Análise do estresse e tópicos discutidos no Twitter durante a pandemia da COVID-19 no Brasil*

Diansley R. S. Peres[1], Gean F. da Silva[1], Elaine R. Faria[1], Maria Camila N. Barioni[1]

[1]Faculdade de Computação – Universidade Federal de Uberlândia (UFU)
Campus Santa Mônica - Bloco 1A - Sala 1A236 – Uberlândia – MG – Brazil

{diansley, gean.fgsilva, elaine, camila.barioni}@ufu.br

Abstract. *This work proposes an application to measure the incidence of stress during the COVID-19 pandemics using the TensiStrength (TS) algorithm adapted for Portuguese and natural language processing techniques in tweets. As a result, it was possible to validate the TS to measure stress and relaxation, as well as to describe discussions related to the pandemics in Brazil through different topic extraction algorithms and word cloud visualization.*

Resumo. *Este trabalho propõe um aplicação para mensurar a incidência de estresse durante a pandemia da COVID-19 por meio do algoritmo TensiStrength (TS) adaptado para o português e de técnicas de processamento de linguagem natural em tweets. Como resultado, foi possível validar o TS para mensurar o estresse e relaxamento, bem como descrever as discussões relacionadas à pandemia no Brasil por meio de diferentes algoritmos de extração de tópicos e visualização de nuvens de palavras.*

1. Introdução

A pandemia do coronavírus (SARS-CoV-2) tem assolado países de todo o mundo desde 2019, quando a doença foi detectada em Wuhan (China)[1]. Somente no Brasil, até setembro de 2022, mais de 34 milhões de casos da doença foram confirmados com aproximadamente 697 mil óbitos registrados[2]. Além dos diversos efeitos físicos da doença, que incluem insuficiência respiratória, febre e, em casos mais graves, a falência de múltiplos órgãos, outro efeito da pandemia pode ser verificado através do desenvolvimento de doenças como alcoolismo, depressão e desenvolvimento de estresse pós-traumático [Afonso 2020].

Pesquisas realizadas em diferentes países forneceram evidências de que houve um aumento na incidência de transtornos mentais desde o início da pandemia. Entre os fatores associados aos transtornos desenvolvidos verificam-se: dificuldades financeiras, o isolamento social, a incidência da doença e óbito de pessoas próximas, a exposição excessiva a notícias sobre a COVID-19 e níveis inferiores de escolaridade dos indivíduos [Afonso 2020][Brown et al. 2020][Taylor et al. 2020].

O estresse pode ser entendido como a resposta do organismo humano à pressão. A condição geralmente surge quando as pessoas vivenciam algo novo, inesperado ou quando

*Agradecemos à FAPEMIG (Fundação de Amparo à Pesquisa do Estado Minas Gerais) pelo apoio financeiro a este projeto (Projeto APQ-00226-21)

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

[2]https://covid.saude.gov.br/

existe a sensação de pouco controle sobre a situação enfrentada. Em excesso o estresse pode ser prejudicial, levando o indivíduo a um estado permanente de fuga ou enfrentamento e afetando sua saúde física e mental[3]. É esperado que as pessoas experienciem estresse durante a pandemia da COVID-19, de forma que uma parte da população seja consideravelmente afetada por essa condição[4].

Durante crises como a pandemia do coronavírus as pessoas utilizaram as redes sociais com diferentes finalidades. Logo, identificar o conteúdo publicado pode contribuir para uma resposta mais adequada por parte das autoridades em contextos de emergência. Analisar o conteúdo postado pelos usuários pode ajudar a identificar demandas relacionadas a medidas de segurança, pedidos de ajuda, combate de rumores e *fake news* e identificação de usuários com diversos fins [Li et al. 2020]. Dessa forma, as redes sociais podem atuar como importante meio de gerenciamento de crises, tomadas de ação e meio propagador da consciência situacional de uma crise por parte da população [Freitas et al. 2020].

Técnicas não supervisionadas de modelagem de tópicos têm sido utilizadas para organizar, entender e sumarizar grandes conjuntos de textos e descobrir tópicos latentes em documentos textuais. A Alocação Latente de Dirichlet (LDA) e a Matriz de Fatorização Não Negativa (NMF) são duas das mais populares técnicas de modelagem de tópicos, além de algoritmos de *Deep Learning* baseados em Representações Codificadoras Bidirecionais de Transformadores (BERT) [Abuzayed and Al-Khalifa 2021].

Redes sociais, como o *Twitter*, são espaços onde os usuários compartilham diariamente suas visões, atitudes e opiniões. Diversas técnicas de modelagem de tópicos têm sido empregadas por trabalhos que analisam dados no contexto da pandemia por meio de redes sociais, como no trabalho publicado por [de Sousa and Becker 2021]. No entanto, não foram identificados trabalhos que associam tópicos derivados de algoritmos de processamento de linguagem natural (PLN) com níveis de estresse do usuário em textos em Português.

