JHI JOURNAL OF HEALTH INFORMATICS

CBIS'20 7 a 11 de Dezembro Foz do Iguaçú - Paraná Evento Virtual

COVID 19: O que sentem os brasileiros de acordo com o Twitter?

COVID 19: What do Brazilians feel according to Twitter?

COVID 19: ¿Qué sienten los brasileños según Twitter?

Giovanni Pazini Meneghel Paiva[1], Elisa Terumi Rubel Schneider[2], Josilaine Oliveira Cezar[3], Lucas Ferro Antunes de Oliveira[1], João Vitor Andrioli[4], Claudia Maria Cabral Moro Barra[5], Emerson Cabrera Paraiso[6], Lucas Emanuel Silva e Oliveira[7], Yohan Bonescki Gumiel[8]

RESUMO

Descritores: Processamento de Linguagem Natural; COVID-19; Mensagens do Twitter

Objetivo: A pandemia causada pelo novo coronavírus (SARS-CoV-2) caracteriza-se como o maior desafio do século 21. Neste contexto, procurou-se levantar um panorama geral de dados de usuários do Twitter, no Brasil, relacionados à COVID-19. **Métodos:** Utilizando de técnicas de Processamento de Linguagem Natural, foi aplicado um modelo Word2Vec CBOW em um conjunto pré-processado de dados públicos em português. Este foi então analisado através de *wordclouds*, tabelas e gráficos t-SNE. **Resultados:** O modelo captou comportamentos e tendências relacionados a COVID-19, como similaridades entre palavras, os unigramas e bigramas mais frequentes e hipóteses baseadas em dados estatísticos recolhidos. **Conclusão:** Este estudo apresenta uma análise inicial de mensagens do Twitter, em português, relacionadas à COVID-19. Os resultados foram promissores e evidenciaram o potencial da aplicação do aprendizado de máquina em assuntos importantes, como uma crise de saúde mundial.

ABSTRACT

Keywords: Natural Language Processing; COVID-19; Twitter messages

Objective: The pandemic caused by the new coronavirus (SARS-CoV-2) is characterized as the greatest challenge of the 21st century. In this context, an attempt was made to raise an overview of data from Twitter users in Brazil, related to COVID- 19. **Methods:** Using Natural Language Processing techniques, a Word2Vec CBOW model was applied to a pre-processed set of public data in Portuguese. Which was then analyzed using wordclouds, tables and t-SNE graphs. **Results:** The model captured behaviors and trends related to COVID-19, such as similarities between words, the most frequent unigrams and bigrams and hypotheses based on collected statistical data. **Conclusion:** This study presents an initial analysis of Twitter messages, in Portuguese, related to COVID-19. The results were promising and highlighted the potential of applying machine learning to important issues, such as a global health crisis.

RESUMEN

Descriptores: Procesamiento de Lenguaje Natural; COVID-19; Mensajes de Twitter

Objetivo: La pandemia causada por el nuevo coronavirus (SARS-CoV-2) se caracteriza como el mayor desafío del siglo XXI. En este contexto, se intentó levantar una visión general de los datos de los usuarios de Twitter en Brasil, relacionados con COVID. - 19. **Métodos:** Utilizando técnicas de procesamiento del lenguaje natural, se aplicó un modelo CBOW de Word2Vec a un conjunto de datos públicos preprocesado en portugués. Luego, esto se analizó mediante nubes de palabras, tablas y gráficos t-SNE. **Resultados:** El modelo capturó comportamientos y tendencias relacionados con COVID-19, como similitudes entre palabras, los *unigrams* y *bigrams* más frecuentes e hipótesis basadas en datos estadísticos recopilados. **Conclusión:** Este estudio presenta un análisis inicial de los mensajes de Twitter, en portugués, relacionados con COVID-19. Los resultados fueron prometedores y destacaron el potencial de aplicar el aprendizaje automático a problemas importantes, como una crisis de salud global.

