técnicas de análise sentimentos a fim de buscar resultados melhores que um modelo treinado pré-treinado. Adicionalmente, pretende-se utilizar algorítmos de modelagem de tópicos em conjunto com técnicas de análise de série temporal, com o objetivo de identificar a relação direta entre eventos ocorridos na mídia e os tópicos que compõem a base. Com isso, também será possível entender como os sentimentos e opiniões dos usuários do Twitter sobre a *MonkeyPox* se comportam ao longo do tempo.

References

Leite, J. (2022). Bolsonaro insinua que quem se vacinar contra monkeypox é gay.

Mueller, A. (2016). GitHub - amueller/word_cloud: A little word cloud generator in Python.

Murthy, K. N. and Scholar, P. (2020). WORD CLOUD IN PYTHON. 24(01):7.

Ortiz-Martínez, Y., Sarmiento, J., Bonilla-Aldana, D. K., and Rodríguez-Morales, A. J. (2022). Monkeypox goes viral: measuring the misinformation outbreak on Twitter. *The Journal of Infection in Developing Countries*, 16(07):1218–1220.

Pinheiro, M. P. C. R. (2021). *Introduction to Statistical and Machine Learning Methods for Data Science*.

Pinto, M. A. S., Jacob Junior, A. F. L., Busson, A. J. G., and Colcher, S. (2020). Relacionando Modelagem de Tópicos e Classificação de Sentimentos para Análise de Mensagens do Twitter Durante a Pandemia da COVID-19. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2020)*, pages 61–64, Brasil. Sociedade Brasileira de Computação - SBC.

Thakur, N. (2022). MonkeyPox2022Tweets: The First Public Twitter Dataset on the 2022 MonkeyPox Outbreak. preprint, MATHEMATICS & COMPUTER SCIENCE.

WHO. Monkeypox.