O presente trabalho tem como objetivo mensurar, através do algoritmo *TensiStrength* adaptado para o Português (TSpt), o nível de estresse presente em publicações realizadas no *Twitter*, bem como descrever através de algoritmos de extração de tópicos as discussões em diferentes momentos da pandemia. Verificou-se que o TSpt gerou resultados mais bem ajustados em comparação aos apresentados pelos autores do TS, bem como uma associação estatisticamente significativa entre os rótulos obtidos por meio do TSpt e da anotação de juízes humanos. A extração de tópicos, realizada a partir dos algoritmos NMF e BERTopic demonstrou discussões relacionadas, por exemplo, à vacinação, ao cenário político no Brasil e à medidas de prevenção.

O restante do artigo está organizado como descrito a seguir. Os principais conceitos necessários para o entendimento do trabalho descrito aqui são apresentados na Seção 2. A Seção 3 apresenta os trabalhos correlatos. O método de trabalho é descrito na Seção 4. A Seção 5 apresenta a discussão dos resultados obtidos. Por fim, a Seção 6 apresenta as conclusões.

[3]https://www.mentalhealth.org.uk/explore-mental-health/a-z-topics/stress

[4]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

2. Conceitos Fundamentais

Os conceitos relacionados com as estratégias empregadas no desenvolvimento do método de trabalho, descrito na Seção 4, são: *LDA*, *NMF*, *BERTopic* e *TensiStrength*. Essas estratégias serão brevemente descritas a seguir.

O *TensiStrength* é um sistema composto por uma abordagem léxica e um conjunto de regras diretas e indiretas para detecção de expressões de estresse e relaxamento [Thelwall 2017]. O algoritmo possui uma lista de termos em inglês relacionados a estresse e relaxamento. Trata-se de termos sinônimos de estresse, ansiedade, frustração e termos relacionados à raiva e emoções negativas, estados ou situações de relaxamento e indicadores de sentimentos positivos. Cada termo do dicionário possui uma pontuação de acordo com o nível de estresse e relaxamento. A cada senteça é atribuído o maior escore do termo de estresse e de relaxamento identificado. No caso de textos com múltiplas sentenças é atribuída a maior pontuação entre todas as sentenças.

Latent Dirichlet Allocantion, ou LDA, é um modelo probabilístico que representa documentos como combinações aleatórias sobre tópicos latentes formados por distribuições de palavras. O LDA supõe que documentos com tópicos similares usam grupos similares de palavras e que tópicos podem ser encontrados ao procurar grupos de palavras que ocorrem juntas na coleção de documentos, além das suposições de que a ordem de ocorrência das palavras e dos documentos não importa e que o número de tópicos é assumido, corrigido e conhecido [Martins et al. 2020].

Non-negative Matrix Factorization (NMF) é um método não probabilístico e decomposicional que utiliza matrizes de fatorização a partir de dados no formato TF-IDF, que medem a importância de uma palavra em uma coleção de documentos. O algoritmo decompõe a matriz de entrada em matrizes não negativas W e H, com os termos extraídos e os pesos de cada termo, respectivamente [Egger and Yu 2022].

O *BERTopic* é um método de extração de representações de tópicos latentes em que documentos são gerados com uma camada de pré-treinamento baseada em modelos de linguagem. As camadas de pré-treinamento são agrupadas e, por fim, representações dos documentos são geradas a partir de uma variação do método TF-IDF [Grootendorst 2022].

3. Trabalhos Correlatos

Vários pesquisadores têm desenvolvido trabalhos de análise de textos publicados em redes sociais a partir de algoritmos de extração de tópicos e análise de vocabulários. Trabalhos recentes foram realizados com esse objetivo no contexto da pandemia da COVID-19 [Menuzzo et al. 2021] [Ljajić et al. 2022][Alzamora et al. 2022][de Sousa and Becker 2022].

[Menuzzo et al. 2021] avaliaram o impacto do discurso de governantes municipais sobre o comportamento da população durante a pandemia. No estudo, foram avaliados a diversidade e coesão dos discursos por meio de publicações realizadas no Facebook combinadas com dados epidemiológicos. Os autores apresentaram dois diferentes modelos baseados em LDA para a análise dos tópicos e da coesão relacionada à evolução da pandemia.

Um estudo recente avaliou possíveis razões para a resistência à vacinação contra a COVID-19 na Sérvia. Os autores coletaram *tweets* com alguma menção à vacinação.

Uma primeira leva de publicações foi avaliada manualmente em termos de relevância e polaridade de sentimento. Um modelo baseado em tranformadores BERT foi então executado para a classificação de sentimentos dos demais *tweets* não anotados. Após a classificação de sentimentos foi realizada a extração de tópicos por meio do LDA e do NMF para avaliar as razões para a resistência à vacinação nos *tweets* com sentimentos negativos [Ljajić et al. 2022].

[de Sousa and Becker 2022] desenvolveram um estudo temporal das posturas pró e contra a vacinação da COVID-19 nos Estados Unidos e no Brasil utilizando dados do *Twitter*. Foram avaliados os principais argumentos para defender cada posicionamento nos EUA, a evolução desses posicionamentos ao longo do tempo e a relação com o cenário brasileiro por meio da modelagem de tópicos BERTopic.

