Um modelo de análise de sentimentos sobre o impacto das notícias no Twitter: um estudo de caso sobre a Covid-19

Suzane J. Menon, Isabel H. Manssour
Curso de Especialização em Ciência de Dados
Pontifícia Universidade Católica do Rio Grande do Sul - Porto Alegre, Brasil
suzanemenon@gmail.com, isabel.manssour@pucrs.br

***Resumo*—O constante aumento do volume de dados que são produzidos e compartilhados todos os dias na internet torna indispensável o uso de ferramentas por jornalistas e criadores de conteúdo para agilizar o processo de investigação. Uma etapa importante neste processo é entender o sentimento das pessoas sobre um determinado assunto. Neste trabalho é apresentado um modelo de análise de sentimentos para classificar postagens do Twitter. Um estudo de caso aplicado ao tema COVID-19 no Brasil em 2020, resultou no desenvolvimento de uma ferramenta que utiliza técnicas de visualização e filtros, propondo relações entre o impacto das notícias no sentimento das pessoas.**

***Palavras-chave*—Análise de sentimentos, máquina de aprendizado, classificação, visualização, Twitter, mídia social, covid-19.**

I. Introdução

O constante aumento do volume de dados gerados e a velocidade com que trafegam ou são disponibilizados nos diversos meios na internet é uma realidade nos dias atuais. Por isso, a internet vem sendo adotada como principal meio de obtenção e compartilhamento de dados e informações [8]. No entanto, devido a essa velocidade, a informação pode facilmente se tornar obsoleta. Portais de notícias de grande repercussão como G1 e Folha de São Paulo, permitem aos leitores estarem bem informados em tempo real e de forma imparcial na maioria das vezes. A internet é, também, o local no qual as pessoas sentem-se livres para se expressarem de forma espontânea. Canais como Twitter e Instagram podem ser utilizados para representar uma amostra da população e seus sentimentos sobre determinando assunto.

Jornalistas e criadores de conteúdo conduzem suas pesquisas e investigações com base em fatos e acontecimentos do mundo. Uma das maiores dificuldades apontadas por Oliveira [9] não é a falta da informação, mas sim o excesso dela. O uso de ferramentas baseadas em técnicas de ciência de dados tornaram-se fundamentais para execução de tarefas como coleta, filtro e análise de dados que normalmente exigem muito tempo do profissional. A busca por cientistas de dados para resolver este problema tem aumentado muito [10]. Estes profissionais atuam como peça chave dentro das organizações filtrando e atribuindo significado aos dados.

Neste trabalho foi desenvolvido um modelo de análise de sentimentos que utiliza mensagens extraídas do Twitter para gerar uma amostra do sentimento dos usuários em relação a uma pesquisa. Também são exibidas notícias postadas no intervalo da observação, contribuindo, assim, para a análise do impacto dessas notícias na opinião das pessoas.

Em uma amostra de, aproximadamente, 60 mil *tweets* em português sobre a COVID-19 postados entre 23 de Fevereiro a 2 de Agosto de 2020, o modelo revelou um aumento crescente de mensagens negativas sobre a doença. Foi identificado, também, que uma das *hashtags* mais comentadas durante este período, além das derivações de COVID-19, foi ”fiqueem-casa”. As notícias, por sua vez, mostram o aumento de casos da doença em diversas regiões do Brasil. Utilizando estes dados como um complemento a pesquisa, jornalistas podem, por exemplo, conduzir a escrita de uma matéria apontando que as pessoas, de um modo geral, estão conscientes sobre as medidas de segurança para evitar a transmissão da doença. No entanto, predomina uma sensação de insatisfação sobre o cenário da pandemia no Brasil.

O restante deste documento está organizado da seguinte forma: A Seção II apresenta uma revisão de alguns trabalhos relacionados. Uma descrição do modelo proposto, incluindo a metodologia utilizada e a arquitetura desenvolvida é apresentada na Seção III. Na Seção IV é apresentado um estudo de caso abordando o cenário da COVID-19 no Brasil em 2020. Por fim, são apresentadas as conclusões, incluindo as dificuldades encontradas durante o desenvolvimento do trabalho e suas possíveis melhorias.

