A COVID-19 no Twitter: correlacionando vocabulário com agravamento e atenuação da pandemia no Brasil

Pedro Loures Alzamora[1], Marcelo Sartori Locatelli[1], Marcelo Ganem[1], Thiago Henrique Moreira Santos[1], Daniel Victor Ferreira[1], Tereza Bernardes[1], Ramon A. S. Franco[1,2], Janaína Guiginski[1], Evandro L. T. P. Cunha[1], Ana Paula Couto da Silva[1], Wagner Meira Jr.[1]

[1]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil

[2]Universidade Federal do Oeste da Bahia (UFOB), Barreiras, BA, Brasil

{ploures.alzamora,janainaguiginski,tbernardesfaria}@gmail.com

{evandrocunha,ana.coutosilva,meira}@dcc.ufmg.br

ramon.franco@ufob.edu.br

Abstract. *This study characterizes the first year of the COVID-19 pandemic in Brazil as a social phenomenon by analyzing the correlation between the aggravation/attenuation of the pandemic and the vocabulary used on Twitter in the weeks that precede these variations. Among other results, we observed that politically motivated terms and words with a negative tone are more prevalent in the weeks that precede the increase in the number of cases/deaths, while the use of terms related to media content (internet, music, television) is intensified in the weeks preceding the drop in the number of cases/deaths. Such results suggest the possibility of using the method introduced here for the analysis of social phenomena using computationally light and totally anonymized data from online social networks.*

Resumo. *O presente estudo busca caracterizar o primeiro ano da pandemia de COVID-19 no Brasil como um fenômeno social por meio da análise da correlação entre o agravamento/atenuação da pandemia e o vocabulário utilizado no Twitter nas semanas que precedem essas variações. Entre outros resultados, observou-se que termos politicamente motivados e com teor negativo são mais prevalentes nas semanas que precedem o aumento do número de casos/mortes, ao passo que o uso de termos relacionados a conteúdos midiáticos (internet, música, televisão) é intensificado nas semanas que antecedem a queda da quantidade de casos/mortes. Tais resultados sugerem a possibilidade de utilização do método aqui introduzido para a análise de fenômenos sociais a partir de dados computacionalmente leves e totalmente anonimizados provenientes de redes sociais online.*

1. Introdução

Com o rápido avanço da COVID-19 nas diversas regiões brasileiras e seu impacto em diferentes esferas da sociedade (economia, política, cultura etc.), tornou-se relevante o desenvolvimento de métodos que auxiliassem na sua compreensão e na previsão da sua propagação no território. A análise de redes sociais online como ferramenta para a

caracterização de fenômenos relacionados à saúde pública vem sendo uma prática amplamente adotada na ciência [França et al. 2014]. O monitoramento dessas redes permite acompanhar, prever e avaliar a repercussão de fenômenos em tempo real, capturando o comportamento de um conjunto de indivíduos em relação a determinado evento durante certo período de tempo. Neste contexto, o Twitter é uma rede comumente utilizada [Gomide et al. 2011, Aiello et al. 2020, Li et al. 2020, Malagoli et al. 2021], uma vez que sua política de coleta de dados permite obter grandes quantidades de postagens, filtradas por palavras-chave relacionadas ao tema de interesse. Adicionalmente, o Twitter é uma rede social online amplamente usada no Brasil e no mundo, além de possuir caráter extremamente reativo e espontâneo, o que favorece estudos em tempo real sobre questões sanitárias e epidemiológicas [Sultana et al. 2021].

Neste trabalho, é utilizado um conjunto de dados obtidos a partir do Twitter para buscar responder a duas questões de pesquisa: (i) é possível caracterizar os temas debatidos na rede apenas a partir da frequência de palavras, isto é, sem ter acesso ao texto completo dos *tweets*?; e (ii) de que forma o tom e a abordagem da COVID-19 no Twitter se modifica nas semanas que precedem a variação dos números de casos/mortes pela doença? Os resultados aqui alcançados indicam que correlações positivas com o agravamento da pandemia evocam temas e abordagens diferentes daqueles associados às correlações negativas, sugerindo que a resposta à primeira questão de pesquisa é positiva. Ademais, também revelam que a semana anterior ao aumento dos casos apresenta maior proporção de termos políticos com forte carga crítica, sugerindo que a segunda questão de pesquisa também possui resposta positiva. Ressaltamos também que essas análises não sugerem causalidade entre os temas abordados no Twitter e o número de casos/mortes, apenas correlação.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta a metodologia introduzida para a realização das análises aqui apresentadas, as quais utilizam um conjunto de dados computacionalmente leves e totalmente anonimizados provenientes do Twitter. A Seção 3 evidencia os resultados alcançados, revelando as palavras cuja variação no uso se correlaciona às variações nos números de casos/mortes por COVID-19 no primeiro ano da pandemia no Brasil. Por fim, a Seção 4 conclui o artigo e apresenta propostas para trabalhos futuros.

2. Metodologia

A metodologia deste trabalho consiste em três etapas principais: (i) coleta e tratamento dos dados; (ii) cruzamento dos dados e seleção das palavras; e (iii) análise das palavras selecionadas. Essas três etapas são detalhadas a seguir.