[Alzamora et al. 2022] buscaram caracterizar o primeiro ano da pandemia no Brasil avaliando a correlação entre o agravamento e atenuação da pandemia e o vocabulário utilizado no *Twitter* nas semanas anteriores às variações.

Também existem diversos trabalhos na literatura com o objetivo de prever o nível de estresse dos usuários por meio de técnicas de mineração de textos publicados nas redes sociais, como no caso de [Lin et al. 2014] e [Wang et al. 2020].

[Lin et al. 2014] pesquisaram a aplicação de um modelo baseado em redes neurais profundas para detectar de forma automática o estresse psicológico dos usuários por meio das redes sociais. Tanto o conteúdo de cada publicação, quanto os atributos estatísticos dos usuários, tais como data e hora da publicação, estilo linguístico e engajamento social, foram levados em consideração no trabalho.

O trabalho de [Wang et al. 2020] se baseou na construção de um *framework* para a detecção do estresse em 3 níveis: genérico, em grupo e individual, usando modelos baseados em redes neurais. Foram avaliadas publicações em Inglês da rede *Sina Weibo*.

Assim, percebe-se que os trabalhos coletaram nas redes sociais publicações em diferentes idiomas combinadas com técnicas de extração de tópicos para resumir a reação pública frente à COVID-19 ou utilizaram algoritmos para a detecção de estresse nos indivíduos em diferentes contextos. Não foram verificados, no entanto, trabalhos que procurassem explicar o estresse presente nas publicações em Português realizadas durante a pandemia utilizando técnicas de extração de tópicos.

4. Método de Trabalho

O trabalho descrito aqui teve como objetivo descrever as publicações realizadas em diferentes momentos da pandemia do coronavírus no Brasil em termos de estresse e relaxamento. A abordagem proposta consiste de um método não supervisionado organizado em 4 etapas: 1) coleta de Dados do *Twitter*; 2) pré-processamento dos *tweets*; 3) aplicação do algoritmo TSpt para a classificação das publicações em termos de estresse e relaxamento e 4) avaliação da proporção de estresse, geração de nuvens de palavras e extração de tópicos dos cenários avaliados.

4.1. Coleta de Dados

Para a realização deste trabalho optou-se por avaliar os *tweets* publicados ao longo de diferentes momentos da pandemia no Brasil. O *Twitter* foi escolhido como rede social

foco da pesquisa devido aos recursos gratuitos disponibilizados pela plataforma para a coleta de dados, realizada por meio da biblioteca *snscrape*[5] no ambiente *Python*. Para a busca das publicações, considerou-se as postagens identificadas por meio da palavra-chave *covid*. Além das bases de avaliação, foram coletadas, num primeiro momento da pesquisa, publicações realizadas entre abril e maio de 2021, selecionadas por meio da mesma palavra-chave. Essa base de dados, composta por 33.703 *tweets* e sujeita às mesmas etapas de pré-processamento que as bases de avaliação, foi mantida no escopo deste trabalho para testes e avaliação de parâmetros. Para a definição dos períodos que seriam avaliados, levou-se em consideração marcos históricos da pandemia no Brasil, identificados por meio de acervos online[6][7].

Os períodos considerados em cada coleta, a quantidade de *tweets* e de usuários coletados podem ser verificados na Tabela 1. A fim de obter um volume maior de publicações para serem avaliadas neste trabalho, realizou-se a coleta em um período de 2 meses em torno de cada evento selecionado.

Tabela 1. Períodos considerados para a coleta de *tweets*.

Período	Descrição	*Tweets*	Usuários.
02/2020 - 03/2020	03/2020: Primeiros registros de transmissão interna do coronavírus no Brasil	26.180	15.182
06/2020 - 07/2020	07/2020: Aproximadamente $\frac{1}{3}$ das mortes registradas desde o início da pandemia ocorreram em julho de 2020	62.624	29.000
11/2020 - 12/2020	12/2020: Início da segunda onda de contágio da COVID-19 no Brasil	45.753	23.104
01/2021 - 02/2021	01/2021: Autorização do uso emergencial das vacinas CoronaVac e Oxford pela Anvisa	42.441	21.366
06/2021 - 07/2021	07/2021: O Brasil atinge a marca de 500.000 mortes causadas pelo coronavírus	33.497	17.280
12/2021 - 01/2022	12/2021: Anúncio da variante Ômicron pela OMS como uma variante de preocupação depois de descoberta na África do Sul	42.702	21.399
04/2022 - 05/2022	04/2022: Desobrigação do uso de máscaras em todos os estados brasileiros, com índice de mortes causadas pela COVID-19 $< 0,3$ mortes por 100.000 habitantes	8.094	5.569

4.2. Pré-processamento

Todo o pré-processamento foi realizado através de bibliotecas *Python*. Foram aplicadas as seguintes tarefas em cada uma das bases de dados: conversão de *emojis* em palavras; remoção de sinais de pontuação; conversão do texto para letras minúsculas; remoção de menções a usuários e *URLs*; remoção de caracteres numéricos; remoção de palavras sinônimas e associadas à COVID-19, tais como *morte*, *covid19*, *coronavírus* e *pandemia*, remoção de palavras de parada e lematização da base de dados coletada. Foram utilizadas, ainda, nuvens de palavras para avaliar e refinar cada etapa do pré-processamento aplicado.

