Análise das Publicações no Twitter Sobre as Vacinas Contra a COVID-19 no Brasil

Douglas Almeida Vidal[1], Adriano Madureira[1], Harold Junior[2], Karla Figueiredo[2], Lucas Mendonça[1], Rita Paulino[3], Yomara Pires[1], Marcos César da Rocha Seruffo[1]

[1]Universidade Federal do Pará

[2]Universidade do Estado do Rio de Janeiro

[3]Universidade Federal de Santa Catarina

{vidalstm998, adrianomadureira, harold.dias, erlucomlpg, rcpauli}@gmail.com, karlafigueiredo@ime.uerj.br, seruffo@ufpa.br

Abstract: *At the beginning of 2020, the world lived with a crisis caused by the emergence of the Corona Virus disease (COVID-19). This pandemic was devastating for several countries, but the impacts suffered and the measures taken to face this crisis distinguished each nation. Among the most effective procedures to combat the disease, the vaccination has become the prevention and control tool during the pandemic. At the same time, Online Social Networks (RS) played an important civic and political role, remaining among the most used sources of news and information. This paper presents an analysis of posts by Brazilians and the president of Brazil on the Twitter platform about vaccines against COVID-19, carried out from August 2020 to March 2021. Machine Learning techniques were used, and the best results showed that the Support Vector Machine (SVM) was the best-performing model. The classification of vaccines that users and the president most mentioned in tweets reached 60.72 % accuracy after selection with ReliefF.*

Resumo: *No início de 2020, o mundo conviveu com uma crise causada pelo surgimento da doença do Corona Vírus (COVID-19). Esta pandemia foi devastadora para diversos países, mas os impactos sofridos e as medidas realizadas para enfrentar esta crise distinguiu cada nação. Entre as medidas mais eficazes para o enfretamento da doença, a vacinação tornou-se a principal ferramenta de prevenção e controle durante a pandemia. Ao mesmo tempo, as Redes Sociais Online (RSO) exerceram um importante papel cívico e político, estando entre as fontes de notícia e informações mais utilizadas no mundo. Este artigo apresenta uma análise das publicações de brasileiros, incluindo o presidente do Brasil na plataforma Twitter sobre as vacinas contra a COVID-19, realizadas no período de agosto de 2020 a março de 2021. Técnicas de Aprendizado de Máquina foram usadas e os resultados mostraram que a Máquina de Vetores-Suporte (Support Vector Machine - SVM) foi o modelo com melhor desempenho. A classificação das vacinas que os usuários e o presidente mais citaram nos tweets alcançou 60,72% de acertos após seleção com ReliefF.*

Palavras-chaves: Covid-19, Aprendizado de Máquina, Redes Sociais Online, Support Vector Machine e Twitter.

e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2023.70419

1. Introdução

Em 2020, foi vivenciada uma situação de calamidade pública e de emergência em saúde com a propagação do novo Corona vírus (COVID-19). Uma catástrofe de escala mundial ocorrida em ondas sucessivas, com variantes distintas do vírus, impondo desafios adicionais ~~impostos~~ palas restrições de circulação globais e saúde econômica dos países. Mundialmente, em 2020, a propagação do vírus gerou 1,8 milhão de mortes[1] e embates por parte dos governantes para estabelecer regras próprias para combater a doença.

Neste cenário, as Redes Sociais Online (RSO) se tornaram um espaço significativo para atividade cívica e política, além de criarem oportunidades para governantes influenciarem as opiniões dos seus públicos [Malinen et al., 2020]. As RSO estão entre as fontes de informação mais utilizadas no mundo. O acesso fácil e econômico à Internet e o elevado número de usuários popularizaram rapidamente estas plataformas, tornando-as uma das formas mais fáceis e eficazes de divulgar a informação. Durante grandes eventos, a resposta geral é uma busca maior por informações, seja um evento esportivo, uma doença ou um desastre natural [Gonzalez-Padilla e Tortolero-Blanco 2020].