2.1. Coleta e Tratamento dos Dados

Dados provenientes do Twitter. Para a realização das análises aqui apresentadas, foram considerados quase 66 milhões de *tweets*, publicados entre 03 de maio de 2020 e 02 de janeiro de 2021, totalizando 35 semanas. A opção por restringir a coleta a *tweets* e não incluir *retweets* se justifica por evitar que textos de usuários mais influentes (por exemplo, com muitos seguidores) sejam mais expressivos na base de dados. O procedimento completo para a construção desse *dataset* é descrito por [Brum et al. 2020][1] e se restrin-

[1] O *dataset* apresentado por [Brum et al. 2020] tem duração de abril a julho de 2020. Neste trabalho, é utilizada uma versão expandida desse *dataset*, coletada empregando os mesmos critérios e infraestrutura.

giu a *tweets* em português que mencionassem ao menos uma destas treze palavras-chave: *covid-19*, *covid19*, *covid*, *coronavirus*, *corona*, *sars*, *confinamento*, *quarentena*, *distanciamento social*, *aglomeração*, *aglomerações*, *cloroquina* e *hidroxicloroquina*. A partir dos *tweets* coletados, foi construída uma base de dados contendo a frequência semanal de todas as palavras utilizadas no período (isto é, não apenas das palavras-chave). Essa base de dados, portanto, não contém o texto dos *tweets* em si, pois se trata de um conjunto de dados anonimizados que conta apenas com a frequência de cada palavra em cada semana.

O tratamento desses dados consistiu na padronização da capitalização e aglutinação das frequências de palavras previamente distintas em função de padrões de capitalização diferentes, que foi então seguida por um processo de filtragem, em que foram removidos todos os caracteres numéricos, além de todas as palavras com três ou menos caracteres (que, em português, frequentemente correspondem a artigos, preposições, interjeições e pronomes). É importante ressaltar que esse pré-processamento pode ter descartado algumas siglas e abreviações com função semântica relevante para a construção de sentido dos *tweets* e que não foi realizado processo de lematização – o qual se pretende implementar em trabalhos futuros.

Com base na frequência absoluta das palavras por semana, foi calculada a proporção (ou frequência relativa) de cada uma delas no total de palavras da respectiva semana. Todas as palavras, bem como essas proporções semanais, foram agrupadas em uma única tabela, a partir da qual foi calculado o fator de crescimento das palavras ao longo do tempo. Esse processo foi feito por meio da divisão da proporção de cada palavra em uma semana pela sua proporção na semana anterior. Os fatores de crescimento (positivos ou negativos) fornecem uma medida instantânea de tendência (de crescimento ou de decrescimento) no emprego das palavras nos *tweets*, possibilitando a comparação direta entre essa medida e o fator de crescimento de casos e mortes por COVID-19.

Vale ressaltar que, como a base de dados não contém o texto dos *tweets* em si, a privacidade dos dados compartilhados pelos usuários do Twitter é garantida, uma vez que foi armazenado apenas o número de vezes em que cada palavra aparece nos *tweets*. Com os dados totalmente anônimos, foi possível sua publicação em repositórios de acesso público [Moreira et al. 2021]. Além disso, o conjunto de dados organizado dessa forma pode ser mais facilmente processado, pois exige menos capacidade de armazenamento e computação. Em contrapartida, esse tipo de tratamento inviabiliza a depreensão do contexto em que cada palavra foi utilizada.

Indicadores epidemiológicos. As informações sobre indicadores epidemiológicos da pandemia de COVID-19 no Brasil foram obtidas no âmbito do projeto de pesquisa Covid Data Analytics, conduzido pelo Departamento de Ciência da Computação da Universidade Federal de Minas Gerais[2]. O cálculo dos indicadores utiliza informações de dois *datasets*: (i) uma base com dados de casos e óbitos compilados diariamente dos boletins epidemiológicos das 27 Secretarias Estaduais de Saúde[3]; e (ii) uma base de dados com a projeção bayesiana da população de todos os municípios do Brasil em 2020 [Freire et al. 2019].

[2]Mais informações em: `http://covid.dcc.ufmg.br/`. Acesso em: 20 maio 2022.
[3]Disponível em: `https://brasil.io/covid19/`. Acesso em: 03 fev. 2022.

Dentre os indicadores epidemiológicos e níveis geográficos disponíveis, optou-se por utilizar os fatores de crescimento de *casos* de COVID-19 e de *mortes* pela doença em todo o Brasil. A escolha desses dois indicadores se deve ao fato de que ambos fornecem uma medida do comportamento imediato da doença, indicando se naquele momento a pandemia estava se agravando ou atenuando. Além disso, o fator de crescimento é um indicador relativo, que normaliza e permite a comparação entre a frequência das palavras e o número de casos/mortes em uma mesma escala. Optou-se por trabalhar apenas com o nível geográfico nacional, uma vez que a base de dados de *tweets* utilizada enumera a contagem de palavras em todo o Brasil, não sendo possível distinguir esses dados conforme sua geolocalização mais precisa.