[1] *Estudantes do curso de Engenharia da Computação, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[2] *Doutoranda do Programa de Pós-graduação em Informática, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[3] *Bibliotecária. Mestranda do Programa de Pós Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[4] *Mestrando do Programa de Pós-graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[5] *Professora do Programa de Pós-graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[6] *Professor do Programa de Pós-graduação em Informática, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[7] *Professor da Escola Politécnica, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*
[8] *Doutor em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná – PUCPR, Curitiba (PR), Brasil.*

Autor Coorespondente: **Giovanni Pazini Meneghel Paiva**
e-mail: **paivagio@hotmail.com**

INTRODUÇÃO

Nos últimos anos, as mídias sociais se tornaram fonte de informações sobre diversos temas, e a sua análise possibilita caracterizar eventos relacionados a determinado tópico. Segundo a Hootsuite[I], em 2019, 57% da população mundial tinha acesso à internet, e 45% eram usuários ativos de redes sociais. Um aumento de 9% de utilização em comparação com 2018, sendo no Brasil 8%, período em que a população aumentou 1,1%. O aumento da utilização das redes sociais e do poder de processamento computacional têm acelerado a extração automatizada de informações para monitoramento da saúde pública e para subsidiar decisões governamentais.

Em dezembro de 2019 foi relatado o início do surto de COVID-19 na China, causado pelo coronavírus, sendo declarado estado de emergência pela Organização Mundial da Saúde (OMS) devido à sua rápida disseminação. O portal CoronavírusBrasil[II], que integra dados das Secretarias Estaduais de Saúde, indica que atualmente temos quase 4,8 milhões de casos confirmados no Brasil, com mais de 144 mil óbitos acumulados e com uma taxa de letalidade de 3,0%. Durante a pandemia, o número de interações nas redes sociais aumentou significativamente. Como muitas atividades e trocas de mensagens relacionadas à pandemia ocorrem por meio de plataformas de mídia social, como Facebook, Twitter e Instagram, é essencial direcionar esforços na análise de conteúdo das redes sociais durante a propagação do vírus, extraindo informações valiosas sobre a crise de saúde e como a população está reagindo a ela. A análise dos dados das mídias sociais representa uma tendência[1], sendo uma fonte alternativa e valiosa para estudos relacionados à saúde.

Técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina vem sendo aplicadas à análise de dados de redes sociais, como na criação de modelos preditivos e na identificação de tendências relacionadas à doenças[2], buscando compreender e analisar linguagens humanas naturais através da linguagem computacional. O PLN, subcampo da Inteligência Artificial, sofreu uma evolução significativa nas últimas décadas, impactando em diversas áreas do conhecimento, como medicina, direito, física, ciências da informação, entre outras. Além disso ele pode ser aplicado para medir e monitorar sentimentos, como a hesitação sobre a aplicação de vacinas[3-4], para identificar comentários sobre uma doença específica (como a doença causada pelo vírus Zika), permitindo levantar dados sobre sintomas, níveis de transmissão, prevenção e tratamento[5], e ainda o sentimento negativo causado por epidemias[6], auxiliando na definição de políticas públicas e na abordagem de profissionais da saúde.

Segundo a HootSuite, o Twitter é uma das plataformas de rede social mais acessadas pela população mundial. Com alta popularidade, a mineração do conteúdo das mensagens do Twitter pode fornecer informações valiosas durante as crises de saúde. No Brasil, durante a pandemia, houve um aumento de 23% de usuários na plataforma.

A análise dos conteúdos de mídias sociais proporciona a compreensão das atitudes, sentimentos, pensamentos e comportamentos das pessoas, auxiliando na tomada de decisão dos profissionais de saúde e governantes. Neste sentido, o objetivo deste trabalho é identificar os tópicos relacionados à COVID-19 mais discutidos entre a população brasileira, empregando técnicas de PLN em *tweets* brasileiros.

Trabalhos Relacionados

Na área da saúde, há muitas pesquisas que propuseram aplicação de PLN e aprendizado de máquina, como Jianqiang, Xiaolin e Xuejun[7], que em 2018 apresentaram um método de *word embeddings* (GloVe-DCNN) obtidos por aprendizagem não supervisionada com base em grandes conjuntos de dados do Twitter, usando relações semânticas contextuais ocultas e estatísticas de co-ocorrência entre *tweets* e palavras.

Medford et al.[4] realizaram uma análise de dados do Twitter para entender o sentimento do público norte americano no início da pandemia. No período de 14 a 28 de janeiro de 2020, coletaram e avaliaram 126.049 *tweets*, concluindo que quase a metade (49,5%) da população expressava medo, e o impacto econômico e político estava entre os tópicos mais discutidos. Em contrapartida, foi possível identificar que, entre as postagens retuitadas, havia um esforço significativo na prevenção, na conscientização da necessidade de quarentena e informações sobre a transmissão da doença. Nesse contexto, Saleh et al.[8] estudaram a percepção pública sobre o distanciamento social identificando emoções e polaridade dos usuários do Twitter. Mostafa[9] corroborou ao analisar sentimentos e estado psicológico de estudantes egípcios durante o período de pandemia, usando Word2Vec[10].