4.3. Aplicação do Algoritmo TensiStrength

Para a detecção dos níveis de estresse e relaxamento nas publicações utilizou-se o algoritmo TS fornecido por [Thelwall 2017]. Foram executadas, em uma amostra de *tweets*, duas versões do algoritmo: a versão original, em inglês, e uma versão adaptada para o

[5]https://github.com/JustAnotherArchivist/snscrape

[6]https://memoriadaeletricidade.com.br/comunicacao-memoria/117830/linha-do-tempo-covid-19

[7]https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil

Português (TSpt). Para executar o *TensiStrength* foi necessário utilizar uma ferramenta de tradução através da biblioteca *googletrans*[8] para traduzir os *tweets* coletados e o dicionário do algoritmo nas aplicações do TS e TSpt, respectivamente. Na versão original traduziu-se a amostra de *tweets* para o Inglês, enquanto na versão TSpt o dicionário do algoritmo foi traduzido para o Português. O desempenho das duas versões do algoritmo foi comparado com base nas mesmas métricas escolhidas pelos autores do TS.

A aplicação do TSpt foi ainda comparada com a rotulagem do estresse e do relaxamento presentes nas sentenças a partir da avaliação de juízes humanos. O manual para a rotulação utilizado no trabalho desenvolvido por [Thelwall 2017] foi traduzido e disponibilizado para que 2 voluntários na pesquisa classificassem o nível de estresse e relaxamento na amostra de 386 *tweets*. A pontuação de estresse, varia de -5 a -1, enquanto a pontuação de relaxamento varia de 1 a 5, tanto nas classificações do TS e TSpt quanto na dos juízes. Para cada *tweet* foi gerada a média dos escores de estresse e relaxamento fornecidos pelos juízes. Somadas as duas pontuações, foi possível classificar os textos com maior incidência de estresse, em caso de um resultado negativo para a soma, relaxamento, no caso de uma soma com resultado positivo, ou neutro, caso a soma seja igual a zero. Os rótulos derivados do TSpt e das avaliações dos juízes foram comparados por meio do teste *chi-quadrado de Pearson* [Plackett 1983] e da *análise de correspondência*, com o auxílio do pacote *FactoMineR*[9].

Validada a versão do algoritmo para aplicação em textos na língua portuguesa, seguiu-se com a classificação das publicações em termos de estresse e relaxamento em toda a base de dados coletada. Uma vez realizada a classificação de cada publicação, observou-se a distribuição da proporção de estresse e relaxamento em cada período. Para descrever cada cenário, foram usados nuvens de palavras e algoritmos de PLN para a extração de tópicos.

4.4. Extração de Tópicos

Para a extração de tópicos foram comparados 3 algoritmos amplamente utilizados na literatura: LDA, NMF e BERTopic. Para a definição do número de tópicos utilizou-se a base de testes inicial de 30 dias extraída entre abril e maio de 2021. O algoritmo inicial, LDA, foi executado sobre essa base de dados, variando o número de tópicos entre 1 e 14, com uma coerência máxima verificada para 5 tópicos. Utilizou-se a mesma quantidade de tópicos entre os 3 algoritmos para efeito de comparação.

Para a quantidade de palavras em cada tópico, levou-se em consideração a literatura correlata, com 10 palavras ou menos por tópico [Ebeling et al. 2021], [Habibabadi and Haghighi 2019]. Optou-se, assim, por utilizar 6 palavras para descrever os tópicos extraídos. Os algoritmos selecionados foram comparados considerando a coerência NPMI (*Normalized Pointwise Mutual Information*) de cada algoritmo em cada um dos 7 períodos. Para a execução do NMF e BERTopic utilizou-se as bibliotecas *Scikit-Learn*[10] e *BERTopic*[11]. A execução do LDA e a verificação da coerência dos algoritmos foi possível graças a funções disponíveis na biblioteca *Gensim*[12].

[8]https://pypi.org/project/googletrans/

[9]http://factominer.free.fr/

[10]https://scikit-learn.org/stable/

[11]https://maartengr.github.io/BERTopic/index.html

[12]https://radimrehurek.com/gensim/auto$_e xamples/index.html$

5. Resultados

5.1. Validação do TSpt

Para avaliar a aplicação do TSpt em relação ao algoritmo original, retirou-se uma amostra de 386 *tweets* estratificada de acordo com a quantidade de publicações presentes em cada uma das 7 bases coletadas (Tabela 1). A amostra de publicações foi traduzida para o Inglês e submetida ao algoritmo TS da forma como foi fornecido pelos autores. Em paralelo foi feita a tradução do dicionário de palavras presente no código do TS para o Português e o algoritmo foi executado considerando as publicações da amostra sem tradução.