2.2. Cruzamento dos Dados e Seleção das Palavras

Para responder às perguntas de pesquisa apresentadas (Seção 1), foi calculado o coeficiente de correlação de Spearman [Spearman 1904] entre o fator de crescimento das palavras em uma determinada semana e o de casos/mortes por COVID-19 n semanas depois. Esse deslocamento de n semanas é chamado de *lag* e se comporta como uma janela deslizante ao longo do período estudado, variando entre 0 e 3. A escolha desses valores considera o tempo médio entre o período de incubação do vírus, o aparecimento dos sintomas e o possível agravamento dos casos [Verity et al. 2020, Sousa et al. 2020, Linton et al. 2020]. Em linhas gerais, esta análise revela quais palavras ganham ou perdem relevância em diferentes períodos que antecedem ou coocorrem com a alteração do número de casos/mortes, possibilitando interpretações que considerem características inerentes à doença.

As análises de correlação foram realizadas em três diferentes intervalos nas 35 semanas consideradas: (i) maio a agosto de 2020 (período 1); (ii) agosto de 2020 a janeiro de 2021 (período 2); e (iii) maio de 2020 a janeiro de 2021 (período total). A Figura 1 apresenta a motivação da divisão do período total de análise em diferentes intervalos. Pode-se notar que os dados coletados seguem um padrão de queda no número total de *tweets* e *retweets* até setembro de 2020, se estabilizando logo em seguida. Essa queda no total de dados coletados pode ter sido ocasionada por pelo menos dois motivos: o "esfriamento" do debate sobre o tema ao longo do tempo, comportamento característico encontrado no Twitter com relação a eventos que são debatidos na plataforma [Weng and Lee 2011, Bae et al. 2014]; ou a própria metodologia de coleta de [Brum et al. 2020], que manteve as mesmas palavras-chave durante todo o período de observação. Independentemente dos fatores que levaram a essa tendência, existe um padrão atípico na quantidade de *tweets* e *retweets* na primeira fase, o que resulta em uma correlação média de apenas 0,06 no período 1 – período do pico de casos da primeira onda de COVID-19 –, em contraste com uma correlação média de 0,17 presente período 2 – período associado ao início da segunda onda de casos da doença. Assim, o conjunto de palavras que denotam correlações significativas com o número de casos/mortes é enviesado pelos valores menos significativos do primeiro período estudado. Esse viés não é desejado, pois neste artigo busca-se encontrar palavras que possam ser utilizadas como reflexos do discurso acerca da pandemia no Twitter durante toda a extensão do primeiro ano da pandemia.

A partir da determinação desse três intervalos, foram geradas três listas ordenadas da maior para a menor correlação entre as variações na proporção das palavras e

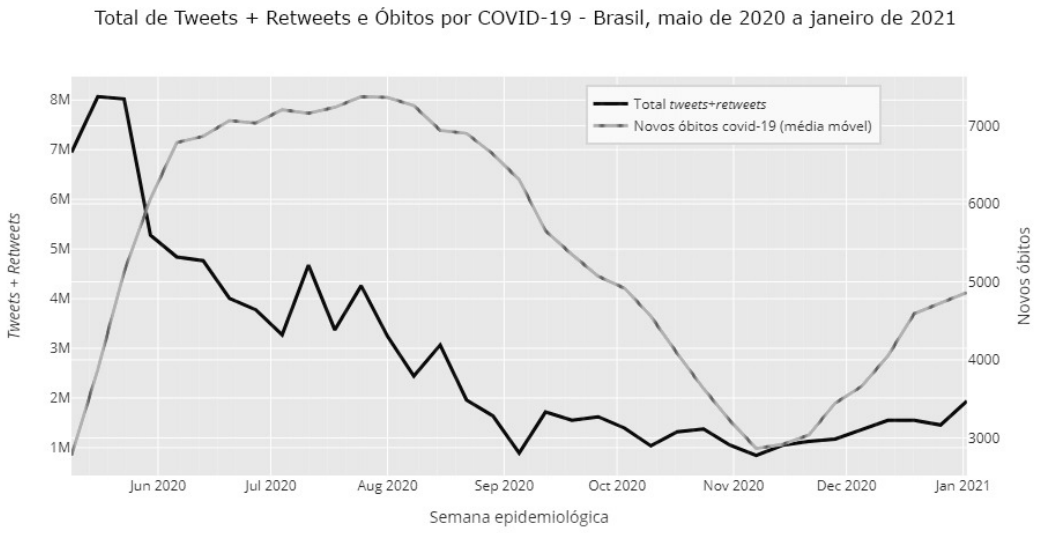

Figura 1. Série temporal do total de *tweets* e *retweets* no *dataset* e de óbitos por COVID-19.

as variações nos números de casos/mortes (uma lista para cada período). Em seguida, conforme mostra a Figura 2, as palavras foram manipuladas de acordo com o seguinte processo: a partir da seleção das n primeiras palavras de cada uma das três listas, foram adicionadas a uma quarta lista as palavras que pertencessem ao conjunto interseção dessas seleções. No caso da figura, a palavra *juntos* é adicionada à nova lista, pois é a única palavra que satisfaz à interseção correspondente a $n = 3$; em seguida, quando $n = 6$, é adicionada a palavra *clube*; e somente com $n = 7$ a palavra *series* é adicionada. Note que *series* é adicionada à quarta lista em uma posição abaixo à de *clube*, mesmo possuindo correlações mais elevadas. Esse mesmo processo é conduzido para as correlações negativas de forma invertida (isto é, a partir das menores até as maiores correlações).