De acordo com a BBC News[2], com o aumento da COVID-19, as reações de combate dos países frente a esta pandemia ocorreram de formas distintas, os países asiáticos são os mais bem sucedidos quanto à contenção da pandemia, seguidos pelas Nações do Oriente Médio e África. O Brasil apresenta um cenário peculiar. Originalmente o governo apostou em medicamentos sem comprovação científica, agora vem enfrentando desafios no que se refere à gestão econômica, de saúde e cooperação internacional. Segundo um estudo[3], o Brasil está entre os piores países no que se refere ao enfrentamento à pandemia. O cenário brasileiro tem sido marcado por inúmeros conflitos e por ações descoordenadas tanto na esfera política quanto na de saúde (Glezer, 2021)

Nesta pesquisa, foram coletados e processados dados dos perfis de cidadãos brasileiros e do presidente do Brasil, na época, com o objetivo de classificar as publicações de acordo com o tipo de vacina mencionada por estes usuários durante a pandemia e encontrar o método mais adequado para este fim, segundo a opinião desses usuários. O acompanhamento foi feito entre os meses de agosto de 2020 a março de 2021, a partir de publicações de usuários em contas próprias na plataforma Twitter e também na conta oficial do chefe de Estado brasileiro.

Assim, este trabalho buscou desenvolver e aplicar uma metodologia de coleta e classificação de *tweets*, utilizando técnicas de Aprendizado de Máquina, aplicando duas propostas e algoritmos para modelos preditivos, por meio das mensagens durante a pandemia,

[1] Dados retirados de: https://brasil.elpais.com/sociedad/2020-12-31/em-2020-18-milhao-de-vidas-levadas-pela-COVID-19-em-2021-a-esperanca-da-vacina.html e: https://news.google.com/COVID19/map?hl=pt-BR&gl=BR&ceid=BR%3Apt-419

[2] https://www.bbc.com/portuguese/brasil-55870630

[3] https://interactives.lowyinstitute.org/features/COVID-performance/

com o intuito de verificar quais foram as vacinas de combate à COVID-19 mais comentadas pela população e pelo presidente durante o período da coleta.

Para este fim, o artigo é estruturado da seguinte forma: a seção 2 fornece uma visão geral de estudos de RSO e de COVID-19. A seção 3 apresenta as duas abordagens de coleta e pré-processamento de dados investigadas, além do uso de algoritmos de classificação. Na seção 4 são apresentados os resultados de classificação das mensagens com as duas abordagens. Por fim, na seção final são apresentadas as conclusões e perspectivas de novos trabalhos.

2. Trabalhos correlacionados

Estudos atuais consideram as RSO como fonte estratégica de apoio à vigilância em saúde durante a COVID-19. Segundo Xavier *et al.* (2020, p. 261), a desinformação é o grande gargalo na comunicação, visto que há um grande esforço dos órgãos de saúde nas atividades de comunicação com a população e combate às notícias falsas que são publicadas em diversos meios e disseminadas especialmente via Internet.

Além disso, Malavé (2020) afirma que no período da pandemia, durante o isolamento social, recomendado pela OMS, o uso das RSO foi potencializado, constituindo o principal canal de comunicação entre os que permaneceram em casa, uma vez que, para o indivíduo é essencial se comunicar e ter o contato com o mundo.