As pontuações de estresse e relaxamento das duas versões do TS foram comparadas considerando o *Erro Absoluto Médio* (EAM) e a *Correlação de Pearson*. De acordo com os autores [Thelwall 2017], essas duas métricas são mais adequadas por levarem em consideração o quanto um valor predito está distante de um valor de referência. As pontuações do TS, já validado na literatura, foram usadas como referência para avaliar o desempenho do TSpt. O resultado das métricas elencadas de comparação estão descritos na Tabela 2.

Verifica-se que as correlações apresentadas, significativas com 5% de significância, foram superiores às obtidas pelos autores para comparar o desempenho do TS em relação a codificadores humanos. No trabalho publicado por [Thelwall 2017], as correlações para estresse e relaxamento foram de 0.465 e 0.422, respectivamente. Os autores defendem, porém, que o EAM é a métrica mais adequada por assumir que a predição e o valor de referência estão na mesma direção. Na amostra selecionada ambos os erros médios absolutos foram menores que os utilizados pelos autores para validar o TS. Os erros verificados indicam que se espera, em média, uma diferença de ± 0.5881 na pontuação do TSpt em relação ao TS em se tratando de estresse. Já as pontuações de relaxamento apresentaram um desvio ainda menor, de ± 0.3005, indicando uma concordância ainda maior entre as duas versões do TS. Considera-se, portanto, que a modificação do TSpt foi próxima ao resultado que seria verificado ao utilizar o algoritmo original.

As pontuações de estresse e relaxamento do TSpt foram também comparadas com as pontuações fornecidas por juízes humanos, treinados a partir do manual fornecido por [Thelwall 2017], traduzido para o Português. Após a classificação, cada publicação foi rotulada de acordo com a soma das pontuações de estresse e relaxamento. O resultado do teste de chi-quadrado de Pearson pode ser visualizado na Figura 1. Verifica-se que, de acordo com *p-valor* obtido no teste, existe uma associação significativa entre os rótulos derivados do TSpt e os rótulos derivados das médias das avaliações dos juízes, com 5% de significância. Particularmente, o resultado da análise de correspondência na Figura 2 revela que o rótulo de estresse derivado do TSpt está mais associado aos rótulos de neutralidade e estresse derivados da avaliação dos juízes. Já os rótulos de relaxamento e neutro derivados do TSpt estão mais associados ao rótulo de relaxamento derivado da anotação dos juízes. A avaliação do estresse e do relaxamento presentes nas publicações foi realizada, portanto, considerando as pontuações obtidas por meio do TSpt.

5.2. Análise da incidência de estresse e extração de tópicos nos *tweets*

Na Figura 3 é possível verificar que, dentre os períodos selecionados, aquele com maior ocorrência de publicações ocorreu entre junho e julho de 2020, período em que ocorreu o pico de mortes no primeiro ano da pandemia. Nesse mesmo período, de acordo com a

LABEL_TS	*LABEL_JZ*			***Total***
	ESTRESSE_JZ	NEUTRO_JZ	RELAXAMENTO_JZ	
ESTRESSE_TS	99	23	10	132
NEUTRO_TS	121	20	32	173
RELAXAMENTO_TS	52	9	20	81
Total	272	52	62	386

χ^2=13.491 · df=4 · Cramer's V=0.132 · p=0.009

Figura 1. Teste de chi-quadrado: TSpt x Juízes

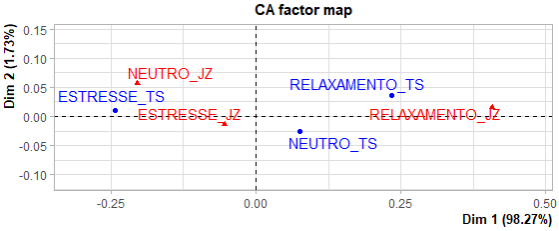

Figura 2. Análise de correspondência: TSpt x Juízes

Figura 4, observou-se a maior proporção de estresse, seguido dos períodos de fevereiro a março de 2020 e de junho a julho de 2021 com 50%, 49% e 49% das publicações rotuladas como estresse, respectivamente. Nota-se uma maior proporção de publicações rotuladas como estresse no início e nos picos de morte durante o primeiro e segundo ano de pandemia no Brasil. O período final selecionado, de abril a maio de 2022, foi o que demonstrou a menor ocorrência de publicações, momento em que ocorreu a flexibilização oficial do uso de máscaras no Brasil, além das menores taxas de óbito desde o início da pandemia.