Maiores correlações maio 2020 / agosto 2020	**Maiores correlações agosto 2020 / janeiro 2021**	**Maiores correlações maio 2020 / janeiro 2021**		**Interseção das n primeiras palavras**
0,76 - series - 3	0,57 - prefeitos	0,67 - genocidio		**1 - Juntos - 0,60**
0,70 - juntos - 1	0,53 - musica	**0,63 - series** - 3		**2 - Clube - 0,39**
0,52 - artista	**0,48 - juntos** - 1	**0,60 - juntos** - 1		**3 - Series 0,54**
0,34 - genocidio	**0,40 - clube** - 2	0,54 - saindo		**...**
0,30 - clube - 2	0,37 - ciencia	0,53 - artigo		
0,29 - esquerda	0,35 - recebi	**0,50 - clube** - 2		
0,25 saindo	**0,34 - series** - 3	0,48 - junho	n	
0,13 eleicoes	0,32 - saindo	0,45 - agosto		
...	...	...		

Figura 2. Ilustração do processo de seleção de palavras.

Em seguida, foram obtidas listas contendo as 25 primeiras palavras que satis-

fizessem a interseção entre as palavras com maiores correlações nos três intervalos de tempo estudados. Esse valor foi selecionado, empiricamente, pois fornece um número satisfatório de palavras para investigação, ao mesmo tempo em que permite uma análise qualitativa de cada termo. Esse processo foi repetido para cada um dos *lags* propostos (de 0 a 3 semanas).

2.3. Análise das Palavras Selecionadas

Por fim, foram analisadas as palavras presentes em cada uma das listas construídas seguindo os passos anteriormente descritos. Foi possível identificar certos temas e domínios recorrentes entre essas palavras, em particular: *sintomas da COVID-19*, *debate político*, *insultos* e *consumo de mídias*. Assim, pôde-se avaliar possíveis mudanças no teor das postagens e na abordagem ao tema da COVID-19 nas semanas que antecedem as variações nos números de casos/mortes.

3. Resultados e Discussão

3.1. Correlações com o Total de Casos

A Tabela 1 elenca as palavras com maior crescimento de uso em relação à semana em que é reportado aumento do número de casos de COVID-19. Os valores apresentados na tabela representam a média das correlações nos três períodos descritos na Seção 2.2. Nesta tabela, pode-se observar que, principalmente na semana que antecede o aumento do número dos casos (isto é, *lag* = 1), há uma presença relevante de termos diretamente relacionados à política (*partido*, *ministros*, *pazuello*[4], *militar*), além de insultos usualmente associados ao debate político (*globolixo*, *verme*, *gado*).

A significativa presença de termos do debate político e insultos nas semanas que antecedem o aumento do número de casos pode ser interpretada de pelo menos três formas diferentes: (a) uma insatisfação dos usuários da rede com relação às causas do iminente aumento do número de casos, como pronunciamentos públicos que incentivam o descumprimento de medidas de prevenção; (b) uma preocupação ou cobrança por medidas preventivas não adotadas, sobretudo em relação a algum evento ou período que possa vir a aumentar o número de contágios; e (c) uma reação aos próprios casos de COVID-19, uma vez que há um atraso para a notificação oficial – isto é, a maior reação pode vir no início do aumento dos casos, o que não corresponde ao período em que se registra, oficialmente, o crescimento.

A Tabela 2 destaca as palavras cuja utilização aumenta nas semanas anteriores a uma queda do número de casos. Nela, foram observados sobretudo dois temas de destaque: consumo de mídias e sintomas da COVID-19. Em relação ao primeiro, observaram-se termos relacionados ao consumo de mídias online, seja no próprio Twitter, seja em *lives* na internet ou em programas televisivos (*mpn*[5], *anygabriellyinspiracao*[6], *livelocalmariliamendonca*, *mtvhottest*[7], *thread*[8]), que não estão presentes na Tabela 1. Observou-se

[4]Ministro da Saúde durante a pandemia.
[5]Premiação "Meus Prêmios Nick".
[6]Referente a Any Gabrielly, da banda Now United.
[7]Evento promovido pela MTV.
[8]Termo referente à interação e produção de conteúdo no Twitter.

Tabela 1. Palavras obtidas com base na correlação positiva com o fator de crescimento de casos.