Tian et al.[11] analisaram os retuítes de boatos que provocaram ansiedade e pânico em massa na população chinesa. Eles conseguiram distinguir informações reais de rumores/boatos retuitados para prever o comportamento dos humanos em relação às "*fake news*". Usaram Redes Neurais Convolucionais (R-CNN), com extração dos *tweets* usando K-means e incorporação de *embeddings* com Word2Vec.

Trabalhos envolvendo saúde também são encontrados para o português brasileiro (pt-br). Em 2018, Araújo et al.[12] apresentaram a metodologia de classificação de sentimento *Sentiment Descriptor Indexing* (SDI) que se baseia na coocorrência de termos do Twitter com descritores do vocabulário ANEW-BR, com foco em "câncer".

Pesquisadores do Centro de Telessaúde do HC-UFMG desenvolveram o Ana, um *chatbot*, para responder perguntas sobre a COVID-19 em pt-br, para além de orientar e informar, fazer uma triagem inicial de casos suspeitos. O Ana funciona em ambiente móvel/*web*, e utiliza árvore de decisão baseada em evidências[13]. Ainda no Brasil, Xavier et al.[14] utilizaram técnicas de PLN para analisar 7,7 milhões de postagens em pt-br no Twitter sobre COVID-19, apoiando na vigilância em saúde.

Existem também estudos relacionados ao uso de produtos não aprovados pelos órgãos competentes sendo vendidos em redes sociais[15], uso de medicamentos sem prescrição médica[16], e de identificação de reações adversas

[I] https://p.widencdn.net/kqy7ii/Digital2019-Report-en
[II] https://covid.saude.gov.br/

aos medicamentos, que utilizam ferramentas de PLN como o Word2Vec, FastText, arquitetura HAN e CNN[17].

MÉTODOS

Para este estudo foi treinado um modelo Word2Vec usando mensagens em português recolhidas do Twitter, relacionadas à COVID-19, entre o período de janeiro a maio de 2020. O estudo foi divido em 4 etapas: (1) extração de mensagens do Twitter, (2) pré-processamento dos textos, (3) treinamento do modelo Word2Vec e (4) análise do modelo obtido, como mostra a Figura 1.

Extração de textos

Para a extração de textos foi utilizado um *dataset* disponibilizado por Melo[18], contendo IDs de mensagens em português de usuários do Twitter sobre a COVID-19[III]. Os *tweets* foram coletados pela biblioteca *Twitterscraper*, filtrando por um conjunto de palavras-chave associadas com a atual pandemia, como quarentena, coronavírus, isolamento, *lockdown* e pandemia. Com a biblioteca *Tweepy*[IV], que utiliza a API do Twitter para extrair informações, foram resgatados e salvos 1.309.017 *tweets*. Todos os *tweets* foram extraídos seguindo a Política de Desenvolvedor do Twitter[V] e as normas previstas na Lei Geral de Proteção de Dados Pessoais (LGPD)[VI], assegurando e preservando a privacidade dos autores das mensagens.

Pré-processamento

Após a extração do texto, os *tweets* foram agrupados em um documento e armazenados em um *dataframe*. Em seguida, foi eliminada do conjunto de dados qualquer acentuação especial, utilizando-se de expressões regulares. Foi realizada a limpeza dos dados inicialmente com a biblioteca *BeautifulSoup*[VII], extraindo a formatação html/xml dos *tweets*, e após esta etapa, foram retirados os caracteres não alfanuméricos, como urls, menção a outros usuários, *emojis* e outros caracteres recolhidos durante o processo de extração de dados. No final, o resultado do pré-processamento foi salvo em formato texto para fácil manipulação.

Treinamento do modelo word2vec

Word2Vec é um modelo de incorporação de palavras, ou seja, um modelo de representação numérica de palavras que captura o significado semântico de cada palavra e as relações entre elas[10]. Os vetores numéricos são gerados através de um treinamento semi-supervisionado com redes neurais, podendo ser treinados com a arquitetura CBOW (Modelo Contínuo de *Bag of Words*) ou Skip-gram. Na arquitetura CBOW, o modelo é treinado para prever a palavra pelo seu contexto, ou seja, maximizar a probabilidade da palavra-alvo examinando as palavras vizinhas. Na arquitetura Skip-gram, o modelo é treinado para prever o contexto dada uma palavra.

