Aplicação de análise de sentimentos no Twitter para avaliação da percepção pública quanto a cloroquina

João Rodrigo W. Olenscki[1], Fernando Xavier[1], André Luis Acosta[2], Antonio Mauro Saraiva[1], Maria Anice Mureb Sallum[3]

[1]Escola Politécnica da Universidade São Paulo

[2]Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology
Faculdade de Medicina da Universidade de São Paulo

[3]Faculdade de Saúde Pública da Universidade de São Paulo

{joao.olenscki,fxavier, saraiva, masallum}@usp.br, andreluisacosta@gmail.com

Abstract. *During the COVID-19's pandemic, several forms of treatment are being tested by researchers. Among them, the main compound of antimalarial drug, chloroquine, stands out, it has gained great repercussion after favorable statements made by the presidents of Brazil and the United States. In order to analyze public opinion regarding this treatment, this research analyzed the use of machine learning algorithms on posts in social networks regarding this medicine. According to our preliminary results, we identified that this method is effective for this research approach, which can also be used in other health-related topics, assisting public management in monitoring and evaluating the effectiveness of their communication activities, as well as fighting fake news.*

Resumo. *Durante a pandemia de COVID-19, diversas formas de tratamento têm sido testadas por pesquisadores. Dentre elas, destaca-se a cloroquina, que ganhou grande repercussão após declarações favoráveis feitas pelos presidentes do Brasil e dos Estados Unidos. Com o objetivo de analisar a opinião pública a respeito da cloroquina, esta pesquisa tem como objetivo analisar postagens a respeito desse medicamento em redes sociais através de aprendizado de máquina. De acordo com resultados preliminares, nota-se que estes métodos podem ser úteis na vigilância em saúde, auxiliando a gestão pública no monitoramento e avaliação da efetividade de suas atividades de comunicação, bem como combate às fake news.*

1. Introdução

Nos tempos atuais, marcados pela conectividade e pela instantaneidade da informação, as redes sociais agregam informações e, por muitas vezes, desinformações. Tais redes possuem alcance muito grande: estima-se que existam hoje cerca de 3 bilhões de usuários ativos em redes sociais [Clement 2020]. Além disso, as redes sociais excercem influência na vida e nas opiniões de seus usuários [Zeitel-Bank and Tat 2014], que interagem e conversam entre si, emitindo opiniões sobre os mais variados assuntos. Com a atual crise sanitária gerada em razão da pandemia do COVID-19 e o distanciamento social físico submetido a quase todos os países do mundo, estas interações virtuais, bem como o número de usuários ativos cresceu vertiginosamente: cerca de 40% para os países em fases intermediárias da pandemia e quase 80% em países nas fases finais [Perez 2020].

Esta rede interligada de informações pode ser facilmente usada para o espalhamento de desinformação sobre mais diversos temas, dentre eles, a saúde é um dos principais alvos. Movimentos como o de anti-vacinação se espalham muito mais rapidamente em ambientes virtuais [WHO 2009, p. 47]. Assim, a análise da opinião pública quanto à aceitação do medicamento para determinada doença, bem como o monitoramento de informações de falsos remédios torna-se um fator essencial para o combate à desinformação e uma importante contribuição para a gestão da saúde pública.

Tendo estes aspectos em consideração, este trabalho propõe um estudo sobre a opinião pública a respeito da inclusão da cloroquina e da hidroxicloroquina entre os medicamentos utilizados no tratamento da COVID-19. Esta tarefa envolveu a análise de sentimento de cerca de 70 mil frases, algo que é inviável de se analisar manualmente. No entanto, pela segregação de uma pequena amostra representativa desta base classificada manualmente, e aplicando os algoritmos de aprendizado de máquina foi possível produzir uma classificação para toda a base. Busca-se, com isso, apresentar uma abordagem que pode ser útil às atividades de vigilância em saúde, extraindo informações de maneira eficiente de grandes bases de dados não-estruturadas.

O trabalho está organizado da seguinte forma: na Seção, 2 apresenta-se os conceitos; na Seção 3 está a descrição dos materiais e métodos usados; na Seção 4 são apresentados os processos de coleta e o pré-processamento dos dados; na Seção 5 é feita a discussão sobre os resultados obtidos; as considerações finais, bem como trabalhos futuros planejados são apresentados na Seção 6.

2. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (em inglês, *NLP*) é uma área de pesquisa tanto da Computação como da Linguística. Pode ser definida como "uma gama de técnicas computacionais criadas para analisar e representar textos naturais [escritos por humanos], em um ou mais níveis de análise linguística, com o objetivo de obter processamento de linguagem semelhante ao humano." [Liddy 2001].

Com o uso de NLP é possível transformar uma análise, antes linguística, em vetorial. As palavras são vetorizadas a partir de si mesmas e de outras próximas, de forma a se criar espaços vetoriais com documentos inteiros, tornando operações algébricas entre estes vetores possível. Assim, é factível calcular distâncias e ângulos entre vetores de palavras e, com estas informações, estimar a proximidade entre palavras ou frases, de maneira que frases com significado similar sejam agrupadas [Zhang and et al. 2016].

3. Materiais e Métodos

3.1. Dados utilizados

Para realização deste estudo, foram coletados dados do Twitter, uma das principais redes sociais, com 330 milhões de usuários ativos mensalmente [Clement 2020]. A escolha dessa plataforma como fonte de dados deve-se à dois fatores:

(i) o Twitter possui uma *Application Programming Interface* (API) que facilita a extração dos dados necessários; e
(ii) o Twitter possui, em sua plataforma, marcadores de atividades ligados a *tags* que são facilmente localizáveis e analisáveis, conhecidas como *hashtags* ('#').

3.2. Metodologia

A metodologia utilizada nesta pesquisa (Figura 1) foi baseada no ciclo de Ciência de Dados proposto por [Shcherbakov and et al. 2014]. Os dados, antes de serem analisados pelos algoritmos de aprendizado de máquina, são submetidos a uma etapa de pré-processamento que, em NLP, tem caraterísticas próprias, descritas na Seção 4.2 Adicionalmente, a metodologia utilizada inclui constante interação com especialistas de domínio que, neste projeto de pesquisa, são pesquisadores da área da Saúde Pública.

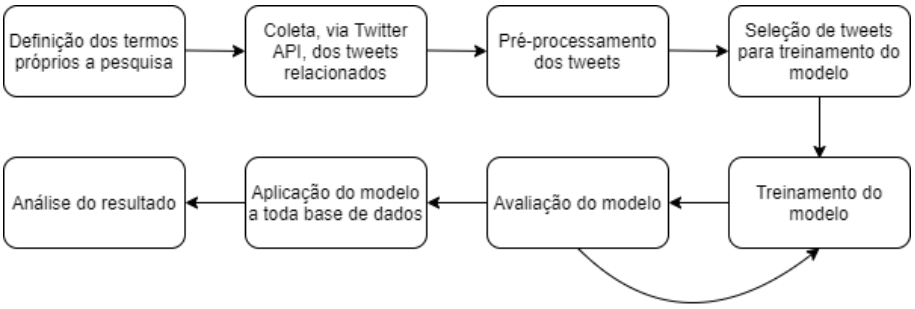

Figura 1. Diagrama de metodologia (elaborado pelos autores)

4. Desenvolvimento

4.1. Coleta de Dados

O primeiro passo para a coleta de dados é definir as palavras-chave que serão usadas como filtros para buscas nas postagens do Twitter. Nessa pesquisa, foram utilizadas diversas palavras-chave, apresentadas na Tabela 1, relacionadas ao nome COVID-19, utilizadas normalmente por usuários brasileiros do Twitter, bem como algumas palavras relacionadas aos sintomas e ao tratamento com a cloroquina, objetos de estudo de nossa análise.

Tabela 1. Tabela das palavras-chave

palavras-chave
'covid19', '#covid19', 'corona', 'covid-19', '#COVID19', '#coronavírus', '#coronavirus', '#covid19brasil', 'coronavirus', 'cloroquina', 'hidroxicloroquina', 'covid', 'febre', 'tosse', 'coriza', 'azitromicina'

Em seguida, foram coletados os dados de postagens que continham pelo menos uma destas palavras-chave. Esta rotina foi executada diariamente entre 26 de março e 21 de abril. Para o treinamento do modelo, optou-se por realizar uma classificação manual em uma pequena, porém representativa, parte da base. Na montagem do *set* de treino foram selecionados, aleatoriamente, 750 *tweets* de 6 dias consecutivos (ou seja, 125 postagens por dia) no intervalo de 08 a 13 de abril. Após testes preliminares com esta base decidiu-se por expandi-la para verificar se o tamanho da mesma era um fator limitante para a acurácia do modelo e assim coletaram-se outros 1250 *tweets* da base completa (postagens do intervalo de 26 de março a 21 de abril), desta forma, obteve-se ao todo 2000 *tweets* para a formação da base de treino.