Correlação positiva com o fator de crescimento de casos			
lag = 0 semana	lag = 1 semana	lag = 2 semanas	lag = 3 semanas
0,44 - coragem	0,45 - muda	0,40 - gostar	0,54 - querido
0,41 - claro	0,39 - veja	0,38 - loja	0,51 - tomando
0,46 - errada	0,32 - brasileiros	0,37 - obvio	0,49 - medicamentos
0,45 - seguindo	0,32 - urgente	0,42 - goias	0,45 - tomar
0,36 - apareceu	0,31 - sonho	0,34 - contando	0,44 - junho
0,38 - razao	0,34 - mudar	0,34 - anticorpos	0,53 - puder
0,41 - ninguem	0,30 - terceiro	0,35 - metro	0,49 - f*****
0,34 - onibus	0,37 - partido	0,34 - sempre	0,46 - posto
0,33 - povo	0,29 - nsfw	0,38 - familias	0,44 - toma
0,38 - usam	0,29 - milhoes	0,32 - maioria	0,49 - brasileiros
0,28 - monte	0,30 - senhor	0,28 - conteudo	0,45 - gado
0,32 - foto	0,29 - verme	0,30 - materia	0,43 - tomei
0,28 - tranquilo	0,32 - hidroxicloroquina	0,34 - rodrigo	0,46 - mostrando
0,32 - mesma	0,34 - usa	0,27 - espalhar	0,41 - curado
0,30 - ignorancia	0,30 - ajudou	0,38 - alta	0,43 - populacao
0,31 - especialistas	0,28 - portal	0,34 - consequencias	0,45 - acha
0,31 - cura	0,27 - ministros	0,30 - pulmao	0,41 - quadro
0,30 - moral	0,29 - obito	0,39 - avos	0,44 - funciona
0,31 - propria	0,31 - comprar	0,33 - mar	0,45 - propaganda
0,29 - divulgar	0,31 - globolixo	0,33 - profissional	0,40 - tomou
0,30 - obvio	0,28 - comprando	0,31 - global	0,47 - tranquilo
0,27 - enquanto	0,38 - maisa	0,27 - feio	0,37 - vai
0,27 - dizendo	0,26 - problemas	0,34 - coragem	0,38 - julho
0,34 - f*****	0,28 - pazuello	0,30 - muitos	0,43 - pesquisa
0,28 - imagem	0,28 - militar	0,33 - vinho	0,39 - duvido

também que, na Tabela 2, não aparecem termos ofensivos (como *verme*, *f*****, *ignorante*, entre outros, presentes na Tabela 1, evidenciando um tom mais pacífico e ameno nas semanas que antecedem às quedas do número de casos.

Além disso, a alta frequência de palavras relacionadas ao consumo de conteúdo midiático (redes sociais online, internet, música, televisão) sugere que os usuários do Twitter possam ter consumido mais mídia em casa nesses períodos, ou que estivessem "mais entretidos", o que pode ter contribuído para um menor espalhamento do vírus.

Todavia, ainda na Tabela 2 foi observada a menção a sintomas de COVID-19 (*garganta*, *febre*, *sintoma*, *tosse*). Isso pode ser considerado surpreendente, por ser um tema que, em outros estudos, costuma ser utilizado como preditor do crescimento de casos de doenças [Gomide et al. 2011, Locatelli et al. 2022], e não do decréscimo. Não está claro o porquê desse padrão, especialmente por ele se concentrar na própria semana de diminuição dos casos, o que motiva trabalhos futuros para que esse fenômeno seja analisado em maior profundidade.

3.2. Correlações com o Total de Mortes

A Tabela 3 indica as palavras obtidas com base na correlação positiva com o fator de crescimento de mortes por COVID-19. Nela, há dois pontos de destaque principais: com pouco ou nenhum *lag*, novamente os dados refletem um crescimento no uso de termos relacionados a temas políticos e insultos (*bolsominion*, *exercito*, *militar*, *ridiculo*, *especialistas*, *arrombado*, *hipocrita*, *trouxa*). Há, também, palavras relacionadas a saúde, inclusive a medicamentos e a protocolos de prevenção contra a doença (*oms*, *remedios*, *anticorpos*, *azitromicina*, *drogas*, *aglomeracao*, *tomando*, *medicamentos*, *colaterais*, *sus*,

Tabela 2. Palavras obtidas com base na correlação negativa com o fator de crescimento de casos.

Correlação negativa com o fator de crescimento de casos			
lag = 0 semana	lag = 1 semana	lag = 2 semanas	lag = 3 semanas
-0,60 - perto	-0,47 - parar	-0,39 - anvisa	-0,53 - tosse
-0,47 - desta	-0,47 - pedindo	-0,41 - completamente	-0,58 - perceber
-0,44 - suspeitos	-0,41 - banco	-0,36 - evento	-0,60 - anygabriellyinspiracao
-0,55 - anygabriellyinspiracao	-0,48 - voltaram	-0,36 - ruas	-0,45 - especial
-0,48 - tosse	-0,43 - dentro	-0,42 - sigo	-0,44 - protecao
-0,39 - garganta	-0,41 - olhar	-0,35 - amanha	-0,46 - atualizacao
-0,39 - novembro	-0,44 - brincadeira	-0,37 - mostrou	-0,43 - mulheres
-0,37 - febre	-0,38 - pegando	-0,31 - mudou	-0,42 - chamei
-0,45 - mpn	-0,42 - enfrentar	-0,32 - vai	-0,46 - portugal
-0,37 - precisava	-0,53 - mpn	-0,36 - gosta	-0,47 - tou
-0,37 - sono	-0,47 - sofrendo	-0,29 - thread	-0,46 - panico
-0,36 - leia	-0,39 - vindo	-0,33 - terceiro	-0,45 - assustou
-0,40 - fazer	-0,34 - ter	-0,39 - lute	-0,41 - novas
-0,34 - chamei	-0,42 - foda	-0,28 - app	-0,42 - bahia
-0,35 - sintoma	-0,41 - situacao	-0,33 - caixa	-0,47 - infetados
-0,38 - proximos	-0,37 - consequencias	-0,30 - educacao	-0,44 - coronavirus
-0,37 - carente	-0,35 - imaginar	-0,27 - menor	-0,49 - apanhar
-0,35 - mulheres	-0,39 - chuva	-0,29 - sistema	-0,36 - google
-0,37 - comigo	-0,40 - aparecer	-0,34 - maio	-0,54 - fiqueemcasa
-0,32 - surtar	-0,35 - feriado	-0,30 - mostrar	-0,43 - livelocalmariliamendonca
-0,39 - conto	-0,34 - surpresa	-0,36 - podia	-0,39 - conhecer
-0,37 - vez	-0,35 - fechado	-0,26 - amem	-0,39 - saudades
-0,34 - existir	-0,35 - espalhar	-0,29 - virus	-0,39 - praticamente
-0,35 - fisica	-0,38 - subir	-0,28 - fantastico	-0,39 - crise
-0,32 - cortar	-0,38 - contagio	-0,31 - cair	-0,38 - mtvhottest