Para fazer o treinamento do modelo Word2Vec,

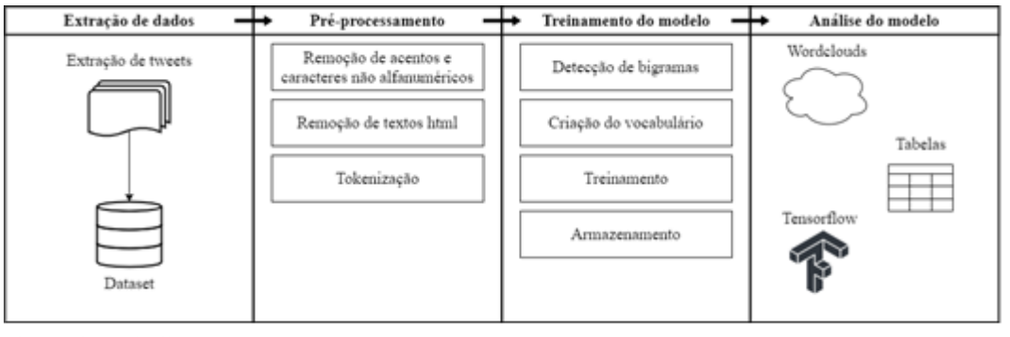
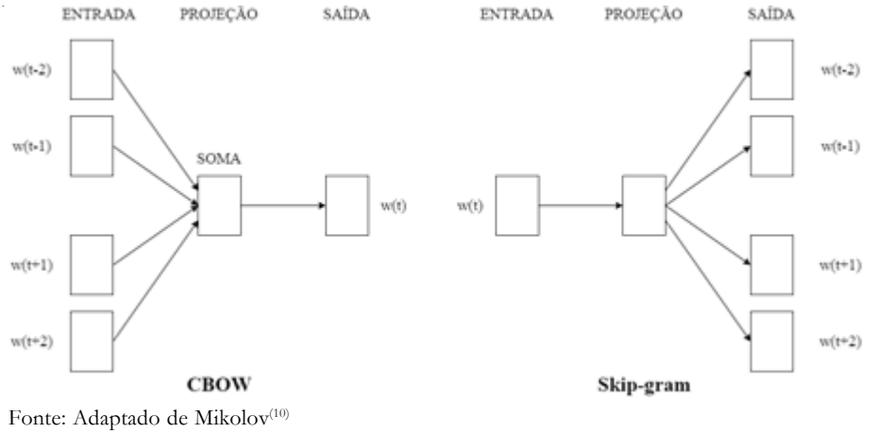


Figura 1 - Fluxo de trabalho



Fonte: Adaptado de Mikolov[10]

Figura 2 - Word2Vec: Arquiteturas CBOW vs Skip-gram

[III] https://data.mendeley.com/datasets/vhxdgjfjnk/3
[IV] https://www.tweepy.org/
[V] https://developer.twitter.com/en/developer-terms/agreement-and-policy
[VI] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm
[VII] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

primeiramente carregou-se o corpus para um *dataframe* para facilitar o processo de treinamento e, utilizando a ferramenta *Phrases*[VIII] da biblioteca *Gensim*, detectaram-se bigramas presentes nos dados. A seguir, com o modelo "pt_core_news_sm" presente na biblioteca *Spacy*[IX], foi construído o vocabulário e treinado um modelo Word2vec, com arquitetura CBOW, configurada para 30 épocas. O modelo gerado foi armazenado no formato binário.

Análise do modelo

A análise do treinamento foi feita utilizando de ferramentas gráficas, como *wordclouds* e gráficos t-SNE, para visualização dos *embeddings* e tabelas comparativas para avaliar a frequência de unigramas e bigramas. Os *wordclouds* foram gerados usando a biblioteca *wordcloud*[X], que recebeu como entrada um dicionário contendo o vocabulário do modelo normalizado com a frequência de cada palavra. Foram retiradas do *wordcloud* preposições, artigos e outras palavras não significativas (como as *stop-words*). As visualizações t-SNE, por sua vez, foram criadas a partir do modelo de der Maaten[XI].

RESULTADOS E DISCUSSÃO

Como em um modelo Word2Vec cada palavra é representada como um vetor numérico, é possível medir a similaridade entre as palavras em um espaço vetorial esparso de alta dimensão, ou seja, capturar o significado e seu relacionamento com outras palavras. Assim, com o modelo treinado, exploramos os principais tópicos relacionados à pandemia COVID-19, como *coronavirus*, *covid*, *pandemia*, *cloroquina*, *virus e lockdown*. Conforme ilustrado na Figura 3, o modelo treinado foi capaz de capturar o significado semântico desses tópicos.