II. Trabalhos Relacionados

Inúmeros trabalhos vêm sendo desenvolvidos na área de análise de sentimentos dos usuários de redes sociais. Estes trabalhos utilizam diferentes técnicas e se caracterizam por buscar melhores formas de extrair e apresentar informações de grandes volumes de dados.

Queiroz et al. [1] apresenta um estudo sobre formas de visualização interativas utilizando dados coletados do Twitter e a opinião de profissionais da área do jornalismo em relação a estas visualizações. Foram utilizadas técnicas de coleta, pré-processamento e treino de um modelo MultinomialNB para classificação de *tweets* que viabilizou criar uma relação com notícias. Os dados para treino do modelo foram classificados a partir da rotulação manual de 30 *hashtags* em positiva ou negativa.

Uma comparação entre dois métodos de classificação de sentimentos sobre temas relacionados a saúde é apresentada por Araujo [2]. O primeiro método implementa um modelo de análise de sentimentos baseado no ANEW-BR, um vocabulário afetivo traduzido e adaptado do *Affective Norms for English Word (ANEW)* para o português brasileiro. Também foram aplicadas abordagens de classificação por *emoticons* e identificação de negação em frases. No segundo, 750 voluntários classificaram manualmente 3.119 mensagens. Os resultados finais mostram que o classificador desenvolvido obteve melhores resultados comparado a classificação feita por humanos. No entanto, a autora conclui que, tratando-se de um tema complexo como saúde, as mensagens foram difíceis de serem rotuladas até mesmo pelos humanos.

Malheiros [3] propõe uma ferramenta simples e intuitiva que captura em tempo real a opinião dos usuários a partir de palavras-chave no Twitter e mostra graficamente um comparativo entre estas palavras. Outra funcionalidade interessante é a possibilidade de acompanhar os comentários coletados na própria ferramenta. Em termos de modelagem, este trabalho utilizou a base de dados Sentiment140 que é uma base rotulada de dados em inglês do Twitter.

Os trabalhos abordados apresentam 3 características importantes que vão ao encontro ao que foi desenvolvido neste artigo: relação entre notícias e sentimento dos usuários, rotulação manual de dados como entrada para um modelo de aprendizado e por fim o desenvolvimento de uma aplicação que utiliza formas de visualização interativas.

III. Descrição do projeto

A seguir são apresentadas as etapas de desenvolvimento deste trabalho, visitando em detalhes cada módulo que compõe o modelo.

A. Metodologia

Foi desenvolvido um modelo composto por três módulos principais: coleta de notícias e mensagens de redes sociais, treino de um modelo de análise de sentimentos e visualização de resultados. A Figura 1 ilustra a conexão entre os módulos.

No primeiro módulo, os dados coletados foram agrupados em dois núcleos distintos de informação: notícias e *tweets*. As notícias são caracterizadas por fatos e/ou eventos sobre um determinado assunto, publicados através de meios de comunicação confiáveis, como jornais e portais de notícias. Inicialmente, foi escolhida como fonte de notícia o site do G1, que publica diariamente conteúdo sobre os principais acontecimentos do mundo. Para coleta de opiniões, foi escolhido o Twitter, ambiente popularmente conhecido por permitir que os usuários se expressem de forma livre através de *tweets*, mensagens de até 280 caracteres.

Para a coleta de notícias e *tweets* é necessário estabelecer uma palavra-chave e um período de tempo. Isto faz com que os dados estejam associados a um mesmo contexto. Utilizando ferramentas de extração automatizada de conteúdo da internet, foi obtida uma base de dados extensa e rica em informações que serão úteis em diversas etapas deste trabalho. A coleta de

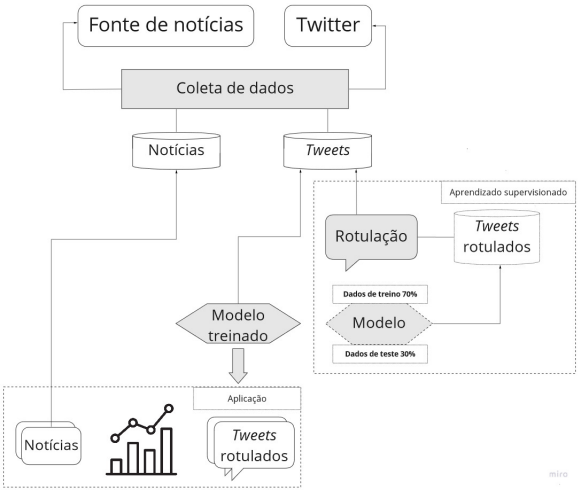

Fig. 1. Definição dos módulos que compõem o modelo.

tweets precede a de notícias, pois o modelo irá utilizar o texto dos *tweets* para rotulação e treinamento. As notícias podem ser coletadas a qualquer momento e quantas vezes for necessário.