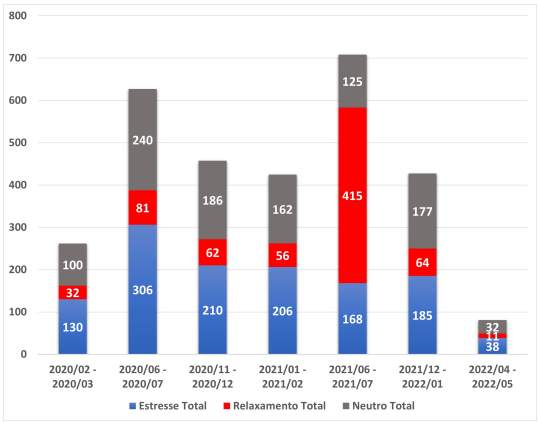

Figura 3. Total de *tweets* (milhares) por rótulo

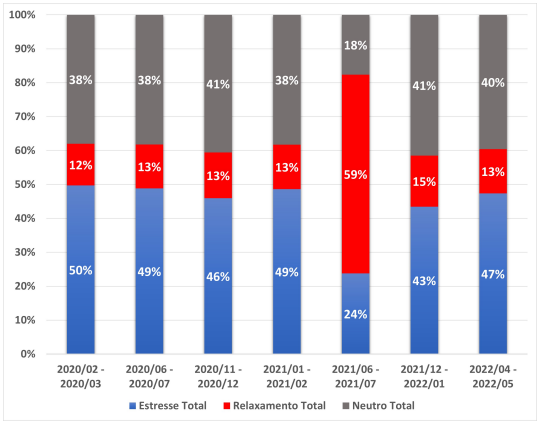

Figura 4. Proporção de *tweets* por rótulo

Para avaliar a extração de tópicos foram comparados os algoritmos LDA, NMF e BERTopic através da coerência NPMI [Bouma 2009]. Essa métrica de avaliação é uma variação da métrica PMI (*Pointwise Mutual Information*), que avalia a associação entre 2 termos. A NPMI é uma medida normalizada no intervalo $[-1, 1]$, em que os resultados -1, 0 e 1 indicam nenhuma coocorrência, independência e total coocorrência entre os termos, respectivamente [Campagnolo et al. 2022]. O resultado da coerência NPMI para os algoritmos selecionados em cada uma das bases de dados coletadas por ser verificado na Tabela 3. Devido ao maior valor médio da coerência NPMI relacionada ao NMF, optou-se por caracterizar os tópicos discutidos nos 4 períodos iniciais por meio deste algoritmo. Para os 3 períodos finais foram considerados os tópicos extraídos através do BERTopic. As Tabelas 4 e 5 apresentam exemplos de tópicos e termos associados a cada conjunto de *tweets* analisado.

Tabela 2. Comparação entre o TS e TSpt

Categoria	EAM	Correlação de *Pearson*
Estresse	0.5881	0.4950
Relaxamento	0.3005	0.6254

Tabela 3. Coerência NPMI dos algoritmos em cada base

Base	LDA	NMF	BERTopic
02/2020 - 03/2020	-0.0443	**0.0189**	-0.0552
06/2020 - 07/2020	0.0130	**0.0623**	0.0169
11/2020 - 12/2020	0.0022	**0.0400**	0.0148
01/2021 - 02/2021	-0.0382	**0.0809**	0.0301
06/2021 - 07/2021	-0.0079	0.0734	**0.0789**
12/2021 - 01/2022	-0.0189	0.0511	**0.0523**
04/2022 - 05/2022	-0.0585	0.0020	**0.0241**
Média	-0.0218	**0.0469**	0.0231

É possível notar que, no período inicial da pandemia no Brasil, houve discussões relacionadas à deflagração do coronavírus no país e à reação do governo brasileiro frente à pandemia. No período com maior pico de casos no primeiro ano discutiu-se sobre o acesso à vacinação e tratamentos defendidos como alternativos por parte do governo brasileiro. Os 3 períodos seguintes foram marcados por discussões relacionadas ao protocolo de distanciamento social, medidas de prevenção, testagem, vacinação e reverberações no cenário político brasileiro. Os períodos de dezembro de 2021 a janeiro de 2022 e de abril a maio de 2022 foram marcados por discussões ainda relacionadas à vacinação bem como por questões relativas a doenças com sintomas similares aos da COVID-19 e à perda de familiares no decorrer da pandemia por parte da população. No período final avaliado, em particular, nota-se a discussão sobre o protocolo de vacinação em crianças.

Tabela 4. Tópicos extraídos via NMF

Base	Tópico	Palavras
02/2020-03/2020	1	caso,confirmar,primeiro,número,suspeito,estado
02/2020-03/2020	2	bolsonaro,pegar,presidente,positivo,jair,comitiva
06/2020-07/2020	1	vacina,risos,teste,contra,achar,bom,tomar
06/2020-07/2020	2	bolsonaro,positivo,cloroquina,exame,sintoma,presidente
11/2020-12/2020	1	teste,negativo,positivo,resultado,maõsjuntas,amanhã
11/2020-12/2020	2	vacina,contra,tomar,querer,pfizer,bolsonaro
01/2021-02/2021	1	vacina,contra,tomar,dose,seringa
01/2021-02/2021	2	tratamento,precoce,existir,cloroquina,ivermectina,médico