A Figura 4 (a) ilustra os principais tópicos relacionados aos medicamentos para COVID-19 de acordo com as mensagens analisadas, evidenciando a relação entre *remedio, (hidroxi)cloroquina, placebo, tamiflu, vermifugo, retroviral* e outros. O ponto vermelho corresponde a palavra em análise, neste caso *remedio*, os pontos verdes e cianos correspondem, respectivamente, às 1ª - 10ª e 11ª - 20ª entidades mais próximas a entidade vermelha.

Na Figura 4 (b), é possível observar as palavras mais relacionadas a um dos sintomas da COVID-19, *tosse,* associando expressões como *tosse seca*, *rinite*, *febre*, *dor* e *enxaqueca.*

Com a análise da frequência de unigramas e bigramas, gerados pelo conjunto de dados pré-processados, é possível perceber as principais tendências e preocupações da população durante a pandemia.

Analisamos a frequência de cada unigrama e bigrama

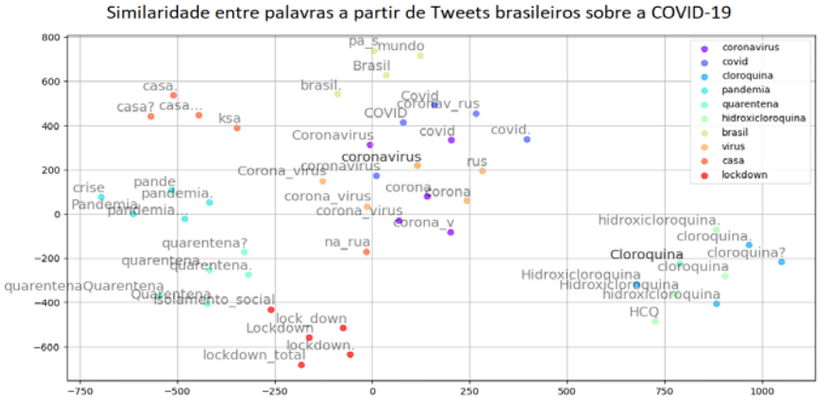


Figura 3 - Similaridade entre os 10 principais tópicos relacionados à COVID-19 de acordo com o modelo gerado.

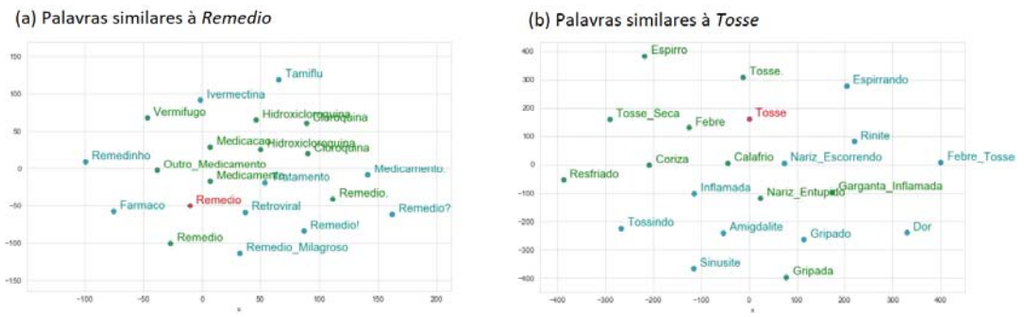


Figura 4 - Similaridade entre os principais tópicos relacionados a remédios para COVID-19 e à tosse, principal sintoma da doença.

[VIII] https://radimrehurek.com/gensim/models/phrases.html
[IX] https://spacy.io/
[X] https://amueller.github.io/word_cloud/references.html
[XI] https://lvdmaaten.github.io/tsne/#implementations

Tabela 1 - Top 10 unigramas e bigramas

Unigrama	Frequência	%	Bigrama	Frequência	%
coronavirus	3.052.569	0,13%	corona_virus	2.027.098	0,27%
covid	2.748.231	0,11%	isolamento_social	1.416.029	0,19%
cloroquina	2.373.050	0,10%	distanciamento_social	1.004.048	0,13%
pandemia	2.352.537	0,10%	fake_news	710.353	0,09%
quarentena	1.834.186	0,08%	nessa_quarentena	619.991	0,08%
hidroxicloroquina	1.576.251	0,06%	em_casa	583.618	0,08%
brasil	1.414.767	0,06%	meu_deus	577.916	0,08%
virus	1.301.652	0,05%	por_favor	563.490	0,07%
casa	1.256.907	0,05%	boa_noite	491.977	0,06%
lockdown	1.193.182	0,05%	sao_paulo	485.614	0,06%