Por *tweets* se tratarem de mensagens de conteúdo textual, no segundo módulo foi definido um modelo de aprendizado supervisionado utilizando Naive Bayes Multinomial. De acordo com [5], este algoritmo de distribuição multinomial é uma das duas variações mais utilizadas para classificação de textos por fazer uso de vetores que agrupam contadores de menção a palavras, ou também chamado de *Bag-of-words*.

Para possibilitar a análise de sentimentos destas mensagens, a técnica escolhida utiliza dados rotulados como entrada. Como muitas dessas mensagens são caracterizadas por possuírem ironia em sua composição, foi necessária uma avaliação humana para a rotulá-las. Um modelo que tornasse o processo de classificação automatizado se tornou necessário. Desta forma, novas mensagens puderam ser classificadas instantaneamente sem ajuda externa.

Cada *tweet* foi submetido a um pré-processamento para remoção de termos e caracteres que não trazem significado para o modelo de aprendizado. No item III-A estas etapas são descritas em maiores detalhes.

A nova base de dados foi dividida em dados de treino (70%) e teste (30%). Com os dados de treino o algoritmo criou associações entre as palavras e os rótulos que foram distribuídos anteriormente. Já com os dados de teste foi feito o cálculo do quão preciso é o modelo para acertar novas entradas, ou seja, novos comentários.

Por fim, o terceiro módulo consiste em um conjunto de técnicas de visualização que possibilitam uma análise visual das notícias e opiniões coletadas no módulo 1 e processadas no módulo 2.

B. Arquitetura e Implementação

A coleta de dados na internet ou *web scraping*, é frequentemente utilizada em ciência de dados para formação de grandes *datasets* [4]. Esta prática faz uso de uma série de

técnicas e ferramentas que podem ser encontradas facilmente na comunidade *open source*, mas também existem soluções poderosas dada a importância desta prática para o mercado nos dias atuais. O processo de *scraping* desenvolvido neste trabalho permitiu criar uma base de dados tanto para notícias quanto para *tweets*. No entanto, foram aplicadas soluções diferentes para cada caso.

A linguagem Python dispõe do módulo *urllib* em sua estrutura para leitura de *URLs*, o que viabilizou o acesso ao conteúdo das páginas de notícias. A extração das partes relevantes deste conteúdo foi resultado de um conjunto de seletores baseados em expressões regulares que serviram de entrada para a biblioteca BeautifulSoup. Já para os *tweets*, foi necessária uma ferramenta que buscasse por publicações antigas. A *API* do Twitter em sua versão pública limita o acesso à comentários anteriores a 7 dias [11]. Sendo assim, buscou-se alternativas na comunidade *open source* e dentre as opções possíveis a biblioteca GetOldTweets3[1] mostrou-se suficiente para esta tarefa.

A partir deste momento, apenas *tweets* são considerados para o pré-processamento. As notícias serão utilizadas posteriormente no módulo de visualização. Cada *tweet* foi submetido às seguintes técnicas:

- *Filtering* - Remoção de qualquer conjunto de palavras que não possuem um significado para o contexto do modelo de aprendizado. Para este trabalho são removidos *hyperlinks*, *URLs hashtags*, menções a outros usuários da rede social e caracteres especiais.
- *Lowercasing* - Técnica de padronização utilizada para converter caracteres em formato caixa alta para caixa baixa, evitando inconsistências.
- *Tokenize* - Quebra de textos em palavras (*tokens*).
- *Stopwords* - Remoção de palavras que não acrescentam significado em uma sentença e que podem ser ignoradas sem afetar o modelo [7]. Artigos, pronomes, preposições e alguns advérbios são exemplos de palavras da língua portuguesa que são removidas utilizando a biblioteca NLTK[2]. Esta é uma das principais bibliotecas Python para construção de programas que trabalham com dados de linguagem humana.