4.2. Pré-processamento

O pré-processamento é uma parte preliminar da construção do algoritmo, nela executa-se uma formatação na apresentação e estruturamento da base. Neste caso específico a base é

gerada como um arquivo .csv, com 10 colunas e cerca de 180000 linhas (por dia). De todas estas informações a única relevante ao estudo é o texto e portanto todos os outros campos devem ser relevados. Além disso, no contexto do uso de NLP, existem algoritmos criados para "limpar" o texto, assim é possível retirar de cada documento (*tweet*) as palavras vazias (em inglês *stop words*), as menções a outros usuários ou mesmo as *hashtags* ('#').

Análise preliminar destes métodos (Tabela 2), aplicando o classificador SVC (*Support Vector Classifier*), indica que, ao contrário do comportamento esperado [Silva and Ribeiro 2003], o melhor tratamento para este tipo de dado é não retirar palavras. Isto pode ser explicado pela limitação de caracteres do Twitter, fazendo com que as frases possuam poucas palavras, aumentando a importância de palavras que não apresentariam, em outros contextos, sentido para a análise. Esta análise é importante na determinação de qual método de pré-processamento deverá ser aplicado posteriormente na análise dos diferents classificadores.

Nessa etapa, também foi realizada a remoção de elementos não relevantes à análise - isto é - *tweets* que apesar de conterem as palavras-chave não mantinham relação com os objetos de estudo, como *tweets* sobre futebol, transmissões ao vivo de cantores, etc.

Tabela 2. Tabela da precisão dos métodos de pré-processamento

Método aplicado	Acurácia (%)
nenhum	56.25
remover '#' e '@'	55.79
remover 'stopwords', '#' e '@'	55.26
remover 'stopwords'	53.16
remover '@'	52.89
remover '#'	50.00

4.3. Análise dos Dados

Quanto aos dados de treino do modelo, estes foram classificados em 3 grupos diferentes: a favor ao uso da cloroquina para o tratamento de SARS-nCoV2, contra o mesmo e neutro. As regras de classificação podem ser encontradas na Tabela 3

Tabela 3. Tabela das regras utilizadas na classificação manual dos tweets

classificação	regra aplicada
1	tweets que afirmavam que a cloroquina era um remédio eficiente
0	tweets sem juízo de opinião que não apresentavam contexto suficiente sofrer avaliação tweets de ataque político ou pessoal a opositores
-1	tweets que afirmavam que a cloroquina não era um remédio eficiente ou que citavam os efeitos colaterais tweets que relacionavam o uso da cloroquina com a morte de pacientes

Foi realizada uma análise da frequência relativa dos termos de interesse (total de aparições de certo termo dividido pelo total de *tweets*), obtendo-se o gráfico presente na Figura 2. É possível observar um aumento brusco no uso do termo perto dos dias 8 e 9 de abril, o que pode ser explicado pelo discurso do presidente do Brasil no dia 8 a respeito do tratamento do COVID-19 [Fernandes and Fabrini 2020]. O pico de citações do termo cloroquina exibido logo após o discurso do presidente é uma evidência da alta sensibilidade da abordagem em capturar efeitos de discursos na opinião pública.

No processamento dos tweets, foram utilizados diversos algoritmos e, com o objetivo de comparar os resultados, calculou-se a acurácia de cada um, com os resultados

expostos na Tabela 4. Foi possível observar que o SVC obteve a maior acurácia, embora os algoritmos *Logistic Regression* e SGD tenham obtido resultados similares.

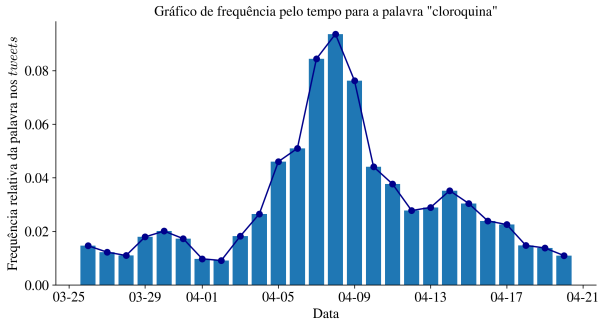

Figura 2. Frequências relativa do termo cloroquina

Tabela 4. Tabela de comparação dos métodos classificadores

Classificador	Acurácia (%)
Support Vector Classifier	56.25
Logistic Regression	55.26
Stochastic Gradient Descent	53.42
Maximum Entropy	52.11
Multinomial Naive Bayes	52.11

Além desta escolha do classificador, executou-se um redimensionamento na base de dados de treino com o intuito de tentar balancear o número de tweets em cada classe.