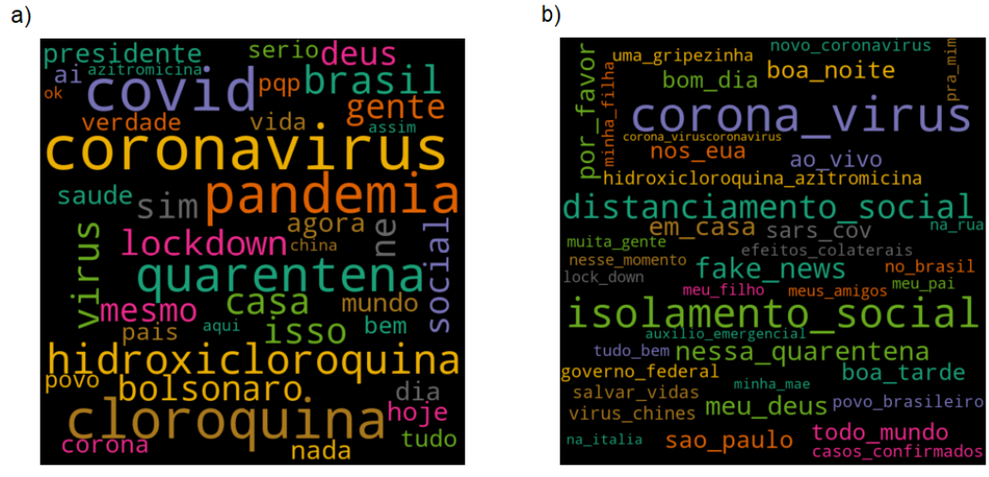


Figura 5 - *Top* 40 unigramas e bigramas, respectivamente

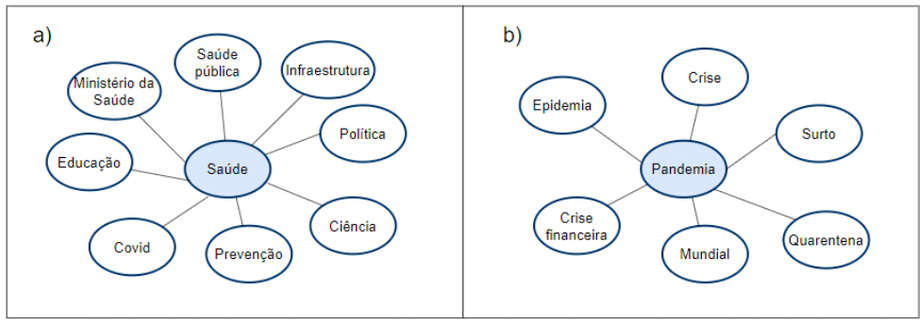


Figura 6 - Palavras próximas a *saúde* e *pandemia*

ao longo de todo conjunto de dados (Tabela 1) e geramos a nuvem de palavras contendo os 40 principais unigramas (a) e bigramas (b), conforme a Figura 5. As palavras unigramas mais frequentes no período foram *coronavírus, covid*, *pandemia*, *quarentena* e *cloroquina*. As palavras bigramas mais frequentes foram *corona vírus, distanciamento social* e *isolamento social*, o que pode indicar a preocupação da população pelo distanciamento tendo em vista o grau de propagação do vírus. É interessante notar que também são detectados, mesmo que em menor escala, os termos *governo federal*, *auxílio emergencial* e até *fake news*.

A Figura 6 mostra as palavras mais próximas aos termos *Saúde* (a) e *Pandemia* (b), de acordo com o modelo treinado. É interessante notar como a população relaciona a *Saúde* com o *Ministério da Saúde*, *Política* e *Saúde Pública*. Também encontramos relação com *Educação*, *Ciência* e *Prevenção*. Outra relação interessante é o termo *Pandemia* com *Crise* e *Crise financeira*.

Os dados levantados nesta pesquisa permitem sugerir hipóteses e tendências de comportamento através de evidências numéricas e estatísticas - não sendo possível, contudo, afirmar com certeza tais afirmações. A seguir, uma visão geral dos principais tópicos discutidos no Twitter, em relação à COVID-19, evidenciados neste estudo:

- A população entende que a pandemia é global e afeta a todos. Os principais países citados foram China, Estados Unidos, Itália, Espanha, Alemanha e Coreia.