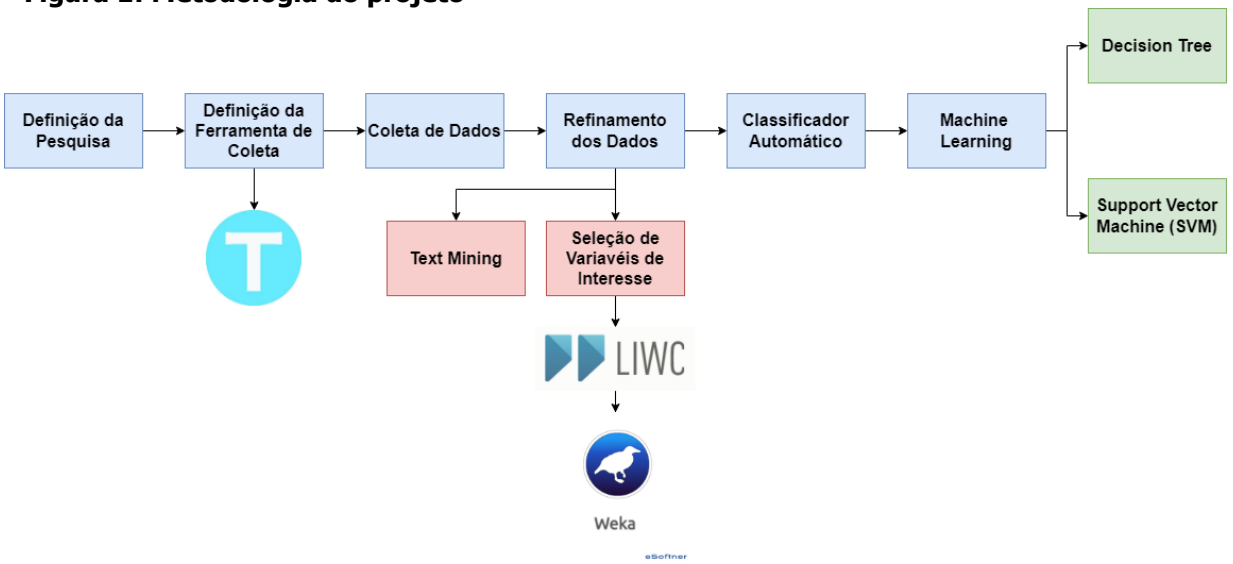
Neste contexto, Dodds *et al.* (2019) dizem que as comunicações presidenciais são um tópico oportuno, visto que dinamicamente evidenciam como o chefe de estado se comunica com o público. De forma geral, há muitos trabalhos que envolvem técnicas de Aprendizado de Máquina e RSO. Em Kaur (2020), foi realizada a análise de sentimento em relação à doença do coronavírus (COVID-19). Bokang *et al.* (2021) usaram uma abordagem combinada de Aprendizado de Máquina para melhorar os detectores automáticos de racismo e discriminação, relacionando os efeitos da COVID-19 sobre as atitudes dos usuários do Twitter, classificando os *tweets* racistas antes e depois da doença ser declarada como pandemia global.

Tomando por base estas referências, a seção seguinte apresenta as técnicas utilizadas para coleta, detecção e classificação automática das publicações no Twitter, com a finalidade de identificar as vacinas contra COVID-19 comentadas nas publicações de brasileiros e do presidente do Brasil.

3. Procedimentos metodológicos

Uma metodologia científica foi seguida, com o intuito de tornar o processo de desenvolvimento da proposta factível de reprodutibilidade.

Figura 1: Metodologia do projeto



A Figura 1 mostra a metodologia utilizada para o desenvolvimento deste trabalho. Para isto foram seguidas duas etapas, sendo estas: 1 - Coleta e Processamento dos Dados; 2 - Aplicação de técnicas de aprendizado de máquina para classificação das vacinas disponibilizadas no Brasil; com a finalidade de responder a seguinte questão-chave de pesquisa: QP- **Qual(is) (são) a(s) vacina(s) mais mencionada(s) nas publicações no Brasil?**

3.1. Coleta e Processamento dos Dados

Nesta etapa, foi utilizada a aplicação *Twint*, escrita em Python, que realiza a extração de *tweets* nos perfis. Este software realiza a coleta baseada em *tweets* de usuários específicos ou *tweets* relacionados a certos tópicos, *hashtags* e tendências. Assim, foi possível obter uma base com 121.380 *tweets* de 85.450 perfis do *Twitter*, sendo 121.362 *tweets* de brasileiros e 18 do presidente Jair Bolsonaro, a coleta dos dados ocorreram no período de agosto de 2020 a março de 2021.

Com a finalidade de analisar o conteúdo específico sobre as vacinas, foram escolhidos termos que representam os imunizantes autorizados e/ou aprovados pela ANVISA (Agência de Vigilância Sanitária) durante o período da coleta dos dados. Desta forma dos 121.362 *tweets* de brasileiros, tem-se: 4.355 para AstraZeneca, 16.229 para Moderna, 10.270 para Pfizer, 12.731 para Sputnik e 77.777 para CoronaVac. Para os *tweets* do presidente: 9 para AstraZeneca e 9 para CoronaVac, totalizando somente 18 *tweets*, mesmo tendo poucas postagens, as publicações possui mais impacto por expressar o posicionamento do chefe de estado sobre a questão.