Com isto, foi possível construir uma matriz esparsa compatível para o uso da função CountVectorizer [6], que tem como entrada um conjunto de vetores de palavras livres de ruídos.

O modelo de análise de sentimentos foi construído com base na *scikit-learn*[3] que é uma biblioteca *open source* de aprendizado de máquina. Dessa forma, teve-se acesso a funções de separação de dados de treino e teste, extração de *features* com CountVectorizer, aplicação do treinamento por Naive Bayes Multinomial e também apresentar as métricas referentes ao modelo treinado.

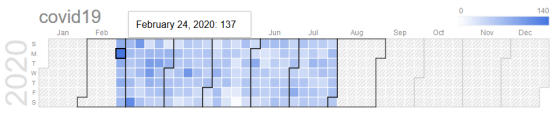

Fig. 2. Visualização de menções ao termo COVID-19.

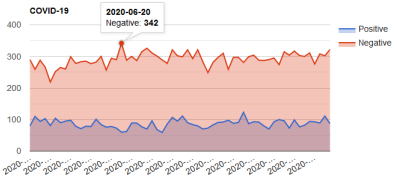

Fig. 3. Evolução de *tweets* classificados como positivo e negativo

Os módulos de extração e manipulação dos dados bem como as técnicas de aprendizado de máquina foram desenvolvidos na linguagem Python no ambiente Google Colab. A representação visual das análises realizadas foram desenvolvidas em uma aplicação Javascript com o uso de bibliotecas gráficas como React Google Charts[4] e React Word Could[5]. A implementação está disponível de forma aberta no GitHub[6].

C. Visualização

Foi desenvolvida uma visualização interativa utilizando as notícias coletadas e a base de dados de *tweets* classificadas pelo modelo. As estruturas escolhidas auxiliam na elaboração de hipóteses através da aplicação de filtros por período de coleta e por *hashtag*. Cada estrutura é detalhada a seguir:

- Gráfico de calor - Este gráfico foi implementado com base no conceito de *heatmap*. Cada ponto da matriz contém o número total de *tweets* que fizeram menção a uma *hashtag*. Com base em cada valor, o gráfico define a intensidade da cor e aplica ao longo de toda a matriz. Esta forma de visualização traz um panorama anual dos dias em que uma *hashtag* foi mencionada e sua frequência como mostra a Figura 2.
- Gráfico de área - Para representar a evolução dos sentimentos positivos e negativos do usuário em um período de tempo, o gráfico em área atende muito bem a este requisito. Uma aplicação deste gráfico foi utilizada para representar os dois sentimentos, sendo a cor vermelha utilizada para o sentimento negativo e azul para o positivo. O eixo x contém os dias do intervalo e o eixo y a quantidade de *tweets* classificados como positivo ou negativo. A Figura 3 apresenta um exemplo desse gráfico relativo a uma amostra capturada entre 2 de Junho a 31 de Julho de 2020.
- Nuvem de palavras - Esta visualização tem se tornado bastante comum na internet, principalmente em redes sociais, para identificar rapidamente os assuntos que estão em evidência no mundo. A nuvem de palavras consiste

[1]https://pypi.org/project/GetOldTweets3/
[2]https://www.nltk.org/
[3]https://scikit-learn.org/
[4]https://react-google-charts.com/
[5]https://react-wordcloud.netlify.app/
[6]https://github.com/suzanemenon/sentiment-analysis/

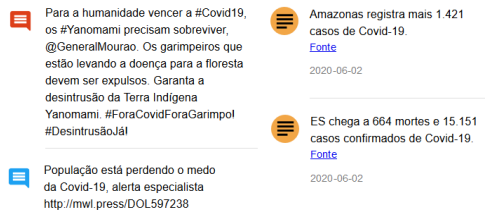

Fig. 4. Representação de *tweets* e notícias

em realçar as palavras de acordo com a sua frequência. Quanto maior o tamanho da palavra no gráfico, maior é a sua ocorrência em relação as outras. Neste trabalho, as *hashtags* obtidas foram utilizadas como entrada para a nuvem de palavras.