ivermectina, *morrem*, *doses*, *lotada*). Essas palavras se concentram na semana anterior ao aumento do número de mortes, o que sugere que isso é uma reação ao aumento do número de casos/mortes. Além disso, três semanas antes do aumento das mortes, é possível observar o aumento do uso das palavras *ivermectina*, *doses* e *medicamentos*.

A respeito das palavras de correlação negativa com as mortes, apresentadas na Tabela 4, pode-se observar os mesmos padrões temáticos que aparecem na Tabela 2, com termos como *olfato*, *garganta* e *nariz*, representando os sintomas da doença logo antes da diminuição das mortes, e *anygabriellyinspiracao*, *mpn*, *mucalol*[9], *shakira*, *exposed*, *taekook*[10] e *fiqueemcasa* como representantes da categoria de mídias que podem ser consumidas em casa e do incentivo ao distanciamento social.

Os resultados sugerem que, a partir das palavras selecionadas em diferentes níveis de *lag* em relação aos casos e, sobretudo, às mortes, é possível identificar diferentes temas abordados nos *tweets* postados pelos usuários da rede. Observa-se que cada conjunto de palavras remete a um contexto diferente, indicando que, dentre os *tweets* postados durante dado período, determinados assuntos foram predominantes. Essa observação responde ao questionamento central do estudo, evidenciando que é possível caracterizar uma relação entre o perfil do discurso na rede, aqui caracterizado pelas palavras apresentadas, e as variações na intensidade da evolução da pandemia ao longo do tempo.

[9] Youtuber brasileiro.

[10] Membro do grupo sul-coreano BTS.

Tabela 3. Palavras obtidas com base na correlação positiva com o fator de crescimento de mortes.

Correlação positiva com o fator de crescimento de mortes			
lag = 0 semanas	lag = 1 semana	lag = 2 semanas	lag = 3 semanas
0,34 - bolsominion	0,44 - fura	0,45 - familias	0,45 - junho
0,34 - exercito	0,32 - agradecer	0,37 - sempre	0,45 - medicamentos
0,35 - mil	0,28 - morrem	0,39 - namorado	0,38 - julho
0,29 - claro	0,31 - terceiro	0,35 - venceu	0,37 - militar
0,28 - ridiculo	0,29 - news	0,36 - obvio	0,39 - crianca
0,32 - obvio	0,37 - drogas	0,37 - igreja	0,32 - ogos
0,35 - povo	0,40 - trouxa	0,36 - presente	0,40 - tomando
0,27 - inves	0,33 - muda	0,38 - primo	0,38 - pesquisa
0,31 - cura	0,32 - atraves	0,32 - acao	0,38 - chegando
0,35 - importa	0,28 - valor	0,41 - rodrigo	0,36 - brasileira
0,24 - especialistas	0,30 - maior	0,35 - achava	0,32 - posto
0,34 - brasileiro	0,28 - lotada	0,34 - passa	0,33 - crime
0,28 - arrombado	0,29 - milhao	0,33 - pressao	0,35 - colaterais
0,25 - resolve	0,26 - azitromicina	0,33 - dou	0,37 - querido
0,36 - aonde	0,32 - militar	0,30 - bons	0,36 - f*****
0,25 - hipocrita	0,28 - aglomeracao	0,31 - feio	0,36 - muda
0,29 - faz	0,27 - problemas	0,31 - parentes	0,29 - sus
0,26 - disso	0,32 - india	0,33 - coragem	0,37 - mostrando
0,24 - propria	0,33 - aparentemente	0,28 - sido	0,30 - tranquilo
0,27 - remedios	0,32 - brasileiros	0,31 - maia	0,39 - trata
0,36 - oms	0,29 - vale	0,29 - combate	0,33 - ivermectina
0,32 - anticorpos	0,31 - avisar	0,31 - errada	0,34 - hora
0,26 - esperava	0,23 - louco	0,30 - fura	0,37 - doses
0,22 - comentario	0,24 - necessidade	0,27 - inveja	0,27 - parte
0,25 - entao	0,31 - policia	0,35 - jesus	0,34 - coreia

Dentre os perfis temáticos traçados a partir da observação das correlações com a evolução da pandemia, destacam-se três achados: (i) é possível observar uma forte carga política e de termos predominantemente negativos nas semanas que antecedem o aumento do número de casos/mortes; (ii) há um aumento no número de referências ao conteúdo midiático de entretenimento nas semanas que antecedem o decréscimo no número de casos/mortes; e (iii) de forma contraintuitiva, a menção a termos relacionados aos sintomas da COVID-19 se relaciona à diminuição, e não ao aumento, dos casos e mortes pela doença, como observado na literatura [Gomide et al. 2011, Shen et al. 2020, Li et al. 2020, Doan et al. 2012, Locatelli et al. 2022].

