doi: 10.1590/s0103-4014.2020.3499.016

Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a Covid-19[1]

Fernando Xavier,[I] *João Rodrigo W. Olenscki,*[II] *Andre Luis Acosta,*[III] *Maria Anice Mureb Sallum*[IV] *e Antonio Mauro Saraiva*[V]

Introdução

A pandemia da Covid-19 está impactando negativamente a população mundial, com graves efeitos na saúde pública e grandes desafios para os governos e tomadores de decisão. São milhões de casos, e boa parte da população mundial está sob alguma forma de isolamento social. Como ainda não há nem vacina, nem tratamento efetivo, o principal desafio tem sido controlar a expansão da doença para não sobrecarregar os serviços públicos de saúde e evitar o aumento das mortes. Por ser doença recente, as atividades de vigilância em saúde têm como um dos focos conhecer os fatores associados à doença, como sintomas, formas de transmissão e fatores de risco de gravidade, além de mapear os casos e lidar com problemas de subnotificação dos casos de contaminações e óbitos.

Além da questão da subnotificação, existem outros desafios para a vigilância em saúde, como análise dos dados que são gerados de maneira muito rápida durante a pandemia e entendimento dos fatores de risco para uma doença ainda pouco conhecida. Dar respostas rápidas às questões e lidar com esses desafios tornam-se preponderantes para aprimorar o combate à doença e seus efeitos. Em paralelo, há 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. Essas publicações podem ter sérios impactos no bem-estar da população, que pode ser levada a utilizar medicamento incorreto e potencialmente perigoso ou não participar de campanhas de promoção à saúde, como vacinação.

Redes sociais são plataformas com alta velocidade na geração de dados, com postagens feitas a todo instante. Em relação à Covid-19, e considerando apenas postagens em português no Twitter, este estudo coletou mais de 7,7 milhões de postagens durante os 62 dias de coleta, resultando em média cerca de

130 mil postagens por dia ou 5.188 postagens por hora. Dado esse volume de dados a respeito da pandemia, um dos desafios está na extração de informações que possam servir de apoio às atividades de vigilância em saúde.

Na era do *Big Data*, a extração de informação útil de grandes e crescentes volumes de dados tem sido o desafio de pesquisadores e empresas. Ele reside não apenas na extração da informação, mas fazê-la de forma eficiente e em tempo adequado, para a tomada de decisão oportuna. Várias abordagens têm sido propostas para essa tarefa de análise de dados de maneira eficiente. Destacam-se as técnicas baseadas em Ciência de Dados (CD), que usam recursos computacionais para analisar grandes massas de dados mediante a execução de algoritmos de aprendizado de máquina.

Nesse cenário, esta pesquisa objetivou aplicar abordagem baseada em CD para uso de redes sociais como ferramenta complementar às atividades de vigilância em saúde. Este artigo traz, inicialmente, uma discussão teórica relacionando a vigilância em saúde e análise de dados em redes sociais, e apresentando trabalhos correlatos. A seguir, são apresentados métodos e dados utilizados nos experimentos realizados com dados do Twitter coletados durante a pandemia da Covid-19. Na seção seguinte, os resultados da pesquisa são apresentados e discutidos. Por fim, são propostas oportunidades de trabalhos futuros nessa área.

Estado da arte e aplicações existentes

Vigilância em Saúde

As atividades de vigilância são fundamentais para a definição e acompanhamento das políticas da saúde, fornecendo subsídios para a atuação dos profissionais no combate às doenças e na promoção da saúde. O conhecimento da situação epidemiológica pode auxiliar no desenvolvimento de novos conhecimentos e, com isso, estratégias de intervenção podem ser aprimoradas. Ademais, a vigilância constitui-se em importante instrumento para monitoramento das ações, trazendo informações para apoio à tomada de decisão (Teixeira et al., 2000; Arreaza; Moraes, 2010).

Ao trazer luz sobre a realidade, a vigilância é uma ferramenta muito importante para a gestão da saúde, sem a qual as ações tomadas seriam meramente um processo de tentativa e erro. O planejamento e acompanhamento deve ser feito de acordo com um diagnóstico correto da situação e isso depende de dados.

São inúmeras as fontes de dados utilizadas para as atividades de vigilância, dados primários e secundários, do setor público e do privado. No entanto, quando as informações dessas fontes são insuficientes para a gestão, novos dados precisam ser coletados e uma forma de fazer isso é a realização de inquéritos para conhecer, entre outras informações, o perfil populacional e a distribuição dos fatores de risco na população, para variáveis como idade, gênero, localização. No entanto, a realização de inquéritos é limitada pelos custos envolvidos e tempo dispendido na sua realização (Malta et al., 2008).

O modelo atual de coleta de dados é baseado em uma forma passiva, pois, em geral, o sujeito fornece os dados apenas quando tem contato direto com os serviços de saúde, seja pela busca de atendimento, seja pela realização de inquéritos. Dessa forma, as ações podem não ter o seu potencial máximo de eficiência por falta de dados suficientes. Em relação à Covid-19, um grande desafio tem sido a subnotificação dos casos e óbitos e diversos estudos têm sido realizados, associados a modelos matemáticos e estatísticos, para estimar esses números. Existem estimativas de que há sete vezes mais casos de pessoas contaminadas do que os oficialmente reportados no Brasil (Ribeiro; Bernardes, 2020), enquanto outros indicam que há onze vezes maiores (Freitas et al., 2020).

Mas apenas ter acesso aos dados pode ainda ser insuficiente, pois há de considerar a questão temporal. Em epidemias, quanto mais rápida a situação epidemiológica for conhecida, mais eficientes podem ser as estratégias de controle. Dessa forma, além de maior cobertura dos dados, há necessidade de que esses dados estejam disponíveis o quanto antes, para minimizar o risco de planejar ações para um cenário que não reflete a realidade.