- Isolamento: expressões como "*fica em casa*", "*distanciamento social*", "*stay home*", "*quarentenou*" e "*estamos juntos*" mostram o entendimento dos usuários da importância do isolamento.

- Os brasileiros entendem que a COVID-19 é causada por um vírus e que é proveniente da China, embora as palavras *bactéria*, *micróbios*, *peste* e *desgraça* também ocorram com certa frequência.

- A cura da doença, bem como a vacinação, são temas bastante abordados pelos usuários. *Cloroquina* e palavras relacionadas a este medicamento tiveram grande número de ocorrências, e vale uma análise de sentimentos para entender qual o sentimento prevalente sobre esta droga.

- Os principais sintomas relatados são tosse, espirro, coriza, febre, dor, enxaqueca, alergia, pneumonia, enjôo, diarréia e nenhum sintoma. Vale destacar que os sintomas não estão, necessariamente, relacionados à COVID-19.

- Na percepção dos usuários do Twitter, a pandemia também está relacionada a uma crise financeira, que de alguma maneira está ligada ao governo, Ministério da Saúde, saúde pública e política.

É importante considerar os riscos envolvendo a pesquisa, pois utilizou-se um *dataset* com informações secundárias, não sendo possível averiguar a procedência e consistência dos dados, o que pode expor a pesquisa a resultados pretensiosos. Além disso, a coleta dos *tweets* corresponde ao período anterior ao pico da pandemia, que ocorreu em maio de 2020, não refletindo assim, o atual comportamento e percepção da população.

CONCLUSÃO

O presente trabalho apresenta uma análise inicial de mensagens do Twitter postadas por usuários brasileiros durante a pandemia de COVID-19, entre janeiro a maio de 2020, tendo por objetivo analisar o cenário iminente ao pico da pandemia. Através do treinamento de um modelo Word2vec, foi possível explorar os tópicos mais comentados sobre a pandemia e extrair informações relevantes sobre o sentimento da população em geral. As incorporações de palavras (w*ord embeddings*) geradas pelo modelo permitiram analisar o significado e o relacionamento entre algumas palavras importantes, como medicamentos, sintomas e isolamento social, através da medição da similaridade entre os vetores de representação. A extração dos unigramas e bigramas mais frequentes permitiu uma análise das principais preocupações e pensamentos da população brasileira que utilizou o Twitter no período analisado. Esta pesquisa evidencia o potencial dos modelos de aprendizagem de máquina aplicados em um grande conjunto de dados na análise de assuntos importantes, como uma crise de saúde mundial. O modelo gerado e os códigos de acesso estão disponíveis publicamente[XII]. Como trabalhos futuros, pretendemos trabalhar com dados de todo o ano de 2020, realizando uma análise de sentimentos da população mês a mês e comparando com o aumento ou diminuição do número de casos da doença no país, bem como um comparativo entre o início e fim do ano.

AGRADECIMENTOS

À CAPES. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. À Fundação Araucária, pelo apoio financeiro. Aos pesquisadores do trabalho de Melo[18] pela disponibilização do *dataset* inicial dos *tweets* em português.