5. Resultados e Discussão

Como indicado na Tabela 4 o melhor classificador obtido para o estudo foi o SVC, apesar dos outros métodos apresentarem resultados próximos, com exceção da Árvore de decisão, que apresentou valor relativamente baixo. As acurácias abaixo de 60% podem ser explicadas pela limitação de caracteres do Twitter, o que consequentemente limita o número de palavras utilizadas em cada postagens. Para uma análise vinculada a um determinado tema frases com sentimentos opostos podem frequentemente compartilhar palavras, tornando o seu processo de distinção difícil. Associado a este mesmo problema, cita-se a grande variação lexical que forma o *corpus* do conjunto de *tweets*, grandeza que torna muito difícil obter uma base de treino representativa desta variação. Pretende-se, no desenvolvimento deste trabalho preliminar, analisar outros métodos de modo a atingir melhores resultados. Além disso, analisa-se a possibilidade do uso de outras métricas para a avaliação dos resultados, tais como a matriz de confusão do classificador, que poderia explicar melhor o funcionamento e as limitações do método aplicado.

Tabela 5. Comparação entre os resultados de treino e teste

classificação	número de ocorrências no *set* de treino	frequência no treino (%)	número de ocorrências no *set* de teste	frequência no teste (%)
-1	506	25.30	17089	27.63
0	819	40.95	22322	36.10
1	675	33.75	22418	36.26

Utilizando o classificador SVC, a partir do *set* de treino, classificou-se os *tweets* com a palavra “cloroquina” no intervalo de tempo de interesse. A Tabela 5 mostra os

resultados em números absolutos e em porcentagem. É importante ressaltar que o *set* de treino foi classificado manualmente, enquanto que o de teste foi classificado via *fitting* com o modelo gerado.

6. Considerações Finais

Este trabalho se propôs a analisar a percepção pública quanto ao medicamento cloroquina através da análise de sentimentos, utilizando-se de classificadores lineares aplicados em postagens do Twitter. Resultados preliminares indicam que esta abordagem pode ser útil na análise desses tipos de dados. Como trabalhos futuros, pretende-se realizar uma revisão bibliográfica mais aprofundada nos temas apresentados, bem como aplicar algoritmos de clusterização para auxiliar a classificação, e também aplicá-los a uma CNN (*Convolutional Neural Network*) utilizando o *Word2Vec*.

As decisões em medicina e saúde pública devem ser tomadas com base em subsídios científicos mas também de dinâmica social e de opinião, muitos dos quais podem ser extraídos das redes sociais. Ademais, além da informação prévia que orienta a decisão a ser tomada, há a informação que resulta da repercussão de uma decisão tomada, portanto, por exemplo, redes sociais podem ser efetivas para avaliar o nível de insatisfação gerado por ações como a quarentena. Pois, além da recomendação médica, que busca maximizar o tempo de isolamento, existe a recomendação econômica e de segurança que também são relevantes ao gestor político. De nada adiantará demandar isolamento se a população se rebelar socialmente contra seus administradores.

Referências

Clement, J. (2020). Number of social network users worldwide from 2010 to 2023. Available: `https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/`. Accessed: 21.04.2020.

Fernandes, T. and Fabrini, F. (2020). Em pronunciamento, bolsonaro defende cloroquina e retoma embate com governadores e prefeitos. *Folha de S. Paulo*.

Liddy, E. D. (2001). Natural language processing. *Encyclopedia of Library and Information Science*, 2nd Ed.

Perez, S. (2020). Report: Whatsapp has seen a 40% increase in usage due to covid-19 pandemic. Available: `https://tcrn.ch/3drizIt`. Accessed: 21.04.2020.

Shcherbakov, M. and et al., S. (2014). Lean data science research life cycle: A concept for data analysis software development. volume 466, pages 708–716.

Silva, C. and Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1661–1666 vol.3.

WHO (2009). *State of the worlds vaccines and immunization*. World Health Organization.

Zeitel-Bank, N. and Tat, U. (2014). Social Media and Its Effects on Individuals and Social Systems. Proceedings of the Management, Knowledge and Learning International Conference 2014, pages 1183–1190. ToKnowPress.

Zhang, Y. and et al., R. (2016). Neural information retrieval: A literature review. arXiv preprint arXiv:1611.06792.