Big Data e Ciência de Dados na Saúde

Com o aumento do poder de processamento computacional e dos dispositivos conectados, como *smartphones, tablets* e sensores, há um grande crescimento na geração de dados, de formatos variados e em velocidades cada vez maiores. Esses fatores representam características do que se convencionou chamar de era de *Big Data,* em que o grande desafio é extrair informação útil desses grandes e variados conjuntos de dados (Dhar, 2013; Concolato, 2017).

Nesse cenário, várias abordagens para análise de dados têm ganhado importância, pois não basta apenas extrair informação útil, mas fazê-lo de forma eficiente. Com alta velocidade de geração dos dados e o grande volume armazenado de diversas fontes, produzir respostas rápidas torna-se fundamental, visto que alto tempo de análise pode gerar diagnóstico de um cenário defasado. Esse fato, em cenário de dispersão acelerada de uma doença, torna-se ainda mais importante, pois os profissionais da saúde precisam dar respostas mais rápidas à população e avaliar as medidas efetivas de intervenção para conter a epidemia. Decisões tardias aumentam o risco de surtos transformarem-se em epidemias ou pandemias.

Dessa forma, destaca-se a contribuição da Ciência de Dados (CD), que pode ser caracterizada como um conjunto de disciplinas e técnicas para extração de informações úteis de maneira eficiente através da atuação de uma equipe multidisciplinar. A visão *data-driven* da CD traz a necessidade da inclusão de outras técnicas e métodos de diversas disciplinas, como Computação, Estatística, Matemática, entre outras. Destaca-se também a participação dos especialistas de domínio, que são pesquisadores e usuários dos dados da área de aplicação. São esses pesquisadores que têm o maior conhecimento acerca dos dados utilizados, e devem participar de todo o planejamento, execução e validação da pesquisa,

pois a simples execução de um algoritmo de aprendizado de máquina pode produzir um resultado que, embora matematicamente explicável, pode não ter relevância ao domínio de aplicação.

Em cenários de epidemias ou pandemias como a da Covid-19, torna-se ainda mais importante a realização de pesquisas que envolvam equipes multidisciplinares, de modo a produzir resultados que tenham menor tempo de implementação nas atividades de vigilância em saúde. Tempo é fator crucial no sucesso das ações de combate às doenças e a CD pode ter importante contribuição, não apenas no cenário de *Big Data*, mas principalmente nele.

Redes Sociais e Vigilância em Saúde

Com a popularização do acesso à internet, as redes sociais estão entre as principais plataformas em número de usuários. Dados estimados de abril/2020 indicam que a principal rede social, o Facebook, tem cerca de 2,5 bilhões de usuários. O Twitter teria cerca de 386 milhões de usuários ativos, com quase 14,5 milhões de usuários no Brasil (Statista, 2020). Com a pandemia e bilhões de pessoas em diversas formas de isolamento, existe um maior uso de redes sociais. Relatório do primeiro trimestre de 2020 do Twitter indica que houve um aumento anual de 24% dos usuários ativos monetizáveis, o maior aumento já registrado anualmente, além de um crescimento de 14% em relação ao trimestre anterior (Twitter, 2020).

Com esse número de usuários, há um volume imenso de dados de postagens e é natural que esses dados sejam usados em pesquisas relacionadas à saúde. As postagens podem conter notícias, opiniões e relatos dos usuários e, em tempos de pandemia, é de se esperar que muitas dessas postagens sejam relacionadas à Covid-19. Dessa forma, algumas questões de pesquisa poderiam ser respondidas, por exemplo:

- Qual a é opinião das pessoas a respeito de tratamento ou medida de controle?
- Quantas pessoas estão relatando um sintoma?
- Que *fake news* estão sendo disseminadas na internet?
- Quais são os locais em que pessoas estão relatando um sintoma?

Mediante métodos automatizados de análise, essas questões poderiam ser respondidas de forma eficiente, possibilitando visão em tempo real para os profissionais da saúde. Nesse sentido, diversas pesquisas vêm sendo desenvolvidas para utilização de dados de redes sociais como apoio à vigilância em saúde. Uma revisão de literatura apontou benefícios de usar dados do Twitter na vigilância, pela disponibilidade em tempo real dos dados e atributos que podem ser usados, como informações sobre a localização dos usuários (Jordan et al., 2019).

Em 2016, foi proposta uma abordagem para monitoramento e previsão de epidemias relacionadas ao *influenza*. Esse projeto usou dados do Twitter e incluiu atividades de coleta de dados, análise de sentimento e visualização para

monitoramento da disseminação do *influenza* (Byrd et al., 2016). Outro trabalho teve como objetivo a criação de um sistema de vigilância em saúde, com métodos para coleta de dados e criação de um *data warehouse* de dados de redes sociais (Garzón-Alfonso; Rodríguez-Martínez, 2018).

Covid-19 e Redes Sociais

Como outras doenças, os dados de redes sociais podem ser usados na vigilância em saúde em relação à Covid-19. Dada a velocidade de disseminação da doença, produzir respostas rápidas para questões como as listadas anteriormente pode gerar informações para tomada de decisão e, com isso, apoiar o planejamento e monitoramento de políticas para a promoção da saúde e controle de doenças.

Trabalho recente analisou cerca de 126 mil postagens durante duas semanas de janeiro de 2020, quando a Covid-19 ainda estava restrita a poucos países. Essa pesquisa teve como objetivo analisar o sentimento das pessoas quanto à doença, assim como avaliar o conteúdo das mensagens. Notou-se que houve predominância da discussão sobre impactos políticos e econômicos da doença do que em relação aos riscos e métodos de prevenção (Medford et al., 2020).

Outra pesquisa teve como objetivo identificar a prevalência de informação de baixa qualidade durante a pandemia da Covid-19. Um dos resultados indicou que a disseminação de informações de baixa qualidade foi potencializada pelo uso de métodos automatizados, os *bots*. A pesquisa identificou também que os principais assuntos nas postagens com informações de baixa qualidade referiam-se à política dos Estados Unidos, *status* da pandemia e questões econômicas. Segundo os autores, os resultados evidenciam uma certa "politização da pandemia" (Yang et al., 2020).