4. Considerações Finais

Este trabalho apresenta uma proposta inicial de análise lexical que correlaciona a variação na utilização de palavras no Twitter com fatores de crescimento de casos/mortes por COVID-19 no primeiro ano da pandemia no Brasil. Análises qualitativas dos resultados obtidos sugerem que é possível, ao menos parcialmente, reconstruir e tecer conjecturas sobre os tópicos presentes no corpo de *tweets* associando as oscilações na proporção das palavras e nos indicadores epidemiológicos, inclusive considerando *lags* temporais de até três semanas. Os resultados encontrados demonstram como o aumento no uso de palavras de tom negativo antecede o aumento do número de casos/mortes, ao passo que o aumento no uso de palavras de tom mais positivo antecede o decréscimo do número de casos/mortes – o que pode se relacionar, inclusive, a questões vinculadas à saúde mental durante o período do primeiro ano de isolamento social causado pela pandemia. Ademais, o resultado que demonstra um aumento no número de referências a conteúdos midiáticos

Tabela 4. Palavras obtidas com base na correlação negativa com o fator de crescimento de mortes.

Correlação negativa com o fator de crescimento de mortes			
lag = 0 semanas	lag = 1 semana	lag = 2 semanas	lag = 3 semanas
-0.46 - conseguir	-0.48 - diria	-0.44 - resolver	-0.58 - assustou
-0.43 - olfato	-0.48 - mpn	-0.40 - podia	-0.51 - perceber
-0.41 - livros	-0.47 - parar	-0.43 - deixar	-0.54 - anygabriellyinspiracao
-0.45 - anygabriellyinspiracao	-0.55 - novamente	-0.43 - anvisa	-0.49 - devia
-0.45 - emprego	-0.42 - aparecer	-0.33 - proteger	-0.46 - chamei
-0.39 - perto	-0.36 - feio	-0.36 - sistema	-0.52 - proximos
-0.34 - senti	-0.41 - brincadeira	-0.32 - voto	-0.42 - eleicoes
-0.49 - taxa	-0.41 - imaginar	-0.30 - praca	-0.39 - dou
-0.36 - garganta	-0.36 - pegando	-0.30 - queda	-0.38 - amor
-0.41 - dificuldade	-0.38 - sofrendo	-0.35 - pfizer	-0.44 - feriado
-0.46 - desta	-0.37 - shakira	-0.31 - aproveita	-0.42 - dias
-0.44 - mpn	-0.37 - confirmou	-0.29 - lado	-0.34 - pontos
-0.37 - horrivel	-0.29 - nariz	-0.26 - exposed	-0.37 - especial
-0.42 - fisica	-0.32 - proxima	-0.30 - amanha	-0.42 - passe
-0.38 - online	-0.42 - anygabriellyinspiracao	-0.35 - educacao	-0.56 - tosse
-0.32 - novembro	-0.35 - pedindo	-0.27 - solta	-0.35 - eleicao
-0.32 - lotados	-0.33 - aumentou	-0.28 - parou	-0.35 - duvida
-0.42 - precisando	-0.36 - alerta	-0.30 - locais	-0.33 - google
-0.34 - ganhar	-0.37 - pede	-0.43 - mpn	-0.36 - f*****
-0.30 - mucalol	-0.33 - chorar	-0.37 - maio	-0.41 - praticamente
-0.29 - cortar	-0.34 - jogou	-0.26 - volta	-0.44 - fiqueemcasa
-0.35 - programa	-0.27 - deram	-0.32 - analise	-0.34 - taekook
-0.36 - suspeitos	-0.39 - homem	-0.25 - vitoria	-0.32 - vice
-0.29 - meninas	-0.27 - luz	-0.24 - negocio	-0.39 - foda
-0.31 - virtual	-0.33 - esperar	-0.25 - prazo	-0.38 - suspeita

de entretenimento nas semanas que antecedem o decréscimo no número de casos/mortes é especialmente relevante, pois indica que esses termos podem ser bons preditores da contenção da doença e reflexos interessantes da adesão às medidas de distanciamento social por parte da população brasileira no primeiro ano da pandemia. Esses resultados seriam beneficiados pela aplicação de métodos de análise de sentimentos/polaridade que pudessem avaliar o teor positivo ou negativo das palavras mais correlacionadas com o agravamento ou atenuação da pandemia [Maia et al. 2021].

Por outro lado, os resultados alertam que, para a COVID-19, a previsão baseada nos sintomas da doença, ainda que possa ser eficaz em alguns contextos, no recorte de *tweets* conduzido nesta pesquisa não se traduz em uma correlação positiva com o número de casos/mortes. Esses resultados, portanto, não parecem ser explicados pela hipótese de que indivíduos estejam relatando, na rede social, sintomas que estão sentindo [Gomide et al. 2011, Shen et al. 2020, Li et al. 2020, Doan et al. 2012, Locatelli et al. 2022]. Sobre esse tema, pretende-se, em trabalhos futuros, explorar um modelo preditivo do número de casos/mortes por COVID-19 com base na variação da frequência de certas palavras no Twitter, mais especificamente aquelas relacionadas ao recorte temático de consumo de mídias online, a ser favorecido por um modelo de interpretação [Molnar 2022, Ribeiro et al. 2016].