3.2. Aplicação de técnicas de aprendizado de máquina para classificação

Nesta etapa modelos preditivos foram empregados para classificar os *tweets* nos cinco tipos de vacinas, permitindo que, em seguida, pudesse ser feita análise das publicações, de usuários

brasileiros e do presidente do Brasil, sobre vacinas contra a COVID-19. Especificamente, para responder a questão de pesquisa QP, é necessário: a) extrair e selecionar os atributos relevantes para a predição; b) treinamento e avaliação dos modelos de classificação. Duas propostas foram desenvolvidas para classificação dos *tweets*, com conteúdo das publicações disseminadas pelo presidente e pela população, relativas aos tipos de vacinas. A primeira utilizou a Ferramenta LIWC (*Linguistic Inquiry and Word Count*), que analisa textos a partir da contagem de palavras em categorias significativas fixadas, utilizando diversos dicionários, e a segunda baseada em *Text Mining* sobre o texto dos *tweets*.

Na primeira proposta, a ferramenta LIWC reuniu um conjunto de atributos, como afeto (alegria, tristeza, entre outros) e cognição (causalidade, discrepância, entre outros) que indicam aspectos morfológicos do texto, contabilizando as palavras em categorias significativas, entre elas, categorias psicológicas (palavras relacionadas a espaço e palavras relacionadas a tempo), categorias cognitivas (certeza e discrepância), categorias estruturais (artigos e preposições), entre outros, utilizando dicionários da língua portuguesa. Segundo Tausczik e Pennebaker (2010), resultados empíricos obtidos com a aplicação LIWC demonstram a capacidade da ferramenta de detectar significados em uma ampla variedade de atributos, para evidenciar o foco do texto, emocionalidade, relações sociais, pensamento e estilos. Com isso, a ferramenta produziu uma tabela com 89 atributos, na qual cada atributo fornece uma pontuação.

Após a extração de 89 atributos, como os tipos de termos extraídos são fixados pela ferramenta, foi realizada uma seleção de variáveis visando à avaliação desses atributos. Nesta etapa do trabalho, foram utilizados os métodos *Information Gain*, *Ratio Gain* e *ReliefF* (Harris, 2002; Liu e Motoda, 2007), com o pacote WEKA[4]. Os resultados obtidos com a aplicação destas técnicas permite a avaliação do nível de importância de cada atributo com relação às classes presentes na base de dados, neste caso, dos tipos de vacinas: “AstraZeneca”,“CoronaVac”, “Moderna”, “Pfizer/BioNTech” e "Sputnik V".

A fim de avaliar a capacidade de identificação dos métodos de seleção, foram empregados os modelos *Random Forest* (RF) para classificar os dados da base de dados, considerando a remoção dos atributos indicados como sendo menos importantes na escala proposta. Após a identificação de um conjunto efetivo de atributos, foi usado o algoritmo *Support Vector Machine* (SVM) (Kecman, 2005) para classificar os *tweets*, visto que este algoritmo tem sido amplamente utilizado em problemas de alta dimensionalidade (muitos atributos) e com múltiplas classes, geralmente com desempenho superior aos de outros classificadores em problemas de predição supervisionada, cujo objetivo é mapear entradas em saída (Piedade, 2020).

A segunda abordagem mencionada no parágrafo anterior se baseia em *Text Mining* e utilizou as técnicas tradicionais de pré-processamento de *Text Mining*: a) Remoção de acentos; b) *Case folding* (descapitalização); c) Remoção de dígitos; d) Remoção de pontuação; e e) *Tokenização* (atomização) (Jurasfsky e Martin, 2020). Em seguida, foi calculado o TF-IDF, que transforma os termos dos documentos em vetores de peso. Esta métrica é composta pela

[4] https://www.cs.waikato.ac.nz/ml/weka/

fusão entre a frequência do termo (TF, *Term Frequency*) e a frequência inversa do documento (IDF, *Inverse Document Frequency*) (Jurasfsky e Martin, 2020). Após todo o pré-processamento, também foi utilizado o SVM visando à comparação dos resultados.

4. Resultados

Esta seção apresenta os resultados obtidos com os modelos de classificação. Como citado anteriormente, foram desenvolvidas duas abordagens de classificação para responder a QP sobre os cincos tipos de vacina. A primeira abordagem foi utilizada a ferramenta LIWC, seguida por métodos de seleção dos atributos. Na segunda foram aplicadas técnicas de *Text Mining* nos *tweets*.