Processamento de Linguagem Natural (NLP)

Natural Language Processing (NLP) é uma área que objetiva a extração de informação de texto, unindo disciplinas de vários campos de conhecimento, como Computação e Linguística, possibilitando também a transformação de texto em dados estruturados (Chowdhury, 2003).

Das aplicações de pesquisa em NLP destacam-se tradução automática, reconhecimento de fala e sumarização de texto. Com aumento da geração de texto em plataformas baseadas na internet, a extração e interpretação de informação útil tem ganhado ainda mais relevância. Com o aumento da pesquisa em Inteligência Artificial, existem diversas linhas de pesquisa para a criação de *chatbots*, utilizados em áreas como Atendimento ao Cliente e Saúde (Smutny; Schreiberova, 2020; Pryss et al., 2019).

Na Saúde existem diversas aplicações de NLP que envolvem tanto tarefas de atendimento a pacientes como processamento de registros de saúde eletrônicos. Em saúde mental, por exemplo, existem aplicações baseadas em NLP para fins terapêuticos e de triagem; em Radiologia NLP pode ser usada para extração de informação de prontuários médicos (Ta et al., 2020; Abd-alrazaq et al.,

2019; Pons et al., 2016). Na Vigilância em Saúde, há histórico de aplicação de NLP e aprendizado de máquina, incluindo análise de dados de redes sociais para criar modelos preditivos e identificar tendências relacionadas às doenças (Dai et al., 2017; Achrekar et al., 2011).

Portanto, dado o histórico da aplicação de NLP em atividades de atendimento de pacientes e em vigilância em saúde, e considerando o volume de dados gerados em redes sociais, existem oportunidades de pesquisa na análise de dados do Twitter durante a pandemia da Covid-19. Além disso, por meio de métodos computacionais, essas informações podem ser acessíveis em tempo real aos gestores e profissionais da saúde. As pesquisas com dados de redes sociais podem ir além de estudos retrospectivos e gerar produtos de apoio à tomada de decisão em tempo real das salas de situação. Algumas possíveis aplicações com NLP aplicada a redes sociais podem incluir:

- Análise da opinião sobre medidas adotadas (Ex.: isolamento social);
- Avaliação do impacto das estratégias de comunicação;
- Identificação de possíveis sintomas relacionados às doenças;
- Identificação e avaliação do impacto das *fake news.*

Materiais e métodos

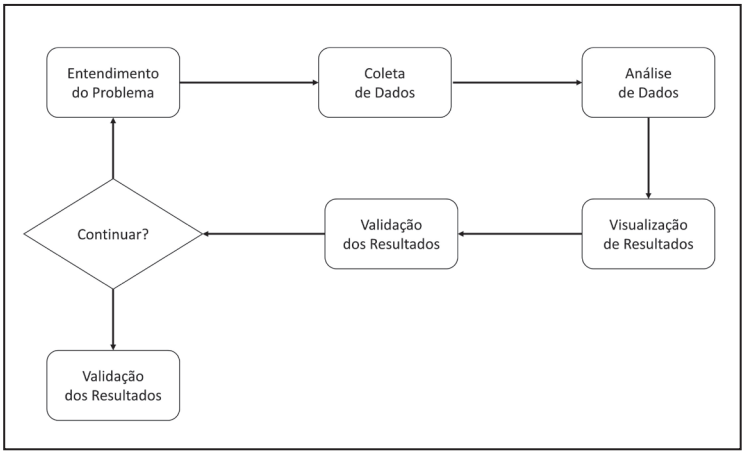
Método baseado em Ciência de Dados

Para execução desta pesquisa, foi adotado um ciclo de vida de Ciência de Dados (Figura 1), que consiste desde a definição do problema de pesquisa até a validação dos resultados junto aos especialistas de domínio.

Nessa abordagem, a participação dos especialistas de domínio é fundamental para o desenvolvimento da pesquisa, não somente na definição dos objetivos e na validação dos resultados. Cada domínio tem suas características específicas e, como isso, pode se refletir nos dados e resultados dos métodos de análise utilizados, a participação dos especialistas ocorre em todas as fases da pesquisa. Dessa forma, tratando-se de vigilância em saúde relativa à Covid-19, este projeto contou com a participação dos pesquisadores das Faculdade de Saúde Pública da USP e da Faculdade de Medicina da USP.

Entendimento do problema

Esse ciclo pode ser realizado em múltiplas execuções, tem fases bem definidas, iniciadas pela definição do problema. Nesse projeto, a pergunta inicial referia-se aos problemas de subnotificação dos casos da Covid-19 no Brasil. No entanto, a cada execução do ciclo, outros experimentos foram realizados, visto que novas perguntas foram geradas ao longo da pesquisa. Dessa forma, esse projeto contou com a execução de quatro experimentos para aplicação de técnicas de análise de dados em postagens do Twitter a respeito da Covid-19.



Fonte: Adaptado de Shcherbakov et al. (2014).

Figura 1 – Ciclo de Ciência de Dados adotado.

Coleta de Dados

• *Período de coleta dos dados*

Foram coletados *tweets* entre os dias 16.3.2020 e 16.5.2020, período em que o Brasil passou de 234 para 233.142 casos confirmados de Covid-19, segundo dados do Ministério da Saúde (2020).

• *Forma de coleta dos dados*

A coleta foi realizada por *script* desenvolvido na linguagem Python, que diariamente coletou dados de postagens no Twitter através da Application Programming Interface (API) disponibilizada pela plataforma. Com ela foi desenvolvido um método automatizado de coleta de dados para busca e *download* dos *tweets*.

• *Filtros utilizados para busca*

A busca dos *tweets* foi executada através de filtros por palavras-chave, listadas na Tabela 1, com termos relacionados a Covid-19, sintomas, medidas e tratamentos. A definição desses termos contou com a colaboração dos pesquisadores da área da saúde. Alguns dos termos listados na Tabela 1 foram adicionados ao filtro ao longo período do estudo, pois passaram a ser citados devido à menção em notícias ou declarações oficiais. Logo, foi esperado que o número total de *tweets* coletados aumentasse ao longo do tempo.