- Listas - A visualização em área detalhada anteriormente permite, também, selecionar um ponto específico no gráfico para mostrar uma lista com os *tweets* postados naquele dia. Ao selecionar um ponto na área azul, são mostradas apenas as postagens classificadas como positivas. O mesmo ocorre para postagens negativas ao selecionar a área vermelha. A cor do símbolo ao lado do texto caracteriza a distinção entre os sentimentos de cada *tweet*. As notícias, por sua vez, são listadas de acordo com o período de tempo definido, sem influência dos demais gráficos. Cada item é formado pelo título original da notícia e sua fonte. O conjunto de notícias e *tweets* é listado a partir do período selecionado. Um exemplo de notícias e *tweets* pode ser visto na Figura 4.

IV. Estudo de Caso

No final do ano de 2019, o mundo se deparou com um novo tipo de coronavirus, a SARS-CoV-2. Denominada como COVID-19, esta doença apresenta um fator de transmissão preocupante na sociedade [13]. De acordo com [14], o Brasil ultrapassou a marca de 100 mil mortes por COVID-19, ficando apenas atrás dos Estados Unidos. O uso de técnicas de aprendizado de máquina e visualização possibilitaram identificar o efeito das notícias sobre a COVID-19 na mudança de sentimento dos usuários da rede social ao longo do tempo.

Notícias e *tweets* foram extraídos no idioma português do Brasil a partir da palavra-chave COVID-19. Foram coletadas aproximadamente 2400 notícias e 60 mil *tweets* publicados entre 23 de Fevereiro a 2 de Agosto de 2020. Foram escolhidos aleatoriamente 500 *tweets* que passaram por um processo de rotulação manual realizada por 3 voluntários. Este método foi utilizado como meio para identificar e interpretar de forma mais precisa a ocorrência de ironia e sarcasmo, cada vez mais utilizados em redes sociais.

Cada um recebeu um mesmo conjunto de *tweets* e foram informados previamente sobre o contexto do estudo. Em seguida, foi solicitado que, para cada mensagem, fosse feita uma interpretação isenta de viés ideológico se o autor estava se referindo ao contexto COVID-19 de forma positiva, negativa ou neutra. Dos 3 questionários respondidos, 2 foram utilizados por completo e o terceiro como critério de desempate. O resultado dos questionários mostra que 55.2% das mensagens foram rotuladas como negativo, 30.4% positivo e 14.4% neutros.

Para o treinamento do modelo, uma amostra de 70% dos *tweets* rotulados foi utilizado. Com os 30% restantes foi realizado o teste do modelo, revelando uma acurácia de 59% no resultado. Estudos feitos em trabalhos relacionados mostram que este valor tende a aumentar a medida que o número de dados de treino aumenta.

De forma a possibilitar a discussão e análise sobre a influência das notícias no sentimento dos brasileiros ao longo do período deste estudo, foi implementada uma visualização contendo elementos compostos por gráficos e filtros. O filtro por *hashtags* mais relevantes mostrou que o termo ”yanomami” foi mencionado 128 vezes. Ao selecionar este termo, o gráfico de calor foi atualizado revelando que entre os dias 2 de Junho a 25 de Julho eventos em relação a esse termo estavam ocorrendo no contexto da COVID-19. Ao definir este período como filtro para notícias e *tweets*, o gráfico de área mostrou um grande número de postagens de sentimento negativo em relação ao tema. Ao final, com a ajuda das notícias listadas, foi possível concluir que se trata de um povo indígena que vive na floresta amazônica e que está sofrendo com a invasão de garimpeiros, aumentando a transmissão do vírus na região.

V. Conclusões e Trabalhos Futuros

Técnicas abordadas em ciência de dados são fortes aliadas nos processos de pesquisa, investigação e análise executados por profissionais do jornalismo. Com o aumento no volume de dados que são gerados e compartilhados todos os dias, torna-se indispensável o uso de ferramentas que agilizem o processo de coleta, transformação e análise destes dados. Outro fator importante neste processo é, também, entender o sentimento do público sobre um determinado assunto.

O objetivo deste trabalho foi implementar um modelo de análise de sentimentos capaz de classificar postagens feitas na rede social Twitter como positivas, negativas e neutras. Também foi desenvolvida uma aplicação que utilizou técnicas de visualização e filtros para criar relações entre o sentimento das pessoas com as notícias publicadas no mesmo período.

