Análise das publicações sobre vacinas contra COVID-19 de brasileiros e do Presidente do Brasil no Twitter

Adriano Madureira[1],Douglas A. Vidal[1], Harold de M. Junior[2], Karla Figueiredo[2], Lucas D. Moreira Mendonça[1],Marcos César da Rocha Seruffo[1], Rita Paulino[3], Yomara P. Pires[1]

[1] Faculdade de Computação – Universidade Federal do Pará (UFPA) – Belém,PA – Brazil.

[2]Universidade Estadual do Rio de Janeiro (UERJ) – Rio de Janeiro, RJ – Brazil.

[3]Universidade Federal de Santa Catarina (UFSC) – Florianópolis,SC – Brazil.

adrianomadureira1@gmail.com, vidalstm998@gmail.com, harold.dias@gmail.com, karlafigueiredo@ime.uerj.br, erlucomlpg@gmail.com, seruffo@ufpa.br, rcpauli@gmail.com, yomara@ufpa.br

Abstract: *Since the beginning of 2020, the world has been experiencing a health crisis caused by COVID-19. Although the pandemic is devastating around the world, the actions to fight it and the impacts it suffers are different among nations. However, the vaccine is one of the main tools for controlling the pandemic. In this scenario, Online Social Networks (OSN) have become a significant space for civic and political activity, being among the most used information sources in the world. This article aims to report an analysis of publications on vaccines against COVID-19 by Brazilian users and the president of Brazil on the Twitter platform. Machine Learning Techniques were used and the results show that the Support Vector Machine model was the one that achieved the best performance with 60.72% accuracy with ReliefF parameter extraction for the analysis of tweets that indicated which vaccines were the most mentioned in the president's and users' profiles.*

Resumo: *Desde o início de 2020 o mundo vive uma crise de saúde ocasionada pela COVID-19. Embora a pandemia seja devastadora em todo o mundo, as ações de enfrentamento e os impactos sofridos são distintos entre as nações. No entanto, a vacina é uma das principais ferramentas para o controle da pandemia. Neste cenário, as Redes Sociais Online (RSO) se tornaram um espaço significativo para atividade cívica e política, estando entre as fontes de informação mais utilizadas no mundo. Este artigo visa reportar uma análise das publicações sobre vacinas contra a COVID-19 de usuários brasileiros e do presidente do Brasil na plataforma Twitter. Técnicas de Aprendizado de Máquina (Machine Learning) foram utilizadas e os resultados mostram que o modelo Support Vector Machine foi o que conseguiu melhor desempenho com 60,72% de acurácia com extração de parâmetro ReliefF para a análise dos tweets que indicavam quais as vacinas mais mencionadas nos perfis do presidente e dos usuários.*

Palavras-chaves: Aprendizado de Máquina, Redes Sociais Online, Support Vector Machine, LIWC.

1. Introdução

No ano de 2020, a humanidade vivenciou uma situação de calamidade pública e de emergência em saúde com a propagação da COVID-19. Uma catástrofe de escala mundial vivenciada em ondas sucessivas, distintas variantes do vírus, desafios nas restrições de circulação globais e saúde econômica dos países. A propagação do vírus gerou, mundialmente em 2020, 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].

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[2], 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.

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

Assim, este artigo aplicou uma metodologia de coleta e classificação de *Tweets*, utilizando técnicas de Aprendizado de Máquina, aplicando duas propostas e algoritmos

[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://interactives.lowyinstitute.org/features/COVID-performance/

para modelos preditivos, por meio das mensagens durante a pandemia, com o intuito de verificar quais as vacinas que combatem a COVID-19 foram mais comentadas durante o período da coleta entre a população e o presidente. Para esse fim, o artigo é estruturado da seguinte forma: na seção *2. Trabalhos correlatos* fornecem uma visão geral de estudos de RSO e de COVID-19. A seção *3. Procedimentos metodológicos* fornece detalhes sobre as duas abordagens de coleta e pré-processamento de dados investigadas neste trabalho, além do uso de algoritmos de classificação. Na seção *4. Resultados obtidos* são apresentados os resultados de classificação das mensagens com as duas abordagens. Por fim, na seção *5. Conclusão* são apresentadas as conclusões e perspectivas de novos trabalhos.