Os termos iniciados com o caracter # representam as *hashtags*, que são marcações utilizadas pelos usuários e que servem como identificação do assunto da postagem. Procurou-se, também, colocar diferentes variações de palavras relacionadas à Covid-19, visto que usuários se referem à doença de diferentes formas. Foram utilizados também, termos relacionados aos sintomas mais comuns, assim como outros relacionados aos tratamentos citados pela mídia ou por órgãos governamentais.

Tabela 1 – Palavras-chave utilizadas

Grupo	Termos Utilizados
Doença	covid19, corona, covid-19, #covid19brasil, #CoronavirusPlantao, coronavirus, covid
Sintomas	febre, tosse, falta de ar, coriza, dor
Tratamentos	cloroquina, hidroxicloroquina, atazanavir, remdesivir, ivermectina, azitromicina
Ações	isolamento, lockdown, quarentena

• *Volume de dados coletados*

Durante o período de coleta, foram armazenados 7.720.408 *tweets,* apenas considerando as postagens no idioma português. Como boa parte dos termos utilizados na busca não são exclusivos de usuários brasileiros do Twitter, era de esperar que fossem coletados *tweets* de usuários de outros países. Para a realização dos experimentos relatados neste artigo foi feita, durante a etapa de pré-processamento dos dados, uma seleção apenas dos *tweets* escritos em português. Essa seleção foi possível pois, dentre os dados retornados pela API do Twitter, há um parâmetro relativo ao idioma da postagem.

Análise de dados

A etapa de análise de dados foi realizada tanto com o uso de métodos estatísticos quanto pela execução de algoritmos de aprendizado de máquina. Para a análise textual do conteúdo dos *tweets*, foram empregadas técnicas de NLP, como Vetorização e cálculo da importância do termo no texto (Term Frequency – Inverse Document Frequency, TF-IDF). O uso dessas técnicas teve como objetivo preparar os dados para execução dos algoritmos de aprendizado de máquina.

Visualização, validação e produção de resultados

Após a execução da análise de dados, os resultados foram apresentados em forma de gráficos e em tabelas, de modo que pudessem ser analisados e discutidos em conjunto com os especialistas do domínio. A última etapa é caracterizada pela geração dos produtos da pesquisa, que podem ser relatórios e pacotes de *software*. Nesse estudo, foram gerados relatórios com os resultados dos experimentos, apresentados ao longo deste artigo.

Resultados e discussão

Descrição geral

Para demonstrar a aplicabilidade da análise de dados de redes sociais para a vigilância em saúde, foram conduzidos quatro experimentos. Inicialmente, foi feita análise exploratória dos dados, com produção de análises estatísticas. Em seguida, foram realizados estudos usando técnicas de NLP e aprendizado de máquina, nos quais o uso de cada técnica estava relacionado ao objetivo de cada experimento.

Cada experimento contou com três etapas: pré-processamento, execução dos algoritmos e pós-processamento. Por ser comum a todos os experimentos, a etapa de pré-processamento é descrita separadamente; as demais são descritas em cada experimento.

Pré-processamento

Essa etapa consiste em preparar os dados para execução das tarefas de análises. Normalmente, essa etapa contém tarefas como seleção de atributos que serão utilizados, tratamento dos casos de dados ausentes, normalização de dados, entre outros.

A busca de *tweets* realizada na API do Twitter retorna, para cada *tweet*, algumas dezenas de atributos. Além do texto e data de criação da postagem, são retornadas informações como: número de vezes que a mensagem foi compartilhada, quantidade de pessoas que marcaram a postagem como favorita, idioma utilizado, entre outras. Além dessas informações, também são retornadas na busca todos os dados relativos ao usuário que postou a mensagem, como número de seguidores, localização, nome, se é uma conta verificada, entre outras.

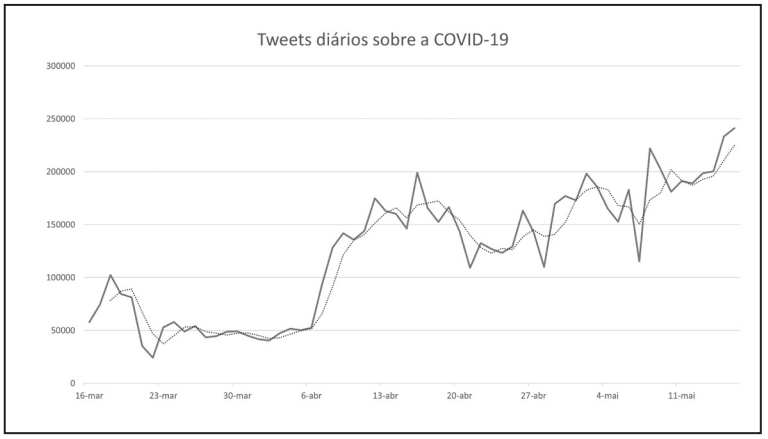
Para os experimentos realizados neste estudo, apenas o atributo relativo ao texto da postagem foi selecionado para a etapa seguinte. Também optou-se por remover os *emojis*, que são figuras normalmente utilizadas em redes sociais para expressar emoções, assim como as *stopwords*, que são palavras que não contribuem para o sentido do texto, como artigos e preposições. Por fim, optou-se por remover os links das postagens, que não eram relevantes para os experimentos planejados.

Experimento 1: Análise exploratória

Foi realizada uma análise exploratória com objetivo de extrair algumas informações básicas, como total de postagens ao longo do tempo, menções a sintomas e formas de tratamento. Na Figura 2, com os dados em escala logarítmica, nota-se aumento na quantidade de postagens relacionadas à Covid-19. Esse fato poderia ser explicado apenas pela inclusão de palavras-chaves no filtro de busca, mas a última alteração nos parâmetros de busca se deu no dia 7 de abril de 2020, o que pode ser verificado pelo salto na linha de tendência. Logo, outros fatores poderiam explicar o aumento do número de postagens a partir do dia 7/4, o que poderia ser verificado através da análise de cada palavra utilizada na busca.