Por fim, é importante destacar que as análises aqui apresentadas foram realizadas com dados intrinsecamente anonimizados, uma questão cada vez mais relevante no contexto de *big data* [Zwitter 2014, Zimmer and Proferes 2014], e de baixo custo com-

putacional, uma vez que tabelas contendo o total de ocorrências das palavras (utilizadas aqui) são muito mais leves e fáceis de se tratar do que milhões de postagens completas. No futuro, espera-se que os resultados possam ser comparados àqueles advindos de métodos bem estabelecidos no âmbito da análise de redes sociais online, assegurando a eficácia desta metodologia.

Agradecimentos. Este trabalho foi realizado com apoio financeiro do CNPq, FAPEMIG, CAPES e dos projetos Covid Data Analytics (PRPq/UFMG e SESU/MEC), MASWEB, INCT-Cyber e Atmosphere.

Referências

Aiello, A. E., Renson, A., and Zivich, P. N. (2020). Social media- and Internet-based disease surveillance for public health. *Annual Review of Public Health*, 41(1):101–118.

Bae, Y., Ryu, P.-M., and Kim, H. (2014). Predicting the lifespan and retweet times of tweets based on multiple feature analysis. *ETRI Journal*, 36(3):418–428.

Brum, P. V., Teixeira, M. C., Miranda, R., Vimieiro, R., Meira Jr., W., and Pappa, G. L. (2020). A characterization of Portuguese tweets regarding the Covid-19 pandemic. In *VIII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 177–184. Sociedade Brasileira de Computação (SBC).

Doan, S., Ohno-Machado, L., and Collier, N. (2012). Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 62–71.

França, T., Faria, F., Rangel, F., De Farias, C., and Oliveira, J. (2014). Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. In Lóscio, B. F., Hara, C. S., and Martins, V., editors, *Tópicos em Gerenciamento de Dados e Informações*, pages 8–45. UFPR; PUC-PR, Curitiba.

Freire, F., Gonzaga, M., and Queiroz, B. (2019). Projeção populacional municipal com estimadores bayesianos, Brasil 2010-2030. *Seguridade Social Municipais. Projeto Brasil*, 3.

Gomide, J., Veloso, A., Meira Jr., W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proc. of the 3rd Web Science Conf. (WebSci'11)*. ACM.

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., and Chen, H. (2020). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance*, 25(10).

Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., Yuan, B., Kinoshita, R., and Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2):538.

Locatelli, M. S., Cunha, E. L. T. P., Guiginski, J., Franco, R. A. S., Bernardes, T., Alzamora, P. L., da Silva, D. V. F., Ganem, M. A. S., Santos, T. H. M., Carvalho, A.

I. R., Souza, L. M. V., Paixão, G. P. F., Chaves, E. F., dos Santos, G. B., dos Santos, R. V., de Freitas, A. C., Flores, M. G., Biezuner, R. F., Cardoso, R. L., Fonseca, R. M., Couto da Silva, A. P., and Meira Jr., W. (2022). Correlations between web searches and COVID-19 epidemiological indicators in Brazil. *Brazilian Archives of Biology and Technology*, 65.

Maia, M., Oliveira, E., and Gallegos, L. (2021). Covid-19 e tweets no brasil: coleta, tratamento e análise de textos com evidências de estados afetivos alterados em momentos impactantes. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 79–90, Porto Alegre, RS, Brasil. SBC.

Malagoli, L. G., Stancioli, J., Ferreira, C. H. G., Vasconcelos, M., Couto da Silva, A. P., and Almeida, J. M. (2021). A look into COVID-19 vaccination debate on Twitter. In *13th ACM Web Science Conference (WebSci'21)*, pages 225–233, New York. ACM.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2 edition.

Moreira, P., Fonseca, R., Alzamora, P., Franco, R. A., Guiginski, J., Cunha, E., Bernardes, T., Chagas, B., Ferregueti, K., Passos, L., Cardoso, L., Schneider, R., Pereira, W., da Silva, A. P., and Jr., W. M. (2021). Covid Data Analytics: Repositório de dados provenientes de múltiplas fontes sobre a pandemia de COVID-19 no Brasil. In *Anais do III Dataset Showcase Workshop*, pages 107–116, Porto Alegre, RS, Brasil. SBC.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM).

Shen, C., Chen, A., Luo, C., Zhang, J., Feng, B., and Liao, W. (2020). Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: Observational infoveillance study. *Journal of Medical Internet Research*, 22(5):e19421.

Sousa, G., Garces, T., Cestari, V., Florêncio, R., Moreira, T., and Pereira, M. (2020). Mortality and survival of COVID-19. *Epidemiology and Infection*, 25(148):e123.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Sultana, A., Tasnim, S., Mahbub Hossain, M., Bhattacharya, S., and Purohit, N. (2021). Digital screen time during the COVID-19 pandemic: a public health concern. *F1000Research*, 10(81).

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 20:669–677.

Weng, J. and Lee, B.-S. (2011). Event detection in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):401–408.

Zimmer, M. and Proferes, N. J. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3):250–261.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2).