A ferramenta LIWC foi capaz de propiciar a obtenção de 89 atributos morfológicos do texto. Com os registros obtidos entre as vacinas de interesse, foi necessário balancear a base de dados devido às grandes diferenças nas quantidades de publicações entre cada vacina. Após o balanceamento, cada uma das cincos vacinas passou a possuir 4.355 *tweets*. Para a identificação dos interesses, foi avaliada a acurácia obtida com o algoritmo RF a partir da remoção dos cinco atributos que obtiveram as menores pontuações para cada método *(InfoGain, RatioGain* e *ReliefF*).

Entre os métodos, o ReliefF foi o apresentou a maior acurácia comparada aos outros métodos. Inicialmente, foram retirado os 5 últimos atributos com menor pontuação para cada método, o resultado final de acurácia entre os métodos de seleção de atributos foram: Base Original 57,23%; InfoGain 57,19%; RationGain 57,15%; e ReliefF 57,30%. Com isso, seguiu-se removendo os atributos indicados por esse método (levando em consideração o *ranking* criado pelo Relief), até que o valor da acurácia continuasse aumentado, só parando de remover quando a acurácia apresentasse uma queda de valor. Finalizando o processo de seleção de atributos, foram retirados 19 dos 89 atributos, com os atributos restantes a porcentagem de 59,42% de instâncias foram classificadas corretamente, 2,19% maior em relação ao caso base.

Após a realização da seleção dos atributos, buscou-se encontrar a melhor parametrização do algoritmo SVM. O algoritmo explorou exaustivamente entre diversos parâmetros os que apresentavam os melhores resultados, com isso, foram selecionados os seguintes parâmetros, considerando validação cruzada: Kernel: {linear, gaussiano, polinomial, sigmoidal}; Custo: {0,025 0,05 0,1 1 2 5 10 20 50}; gamma (gaussiano, polinomial e sigmoidal) = {0,0001 0,001 0,01 0,1 1 10}; coef (sigmoidal e polinomial) = {0,1 1 10 100} e grau(polinomial) = {1 2 3}.

O melhor resultado de acurácia média, encontrado na validação cruzada com SVM foi 60,72%. Este teve configuração SVM= {kernel=polinomial, C=1, gamma=1, coef=1, grau=2}. A Tabela 1 apresenta a matriz de confusão da base teste, com o modelo SVM escolhido para a primeira abordagem.

Na segunda abordagem (*Text Mining* sobre os *tweets*), o SVM (considerando os mesmos parâmetros indicados para a primeira abordagem) foi aplicado à base dos *tweets* pré-

processados, obtendo 90,45% de acurácia média com configuração: SVM= {kernel=sigmoidal, C=1, gamma=1, coef=10}. Para esta configuração de treinamento, a matriz de confusão da base teste é apresentada na Tabela 2.

Tabela 1: Matriz de confusão do SVM na primeira abordagem

Predição - SVM						
Real		AstraZeneca	Moderna	CoronaVac	Pfizer	Sputnik
	AstraZeneca	**54,88%**	8,61%	14,47%	13,78%	8,27%
	Moderna	3,44%	**8,65%**	4,25%	4,59%	2,07%
	CoronaVac	11,83%	7,58%	**5,21%**	19,29%	10,22%
	Pfizer	12,86%	9,64%	15,50%	**51,66%**	10,33%
	Sputnik	10,10%	4,13%	12,51%	12,97%	**60,28%**

Tabela 2: Matriz de confusão da SVM na segunda abordagem utilizando Text Mining

Predito						
Real		AstraZeneca	Moderna	CoronaVac	Pfizer	Sputnik
	AstraZeneca	**88,98%**	1,04%	3,25%	6,03%	0,70 %
	Moderna	1,69%	**9,86%**	1,92%	3,95%	1,58 %
	CoronaVac	1,59%	0,68%	**9,.51%**	5,21%	1,02 %
	Pfizer	6,36%	3,82%	2,54%	**85,66%**	1,62 %
	Sputnik	1,28%	0,81%	0,47%	2,21%	**95,23 %**

5. Conclusão

Este artigo apresenta uma análise das publicações em relação às vacinas que combatem a COVID-19 nos perfis no *Twitter* da população brasileira, incluindo o presidente do Brasil na

época avaliada, com o objetivo de classificar as vacinas mais mencionadas nas publicações. Para isto, foi realizada uma coleta de *tweets* com termos sobre as vacinas. Com a base de dados criada, foi usada a ferramenta LIWC com o objetivo de extrair os atributos dos textos que seriam utilizados em modelos de classificações.