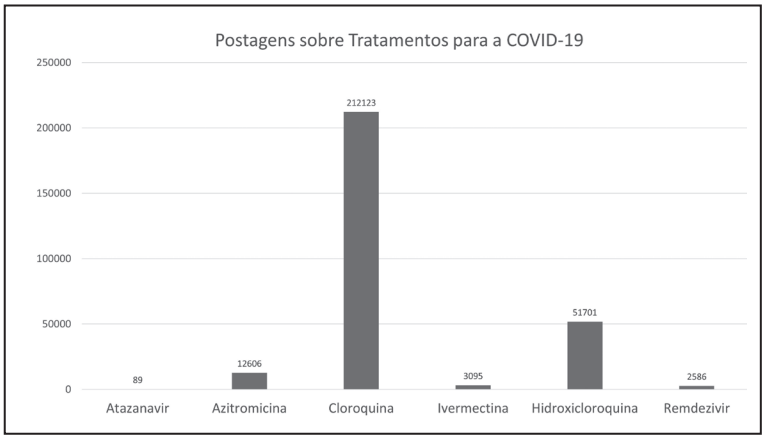
Além de palavras relacionadas ao termo Covid-19, foram executadas coletas de dados utilizando termos relativos aos possíveis tratamentos veiculados pela mídia assim como sintomas mais conhecidos. Em relação às postagens sobre tratamentos, dos seis analisados neste estudo, notou-se predominância das postagens relativas à cloroquina (Figura 3), com referência em cerca de 75% das postagens desse grupo, o que pode ser explicado pela repercussão dos discursos do presidente do Brasil a respeito da cloroquina.

Quando se analisa a quantidade de postagens a respeito da cloroquina ao longo do tempo, nota-se uma tendência no aumento das postagens (Figura 4).



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 2 – Quantidade de postagens sobre a Covid-19, com média móvel de três dias em linha tracejada.

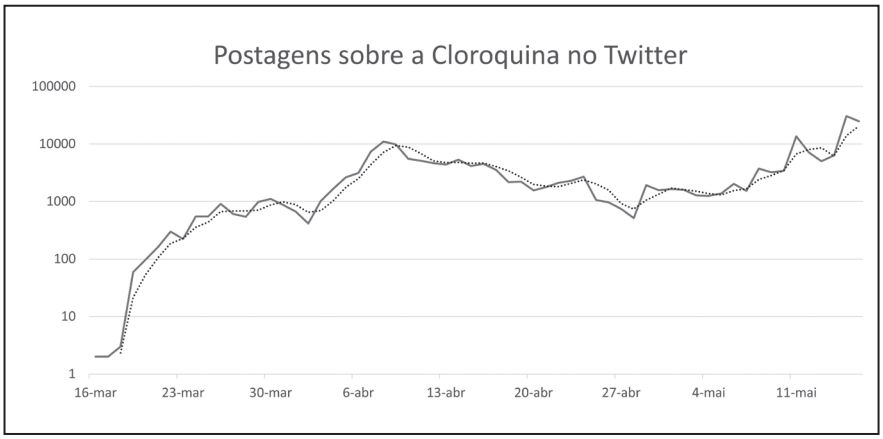


Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 3 – Quantidade de postagens sobre cada tratamento.

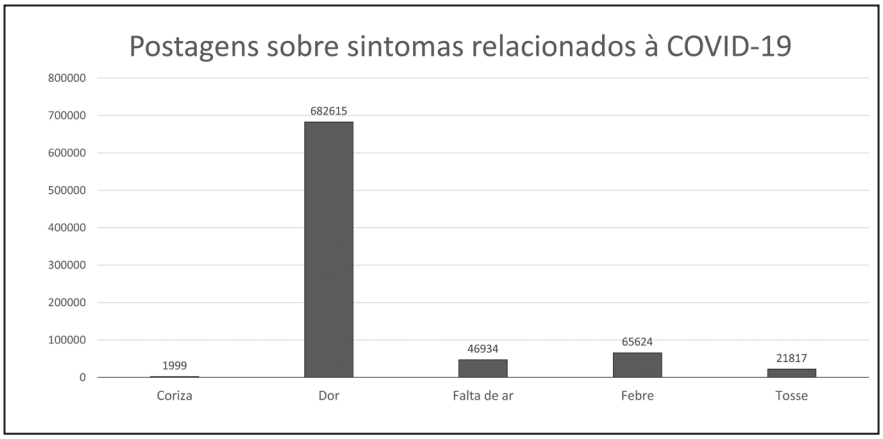
Uma análise futura poderia indicar quais fatores contribuíram para o aumento da discussão sobre o tema incluindo; por exemplo, quais impactos os discursos oficiais tiveram junto à população. Embora o objeto de estudo tenha sido relativo aos tratamentos para a Covid-19, a mesma abordagem poderia ser aplicada para outras ações realizadas.

Buscou-se também realizar uma análise exploratória dos dados em relação aos sintomas mais comuns, com as buscas realizadas com termos relativos a cinco sintomas. Na Figura 5, há predomínio das postagens sobre "dor". Esse sintoma, assim como os outros, não é exclusivo para a Covid-19, o que poderia trazer postagens relativas a outros problemas de saúde. Especialmente no caso de "dor", o uso desse termo não necessariamente significaria um relato de dor física. Logo, análises mais aprofundadas sobre as postagens relativas a sintomas



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 4 – Evolução das postagens sobre cloroquina.