A partir de um estudo de caso envolvendo dados de notícias e *tweets* sobre a COVID-19 em 2020, foi identificado o caso de invasão de garimpeiros em terras indígenas do povo Yanomami e também a preocupação das pessoas a respeito do assunto. Fazendo uso deste modelo em outros contextos permite, então, que o criador de conteúdo tenha em suas mãos uma ferramenta simples e intuitiva para descobrir os eventos que cercam a sua pesquisa.

Durante o desenvolvimento deste trabalho foram identificadas diversas oportunidades de melhoria. Dentre elas, aumentar o número de fontes de coleta de dados, utilizar técnicas combinadas de aprendizado de máquina para refinar o modelo, incluir novos gráficos na aplicação e torná-la disponível online em formato colaborativo, aumentando o uso da ferramenta entre jornalistas e produtores de conteúdo.

REFERÊNCIAS

[1] Queiroz Santos, C., Cunha, H., Teixeira, C., Ramos de Souza, D., Tietzmann, R., Manssour, I., ... Barros, R, "Media professionals' opinions about interactive visualizations of political polarization during Brazilian presidential campaigns on Twitter", 2017 50th Hawaii International Conference on System Sciences.

[2] ARAUJO, Gabriela Denise. Análise de sentimento de mensagens do Twitter em português brasileiro relacionadas a temas de saúde. 2014. 84 f. Dissertação (Mestrado) – Escola Paulista de Medicina, Universidade Federal de São Paulo. São Paulo, 2014.

[3] MALHEIROS, Yuri. Emotte: Uma Ferramenta De Análise de Sentimentos para o Twitter. In: WORKSHOP DE FERRAMENTAS E APLICAÇÕES - SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB (WEBMEDIA) , 2014, João Pessoa. Anais Estendidos do XX Simpósio Brasileiro de Sistemas Multimídia e Web. Porto Alegre: Sociedade Brasileira de Computação, nov. 2014 . p. 62-65. ISSN 2596-1683.

[4] ScrapingHub. "What is web scraping?" scrapinghub.com https://www.scrapinghub.com/what-is-web-scraping/(acessado em 17 de Agosto de 2020).

[5] scikit-learn Machine Learning in Python. "Multinomial Naive Bayes" scikit-learn.org/ https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes(acessado em 17 de Agosto de 2020).

[6] scikit-learn Machine Learning in Python. "CountVectorizer" scikit-learn.org/ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (acessado em 28 de Agosto de 2020)

[7] NLTK 3.5 documentation. "Examples for Portuguese Processing" nltk.org http://www.nltk.org/howto/portuguese_en.html (acessado em 17 de Agosto de 2020)

[8] SCHONS, Claudio H.. O volume de informações na Internet e sua desorganização. Inf. Inf., Londrina, v. 12, n. 1,jan./jun. 2007

[9] OLIVEIRA, Ângela M; NOVAIS, Eunice S; SILVA, Ivani da. Sistema de informação de marketing em unidades de informação. Revista Biblios. Ano 5, n. 18- 19, abr./set. 2004.

[10] Data Science Academy. "O Impacto da Pandemia nas Carreiras em Data Science" datascienceacademy.com.br http://datascienceacademy.com.br/blog/o-impacto-da-pandemia-nas-carreiras-em-data-science/ (acessado em 28 de Agosto de 2020)

[11] Twitter Developers Documentation. "Standard search" developer.twitter.com https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/standard (acessado em 28 de Agosto de 2020)

[12] Veja Saúde. "Coronavírus: "O Brasil transformou a crise sanitária em crise política" saude.abril.com.br https://saude.abril.com.br/medicina/coronavirus-o-brasil-transformou-a-crise-sanitaria-em-crise-politica/ (acessado em 28 de Agosto de 2020)

[13] Ministério da Saúde. "COVID-19" coronavirus.saude.gov.br https://coronavirus.saude.gov.br/sobre-a-doencao-que-e-covid (acessado em 28 de Agosto de 2020)

[14] BBC Brasil. "Brasil passa dos 118 mil mortos e 3,7 milhões de infectados por covid-19" www.bbc.com https://www.bbc.com/portuguese/brasil-51713943 (acessado em 28 de Agosto de 2020)