Tabela 5. Tópicos extraídos via BERTopic

Base	Tópico	Palavras
06/2021-07/2021	1	teste,nariz,cotonete,horrível,exame,mãosjuntas
06/2021-07/2021	2	máscara,usar,distanciamento,pegar,bolsonaro,queiroga
12/2021-01/2022	1	dose,gripe,vacina,terceiro,influenza,reforço
12/2021-01/2022	2	família,mãe,criança,pai,irmã,perder
04/2022-05/2022	1	gripe,dor,garganta,febre,sinusite,alergia
04/2022-05/2022	2	dose,vacina,criança,gripe,vacina,terceiro

Para a representação das publicações por meio de nuvens de palavras foram selecionados os 3 períodos com maior ocorrência de publicações. As Figuras 5, 6 e 7 exibem as nuvens de palavras para a base de dados coletada entre junho e julho de 2020 com rótulos de estresse, relaxamento e neutro, respectivamente. A mesma sequência de nuvens de palavras pode ser visualizada nas Figuras 8, 9 e 10, para o período de novembro a dezembro de 2020, e nas Figuras 11, 12 e 13 para o período de dezembro de 2021 a janeiro de 2022.

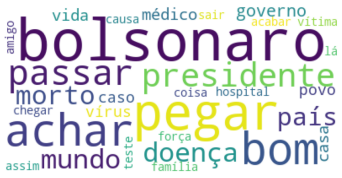

Figura 5. Estresse: 06/20-07/20

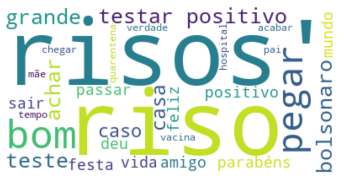

Figura 6. Relaxamento: 06/20-07/20

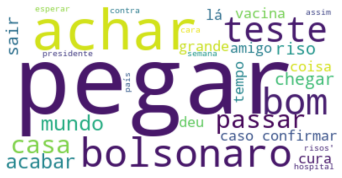

Figura 7. Neutro: 06/20-07/20

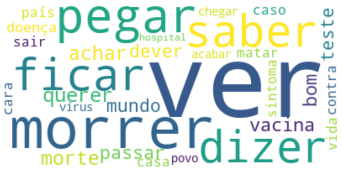

Figura 8. Estresse: 11/20-12/20

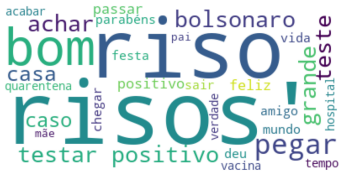

Figura 9. Relaxamento: 11/20-12/20

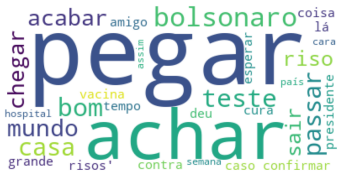

Figura 10. Neutro: 11/20-12/20

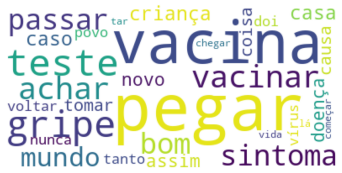

Figura 11. Estresse: 12/21-01/22

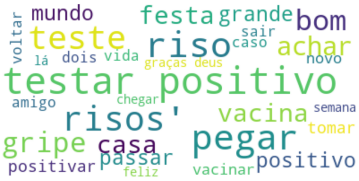

Figura 12. Relaxamento: 12/21-01/22

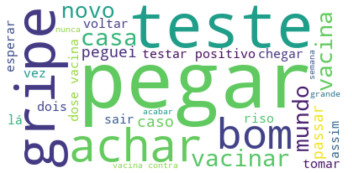

Figura 13. Neutro: 12/21-01/22

É possível notar que, de acordo com as nuvens de palavras geradas a partir dos *tweets* rotulados como estresse, houve uma reação da população a questões no cenário político brasileiro no período com maior pico de mortes no primeiro ano da pandemia. No período de novembro a dezembro de 2021 o estresse presente nas publicações relaciona-se com o diagnóstico e à morte pelo coronavírus. O estresse nas publicações realizadas entre dezembro de 2021 e janeiro de 2022 está relacionado à vacinação, testagem e ocorrência de doenças com sintomas similares aos da COVID-19. Os *tweets* rotulados como relaxamento apresentaram conteúdo relacionado ao então presidente brasileiro, testes positivos realizados e com a volta de eventos sociais. As publicações com o rótulo neutro, devido ao empate entre estresse e relaxamento, indicaram conteúdos relacionados a testes e ao contágio da COVID-19 e doenças similares.

6. Conclusões

A partir da busca de publicações realizadas no *Twitter*, identificadas por meio da palavra-chave *covid* e coletadas em diferentes momentos da pandemia do coronavírus, foi possível observar uma variação no volume de postagens capturadas. Dentre os períodos elencados para avaliação observou-se um pico de *tweets* entre junho e julho de 2020, momento em que $\frac{1}{3}$ dos óbitos ocorridos no primeiro ano de pandemia no Brasil foram registrados. Já o período final de coleta, de abril a maio de 2022, demonstrou o menor volume de publicações coletadas, com aproximadamente 13% do volume coletado entre junho e julho de 2020, refletindo o momento de flexibilização da pandemia no Brasil com a desobrigação do uso de máscaras e queda drástica no número de óbitos.