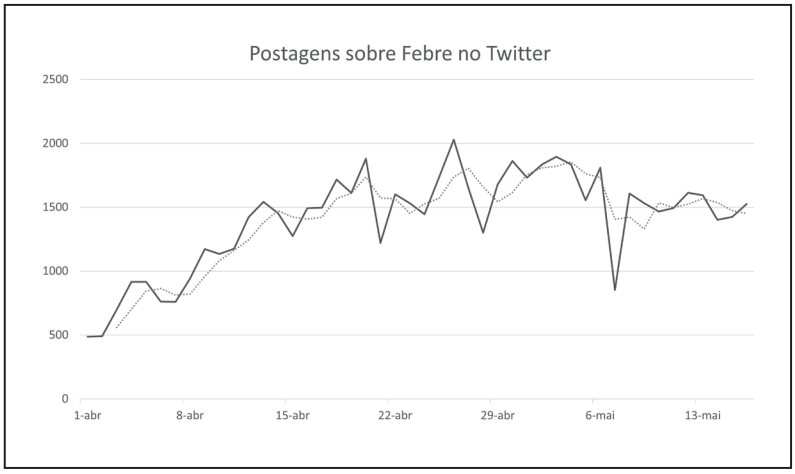
Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 5 – Citações de sintomas no Twitter.

devem considerar aspectos linguísticos, de modo a identificar o que de fato é relato de um sintoma de determinada doença.

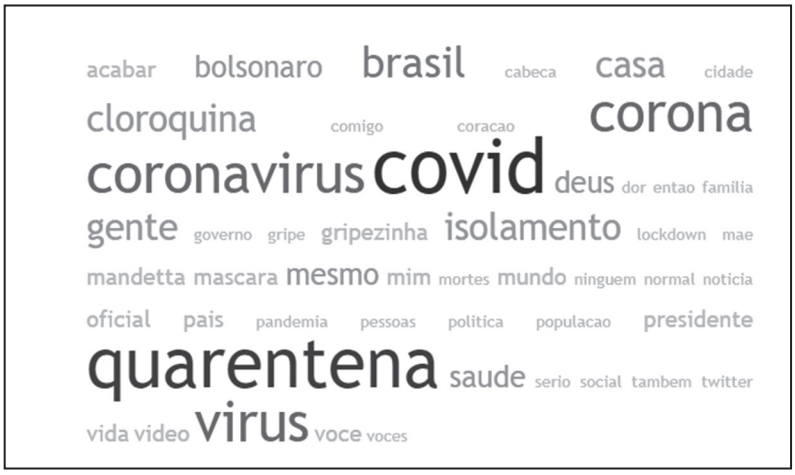
A busca por sintomas nas postagens, no entanto, pode revelar tendências, especialmente se forem consideradas escalas temporal e espacial. O monitoramento constante das postagens por sintomas pode revelar, por exemplo, o surgimento de casos de uma doença em um local antes de se transformar em epidemia, o que permitiria às equipes de saúde tomar as devidas ações com antecedência adequada. Obviamente, a análise espacial depende da disponibilidade dessas informações nos perfis dos usuários.

Mas, em relação à análise dos sintomas em escala temporal, notou-se no gráfico apresentado na Figura 6 um claro aumento no número de postagens relativas à febre. Esse tipo de análise poderia ser usado como indicador de alerta



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 6 – Evolução das postagens sobre febre, com média móvel de três dias na linha tracejada.



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 7 – Nuvem de tags com as cinquenta palavras mais usadas nas postagens.

para a vigilância em saúde e auxiliar os gestores a decidir pela realização de inquéritos em uma determinada região. Com o monitoramento em tempo real desses indicadores, poderiam ser realizadas ações antes de uma epidemia acontecer.

Outra abordagem que pode ser adotada para análise dos dados é o uso de técnicas de visualização computacional, em que se representam informações mediante recursos visuais. Para demonstrar essa abordagem, foi elaborada uma nuvem de *tags*, que é uma representação visual da ocorrência das palavras no texto. Na Figura 7, são representadas as 50 palavras mais citadas nos *tweets* coletados neste estudo; quanto maior o tamanho da palavra maior é a frequência de uso no texto.

Experimento 2: Análise do conteúdo das postagens

Uma aplicação comum de NLP refere-se à análise do conteúdo do texto para extração de significados das mensagens, em que os termos são analisados dentro do contexto em que são utilizados. Uma dessas análises é a extração dos *n-grams*, que são usados para identificar as palavras mais empregadas em relação a um determinado termo. Esse tipo de análise tem aplicações em reconhecimento de fala, tradução de texto e em editores de texto (Abdolahi; Zahedh, 2017).

Para demonstrar a aplicação dessa técnica, foi realizada uma análise dos bigramas (*n-grams* de duas palavras) em relação ao termo cloroquina. A Tabela 2, apresenta os cinco bigramas mais utilizados para cloroquina e, com isso, pode-se identificar os assuntos das principais discussões a respeito da cloroquina nas redes sociais. Notou-se que a maior parte das discussões se refere ao uso desse medicamento e medicamentos associados.

Tabela 2 – Bigramas com termos mais comuns associados à cloroquina

Palavra 1	Palavra 2	Ocorrências
sobre	cloroquina	3827
hidroxicloroquina	azitromicina	3260
tomar	cloroquina	2889
usar	cloroquina	2331
tomou	cloroquina	1711

Tabela 3 – Trigramas com termos mais comuns associados à cloroquina

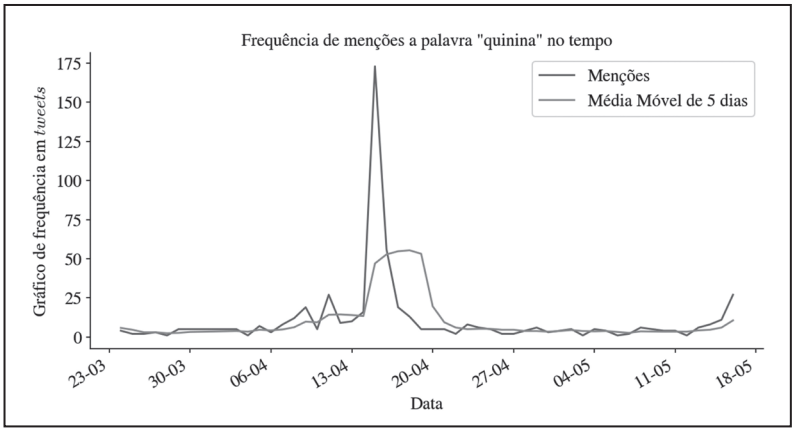
Palavra 1	Palavra 2	Palavra 3	Ocorrências
hidroxicloroquina	azitromicina	zinco	790
efeitos	colaterais	cloroquina	455
cloroquina	tratamento	covid	374
morrido	usando	hidroxicloroquina	341
cloroquina	está	sendo	309

Quando se analisa o uso da palavra cloroquina no contexto de outras duas palavras (trigramas), é possível extrair informações mais detalhadas sobre a discussão a respeito da cloroquina, conforme apresentado na Tabela 3. Através dos dois trigramas mais frequentes, é possível notar que a discussão se concentra a respeito dos outros medicamentos associados à cloroquina e de possveis efeitos colaterais. Esse tipo de informação poderia ser usado, por exemplo, para planejamento pelos gestores da saúde das ações de comunicação.