Uma vez que os atributos extraídos dessa ferramenta são fixados, julgou-se necessária a seleção de variáveis, em que o método *ReliefF* obteve o melhor desempenho com 70 atributos selecionados, aumentando em 2,19% a acurácia em relação ao caso base, com os 89 atributos extraídos pelo LIWC. Após a escolha dos atributos, foi aplicado a Máquina de Vetores-Suporte (*Support Vector Machine* - SVM) para se desenvolver um modelo de classificação automática de *tweets* sobre vacinas, seguido pela análise de sentimentos dos *tweets* classificados. Para esta abordagem, a melhor configuração do algoritmo SVM apresentou uma acurácia de 60,72%.

Os resultados da classificação mostraram que é possível utilizar um modelo de Aprendizado de Máquina com procedimento de seleção de atributos mais relevantes. Com isto, foi possível responder a questão-chave da pesquisa: a vacina Coronavac foi a mais mencionada nos *tweets*, seguida da Pfizer, Moderna, Sputnik V e Astrazeneca nos perfis brasileiros. Por outro lado, no perfil do presidente só foram encontradas menções a respeito das vacinas Coronavac e Astrazeneca.

Na segunda abordagem, com pré-processamento clássico de *Text Mining*, a melhor parametrização do SVM para a classificação dos *tweets* sobre vacinas, registrou 90,45% de acurácia sobre a validação cruzada, indicando que essa metodologia é muito superior à primeira abordagem. Como perspectivas de novos trabalhos, pretende-se utilizar modelos de redes recorrentes (Jurasfsky e Martin, 2020) para aumentar ainda mais a acurácia da classificação sobre as vacinas, além de identificar um *corpus* mais adequado.

6. Referências

Bokang et al. An Ensemble Machine Learning Approach to Understanding the Effect of a Global Pandemic on Twitter Users' Attitudes, mar. 2021.
doi: https://doi.org/10.15837/ijccc.2021.2.4207.

Dodds P.S, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth. Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God,
countries, and Kpop, Sept. 2019.

Glezer, Rubens. As razões e condições dos conflitos federativos na pandemia de Covid-19: coalizão partidária e desenho institucional. Suprema-Revista de Estudos Constitucionais 1.2 (2021): 395-434.

Gonzalez-Padilla, Daniel A. and Tortolero-Blanco, Leonardo. Social media influence in the COVID-19 Pandemic. Epub July 27, 2020.
https://doi.org/10.1590/s1677-5538.ibju.2020.s121

Harris, E. (2002). Information Gain Versus Gain Ratio: A Study of Split Method Biases.

Jurasfsky, D.; Martin, J. H. Speech and Language Processing: An Introduction to

Natural Language Processing, Comp. Linguistics, and Speech Recognition. 2000.

Kaur, Chhinder; Sharma, Anand. Twitter Sentiment Analysis on Coronavirus using Textblob. EasyChair, 2020.

Kecman, Vojislav. (2005). Support Vector Machines – An Introduction 10.1007/10984697_1.

Liu, H., & Motoda, H. (Eds.). (2007). *Computational methods of feature selection*.Malavé,

Mayra. O papel das redes sociais durante a pandemia.
URL: http://www.iff.fiocruz.br/index.php/8-noticias/675-papel-redes-sociais.

Malinen S. Koivula A. and Koiranen I. (2020) How do Digital Divides Determine Social Media Users' Aspirations to Influence Others?. July 22–24, 2020, https://doi.org/10.1145/3400806.3400823

Piedade, Márcio Palheta. Uma abordagem de aprendizagem profunda que usa funções assimétricas para modelagem de pontuação de crédito no varejo. (2020).

Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 2010;29(1):24-54. doi:10.1177/0261927X09351676

Xavier, Fernando et al. Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a COVID-19.Epub July 10, 2020. http://dx.doi.org/10.1590/s0103-4014.2020.3499.016.