REFERÊNCIAS

1. Denecke K, Nejdl W. How valuable is medical social media data? Content analysis of the medical web. Information Sciences. 2009 May;179(12):1870-1880. doi: 10.1016/j.ins.2009.01.025. Available from: https://www.researchgate.net/publication/223887750_How_valuable_is_medical_social_media_data_Content_analysis_of_the_medical_web
2. Dai X, Bikdash M, Meyer B. From social media to public health surveillance: word embedding based clustering method for twitter classification. Proceedings SoutheastCon; 2017 30 March-2 April; Charlotte, NC: IEEE; 1017: 1-7. doi: 10.1109 / SECON.2017.7925400. Available from: https://ieeexplore.ieee.org/document/7925400
3. Bahk CY, Cumming M, Paushter L, Madoff LC, Thomson A, Brownstein JS. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. Health Aff (Millwood). 2016 Feb;35(2):341-7. doi: 10.1377/hlthaff.2015.1092. Available from: https://pubmed.ncbi.nlm.nih.gov/26858390/
4. Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU. An "Infodemic": leveraging high-volume Twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak. Open Forum Infectious Diseases; 2020 Jul;7(7). doi: https://doi.org/10.1093/ofid/ofaa258. Available from: https://academic.oup.com/ofid/article/7/7/ofaa258/5865318
5. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What Are people Tweeting about Zika? An exploratory study concerning Its symptoms, treatment, transmission, and prevention. JMIR Public Health Surveill. 2017 Jun 19;3(2):e38. doi: 10.2196/publichealth.7157. Available from: https://pubmed.ncbi.nlm.nih.gov/28630032/
6. Mamidi R, Miller M, Banerjee T, Romine W, Sheth A. Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 Zika epidemic [internet]. JMIR Public Health Surveill. 2019 Jun 4;5(2):e11036. doi: 10.2196/11036. Available from: https://pubmed.ncbi.nlm.nih.gov/31165711/
7. Jianqiang Z, Xiaolin G, Xuejun Z. Deep Convolution Neural Networks for Twitter sentiment analysis [Internet]. IEEE Access. 2018 Jan 01; 6:23253-23260. doii: 10.1109/ACCESS.2017.2776930. Available from: https://ieeexplore.ieee.org/document/8244338
8. Saleh SN, Lehmann CU, McDonald SA, Basit MA, Medford RJ. Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter [internet]. Infect Control Hosp Epidemiol. 2020 Aug;6(1-8). doi: 10.1017/ice.2020.406. Available from: https://pubmed.ncbi.nlm.nih.gov/32758315/
9. Mostafa L. Egyptian student sentiment analysis using Word2vec during the Coronavirus (Covid-19) pandemic. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020 - AISI 2020. 2020 Set 20. Advances in Intelligent Systems and Computing: Springer. doi: https://doi.org/10.1007/978-3-030-58669-0_18. Available from: https://link.springer.com/chapter/10.1007/978-3-030-58669-0_18
10. Mikolov T, Chen K, Corrado G, Deanet J. Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations - ICLR 2013. 2013 Set 7. Available from: https://arxiv.org/abs/1301.3781

[XII] https://github.com/HAILab-PUCPR/Word2Vec-COVID19-Twitter

11. Tian Y, Fan R, Ding X, Zhang X, Gan T. Predicting rumor retweeting behavior of social media users in public emergencies. IEEE Access. 2020 Abr 21; 8:87121-87132. doi: 10.1109/ACCESS.2020.2989180. Available from: https://ieeexplore.ieee.org/abstract/document/9075170
12. Araujo GD, Teixeira FO, Mancini F, Guimarães MP, Pisa IT. Sentiment analysis of Twitter's Health messages in Brazilian Portuguese. 2018 Jan-Mar; 10(1):17-24. Available from: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/566
13. Ferreira TC, Marcolino MS, Ramos I, Prates R, Ribeiro LB, Reis Z et al. ANA: a brasilian chatbot assistant about COVID-19. Proceedings of Conference ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID). 2020 Jun; Seattle, Washington, United States. Available from: https://www.researchgate.net/publication/342869406_Ana_a_brazilian_chatbot_assistant_about_covid_19
14. Xavier F, Olenscki JRW, Acosta AL, Sallum MAM, Saraiva AM. Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a Covid-19. Estudos avançados. 2020 May 10;34(99). doi: https://doi.org/10.1590/s0103-4014.2020.3499.016. Available from: https://www.scielo.br/scielo.php?pid=S0103-40142020000200261&script=sci_arttext
15. Mackey TK, Li J, Purushothaman V, Nali M, Shah N, Bardier C et al. Big Data, Natural Language Processing, and Deep Learning to detect and characterize illicit COVID-19 product sales: infoveillance study on Twitter and Instagram. JMIR Public Health Surveill. 2020 Aug 25;6(3):e20794. doi: 10.2196/20794. Available from: https://publichealth.jmir.org/2020/3/e20794/
16. O'Connor K, Sarker A, Perrone J, Gonzalez Hernandez G. Promoting reproducible research for characterizing nonmedical use of medications through data annotation: description of a Twitter corpus and guidelines. J Med Internet Res. 2020 Feb 26;22(2):e15861. doi: 10.2196/15861. Available from: https://pubmed.ncbi.nlm.nih.gov/32130117/
17. Rezaei Z, Ebrahimpour-Komleh H, Eslami B, Chavoshinejad R, Totonchi M. Adverse drug reaction detection in social media by deep learning methods. Cell J. 2020 Oct;22(3):319-324. doi: 10.22074/cellj.2020.6615. Available from: https://pubmed.ncbi.nlm.nih.gov/31863657/
18. Melo T, Figueiredo CMS. A first public dataset from brazilian twitter and news on COVID-19 in portuguese.Data in brief. 2020 Oct;32:106179. doi: https://doi.org/10.1016/j.dib.2020.106179. Available from: https://www.sciencedirect.com/science/article/pii/S2352340920310738.