Experimento 3: Análise de Fake News

Neste experimento, procurou-se identificar notícias falsas relativas à Covid-19 assim como analisar a sua disseminação nas redes sociais. Com esse objetivo, foi feita uma análise da disseminação de uma notícia a respeito de supostos benefícios do uso do/a quinino/a como medida de combate ao coronavírus.

Inicialmente, executou-se uma busca na base de dados pelos *tweets* que contivessem as palavras "quinina" ou "quinino". Após uma análise preliminar dos resultados, adicionou-se uma nova restrição: o *tweet* não poderia conter a palavra "cloro", que estava sendo usada para ironizar essa notícia específica (criando uma nova *fake news* do tipo paródia). Em seguida, analisou-se a repercussão dessa notícia ao longo do tempo (Figura 8).



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 8 – Frequência de menções da palavra "quinina" (e suas variantes) ao longo do tempo.

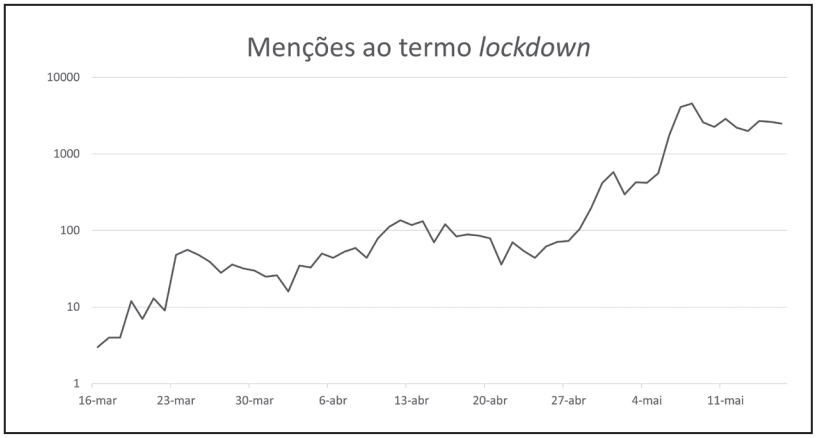
Os resultados indicam que essa notícia teve um pico na semana de 13 de abril, tendo sido muito mencionada nesse período e com grande queda nas semanas posteriores. É importante notar que o aumento na frequência ocorrido na semana do dia 11 de maio é causado não pela notícia em si, mas por sua repercussão. Ou seja, tratou-se de usuários comentando a notícia ao invés de disseminando-a.

Esse fator da repercussão de uma certa notícia na rede é o maior ruído gerado nesta análise. Quando certa notícia falsa surge e tem grande disseminação, ela começa a ser citada em forma de ironia (como por exemplo a associação da quinina com o cloro que foi filtrada durante o experimento), o que representa um desafio para que se tenha noção da real dimensão do alcance da notícia original.

Experimento 4: Análise da opinião sobre políticas públicas

Durante a pandemia da Covid-19, uma das medidas com maior discussão nas redes é aquela referente ao bloqueio total, ou *lockdown*, com posicionamen-

tos favoráveis ou contrários à adoção dessa medida. Foram feitas, nas postagens coletadas neste estudo, 35.467 menções ao termo, com expressivo aumento a partir de maio de 2020, conforme Figura 9. É importante ressaltar que a inclusão do termo *lockdown* nos filtros de coleta de postagens foi feita no início de abril de 2020.



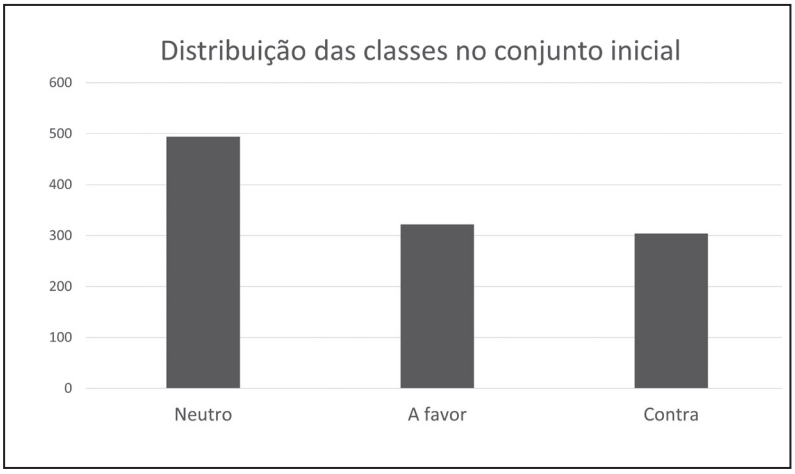
Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 9 – Evolução das menções ao termo *lockdown* (em escala logarítmica).

Uma das técnicas empregadas como medida de opinião sobre determinado assunto é a análise de sentimentos em que, pela execução de algoritmos de aprendizado de máquina, classifica-se cada documento quanto à sentimento do autor (por exemplo: positivo ou negativo). As classes utilizadas podem ser definidas antes da análise ou podem ser aprendidas pelo algoritmo durante a análise. Para demonstração dessa técnica, optou-se por categorizar inicialmente os sentimentos dos usuários quanto à adoção do *lockdown* em três classes: (i) a favor; (ii) neutro ou (iii) contra.