A comparação entre o TS, fornecido pelos autores, e o TSpt a partir do EAM e da *Correlação de Pearson* demonstrou correlações superiores e erros inferiores aos apresentados por [Thelwall 2017] no trabalho validado na literatura. Ao utilizar a avaliação

humana como referência para o TSpt, observou-se uma associação entre as classificações com predominância de estresse e com predominância de relaxamento entre o TSpt e as classificações dos juízes humanos. Os períodos com maior proporção de estresse aconteceram nos 2 primeiros meses da pandemia e durante os períodos com picos de mortes observadas no primeiro e no segundo ano de pandemia no Brasil.

Os tópicos extraídos por meio dos algoritmos NMF e BERTopic demonstraram o conteúdo publicado relacionado a: medidas de prevenção à COVID-19, reverberações no cenário político, testagem, vacinação, óbitos e ocorrência de doenças com sintomas parecidos com os apresentados em decorrência do contágio pelo coronavírus.

Como sugestão para trabalhos futuros seria importante avaliar o desempenho do TSpt na classificação do estresse presente nas sentenças em relação a algoritmos de classificação já consolidados na literatura. Além disso, uma amostra maior de publicações rotuladas por um número superior de juízes treinados para esse fim, a exemplo dos resultados apresentados por [Thelwall 2017], poderia fornecer mais evidências do desempenho do TSpt. Uma maior volumetria de rótulos gerados por juízes humanos, poderia ainda possibilitar a comparação do TSpt em relação a diversos algoritmos, bem como a avaliação dos escores gerados, ao invés da avaliação somente da tendência de estresse e relaxamento, obtida a partir da soma dos escores.

Referências

Abuzayed, A. and Al-Khalifa, H. (2021). Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia computer science*, 189:191–194.

Afonso, P. (2020). The impact of the covid-19 pandemic on mental health. *Acta medica portuguesa*, 33(5):356–357.

Alzamora, P. L., Locatelli, M. S., Ganem, M., Santos, T. H. M., Ferreira, D. V., Bernardes, T., Franco, R. A., Guiginski, J., Cunha, E. L., da Silva, A. P. C., et al. (2022). A covid-19 no twitter: correlacionando vocabulário com agravamento e atenuação da pandemia no brasil. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 157–168. SBC.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Brown, S. M., Doom, J. R., Lechuga-Peña, S., Watamura, S. E., and Koppels, T. (2020). Stress and parenting during the global covid-19 pandemic. *Child abuse & neglect*, 110:104699.

Campagnolo, J. M., Duarte, D., and Dal Bianco, G. (2022). Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management*, 13(4).

de Sousa, A. M. and Becker, K. (2021). Pro/anti-vaxxers in brazil: a temporal analysis of covid vaccination stance in twitter. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*, pages 105–112. SBC.

de Sousa, A. M. and Becker, K. (2022). Comparando os posicionamentos a favor/contra a vacinação covid nos estados unidos da américa e no brasil. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 65–77. SBC.

Ebeling, R., Sáenz, C. A. C., Nobre, J., and Becker, K. (2021). The effect of political polarization on social distance stances in the brazilian covid-19 scenario. *Journal of Information and Data Management*, 12(1).

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7.

Freitas, D. P., Borges, M. R., and Carvalho, P. V. R. d. (2020). A conceptual framework for developing solutions that organise social media information for emergency response teams. *Behaviour & Information Technology*, 39(3):360–378.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Habibabadi, S. K. and Haghighi, P. D. (2019). Topic modelling for identification of vaccine reactions in twitter. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–10.

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., Duan, W., Tsoi, K. K.-f., and Wang, F.-Y. (2020). Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on Computational Social Systems*, 7(2):556–562.

Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., and Feng, L. (2014). User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 507–516.

Ljajić, A., Prodanović, N., Medvecki, D., Bašaragin, B., and Mitrović, J. (2022). Uncovering the reasons behind covid-19 vaccine hesitancy in serbia: Sentiment-based topic modeling. *Journal of Medical Internet Research*, 24(11):e42261.

Martins, J. S., Lenz, M. L., and Silva, M. B. F. (2020). *PROCESSAMENTOS DE LINGUAGEM NATURAL*. Grupo A, Porto Alegre.

Menuzzo, V. A., Santanchè, A., and Gomes-Jr, L. (2021). Evaluating the cohesion of municipalities' discourse during the covid-19 pandemic. In *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 295–300. SBC.

Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International statistical review/revue internationale de statistique*, pages 59–72.

Taylor, S., Landry, C. A., Paluszek, M. M., Fergus, T. A., McKay, D., and Asmundson, G. J. (2020). Development and initial validation of the covid stress scales. *Journal of Anxiety Disorders*, 72:102232.

Thelwall, M. (2017). Tensistrength: Stress and relaxation magnitude detection for social media texts. *Information Processing & Management*, 53(1):106–121.

Wang, X., Zhang, H., Cao, L., and Feng, L. (2020). Leverage social media for personalized stress detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2710–2718.