2. Trabalhos correlatos

Estudos recentes 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 executivo se comunica com o público. De forma geral, há muitos trabalhos que envolvem Aprendizado de Máquina e RSO. Em Kaur (2020), foi realizada a análise de sentimentos em relação à doença do coronavírus (COVID-19). Bokang et al (2021) usou 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

Para desenvolvimento deste trabalho, foram seguidas duas etapas: 1. Coleta e Processamento dos Dados; e 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: Q. **Qual(is) 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.

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 temos: 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*.

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 possa 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 Q, 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, 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 que indicam aspectos morfológicos do texto, contabilizando as palavras em categorias significativas utilizando dicionários da língua portuguesa. Segundo Tausczik e Pennebaker (2010), resultados empíricos obtidos com a aplicação LIWC demonstram a capacidade desta 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, onde 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[3]. 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".

Para avaliar a capacidade de identificação dos métodos de seleção, foram empregados os modelos 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 SVM (Kecman, 2005) para classificar os *Tweets*, visto que este algoritmo têm sido amplamente utilizados 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.

A segunda abordagem mencionada acima baseia-se em *Text Mining* e utilizou as tradicionais técnicas 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 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 comparação dos resultados.

4. Resultados obtidos

4.1 Resultados Obtidos com os Modelos de Classificação

Conforme mencionado anteriormente, para responder a Q foram desenvolvidas duas abordagens de classificação dos cinco tipos de vacina, abordado no conteúdo das publicações. A primeira abordagem envolveu a utilização da ferramenta LIWC, seguida por métodos de seleção dos atributos, enquanto que na segunda foram aplicadas técnicas de Text Mining diretamente aos *tweets*.

A ferramenta LIWC propiciou a extração de 89 atributos morfológicos do texto. Devido às diferenças de quantidades de registros extraídos entre as vacinas de interesse, foi necessário balancear a base de dados. Com isso, cada uma das cinco vacinas passou a possuir 4355 *tweets*. Para a identificação dos atributos de interesse, foi avaliada a

[3] **Weka:** https://www.cs.waikato.ac.nz/ml/weka/

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*).

O método *ReliefF* foi o que apresentou uma acurácia maior em comparação aos outros métodos. Os resultados finais de acurácia para os métodos de seleção de atributos, vale ressaltar que para os métodos foram retirados os últimos 5 atributos, foram: Base Original 57,23%; *InfoGain 57,19%*; *RationGain* 57,15%; e *ReliefF* 57,30%. Assim, seguiu-se removendo os atributos indicados por esse método (tendo como base o *rank* criado pelo *ReliefF*), enquanto o valor da acurácia continuou aumentando, até que a acurácia apresentou queda de valor. No final do processo de seleção dos atributos, chegou-se a um total de 70 atributos, obtendo a porcentagem de 59,42% de instâncias classificadas corretamente, ou seja, com 2,19% maior do que o caso base. Uma vez realizada a seleção dos atributos, buscou-se identificar qual a melhor parametrização do algoritmo SVM. O algoritmo SVM explorou exaustivamente 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%, 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.

Tabela 1: Matriz de confusão da SVM - Base Teste

Previsão - 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%**

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 essa configuração de treinamento, a matriz de confusão da base teste é apresentada na Tabela 2.

Tabela 2: Matriz de confusão da SVM - Base Teste

Previsto						
		AstraZeneca	Moderna	CoronaVac	Pfizer	Sputnik
Real	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 trabalho apresenta uma análise das publicações em relação às vacinas que combatem a COVID-19 dos perfis no *Twitter* do presidente do Brasil Jair Bolsonaro e da população brasileira com o objetivo de classificar quais 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.

No entanto, como os atributos extraídos dessa ferramenta são fixados, julgou-se necessária a seleção de variáveis, onde o método *ReliefF* teve 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 o algoritmo *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 essa abordagem, a melhor configuração do algoritmo SVM apresentou uma acurácia de 60,72%.

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

Já a melhor parametrização do SVM para a classificação dos *tweets* sobre vacinas, utilizado pré-processamento clássico de *Text Mining 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.

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

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.