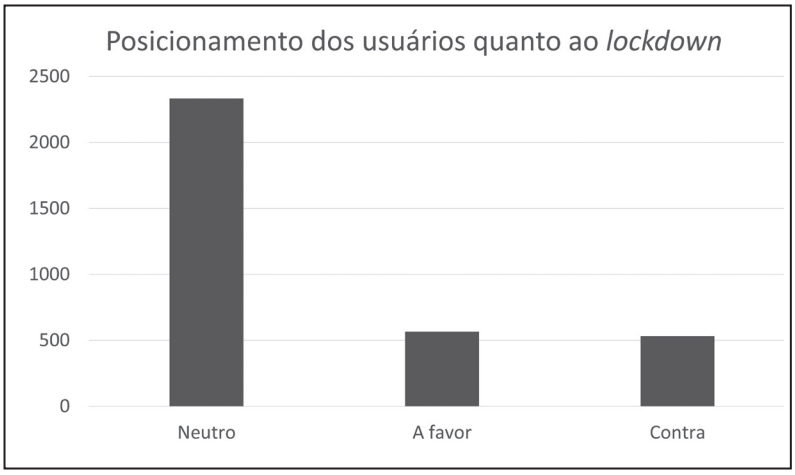
O método utilizado para análise foi o aprendizado supervisionado, ou seja, o algoritmo aprende o modelo a partir de uma espécie de "gabarito" e, a partir do modelo gerado, pode classificar novos conjuntos de dados. Para isso, foram classificadas manualmente 1.120 postagens para compor os conjuntos de treino e teste, com a distribuição das classes listadas na Figura 10.

Na etapa de análise, foram realizadas execuções de diversos algoritmos para encontrar aquele que tivesse a melhor acurácia na avaliação do conjunto de dados de teste a partir do modelo aprendido com os dados do conjunto de treino. Dentre os algoritmos avaliados, o que alcançou melhores resultados foi o Support Vector Machine com acurácia de 0.89. Após essa etapa, aplicou-se o modelo em um conjunto de dados ainda não classificado, com 3.431 postagens com o termo *lockdown*. Conforme gráfico na Figura 11, houve predominância de opiniões classificadas como neutras, enquanto o número de postagens com opiniões a favor ou contra a adoção de *lockdown* foi semelhante.



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 10 – Postagens classificadas manualmente sobre adoção de *lockdown*.



Fonte: Elaboração própria a partir de dados coletados no Twitter.

Figura 11 – Postagens novas classificadas pelo algoritmo.

Dado que nem todas as postagens representam necessariamente um posicionamento do usuário a respeito de um tema, poderiam ser conduzidos outros estudos para identificação de outras classes além das três definidas neste experimento. Para isso, poderiam ser executados métodos de aprendizado não-supervisionado, como os algoritmos de *clustering*, de modo que se identifique: 1) o número ideal de classes; e 2) quais postagens estão identificadas em cada classe. Além do conteúdo das postagens, outras variáveis poderiam ser incluídas na análise, como atributos relativos aos autores das postagens.

Desafios para Análise de Dados de Redes Sociais

O processamento de texto traz alguns desafios adicionais em relação a outros formatos de dados, já que não basta analisar as palavras isoladas, mas também no contexto em que são usadas. As técnicas de NLP são usadas para lidar com esses desafios, por meio de métodos estatísticos e computacionais, com alguns exemplos demonstrados neste artigo. No caso de dados do Twitter, no entanto, existem outros desafios:

- *Textos curtos*: cada postagem no Twitter tem limite máximo de 280 caracteres, o que pode dificultar ainda mais a inferência do significado de uma palavra ou frase. Dependendo do objetivo, apenas um *tweet* pode não ser suficiente para extração de informação útil. Pode-se usar incluir outros atributos na análise, como informações do usuário.
- *Linguagem informal*: nas postagens em redes sociais, também por conta do limite de caracteres, é natural que se faça uso de comunicação informal, com uso de gírias, abreviações e expressões que são características da comunicação via Internet. Incluir esses termos, ou não, na análise vai depender do objeto de estudo. Mas, ressalta-se que a exclusão desses termos pode diminuir ainda mais o conteúdo da postagem a ser analisada, o que pode ter impacto na qualidade do resultado obtido. Esse fato pode ser ainda mais importante em aplicações de análise de sentimento.
- *Uso de imagens*: nas redes sociais é muito comum o uso de imagens associadas às postagens, como os *emojis*. Excluir esses elementos da análise também depende do objeto de estudo visto que, por exemplo, um *emoji* pode ser fundamental para se identificar a percepção do usuário sobre determinado assunto. Estudos que tenham a saúde mental como tema, poderiam incluir esses elementos na análise para identificar com mais precisão o humor dos usuários e, com isso, gerar indicadores do humor da população.
- *Postagens impulsionadas*: é comum o uso de métodos automatizados para impulsionar assuntos de interesse de algum grupo. O uso de *bots* é algo que vem sendo combatido pelas empresas proprietárias das plataformas, mas nota-se ainda um grande uso desses métodos automatizados. Logo, a análise da opinião sobre determinado assunto deve considerar a origem de cada postagem (humano ou *bot*).
- *Veracidade*: a análise epidemiológica de dados de redes sociais também deve ter métodos para identificação da veracidade da mensagem ou, ao menos, um índice do grau de confiabilidade. Caso contrário, os indicadores podem estar baseados em dados incorretos. Por exemplo: nas postagens analisadas neste estudo, identificou-se 33.369 citações da palavra "morreu". Seriam essas ocorrências relativas a óbitos relacionados à Covid-19, cujos dados seriam úteis para lidar com a questão da subnotificação, ou muitas dessas postagens contêm relatos falsos ou

repetidos? Responder essas questões é fundamental para que possa utilizar esses dados na vigilância em saúde.

- *Viés ideológico*: é natural que as pessoas usem as redes sociais para se posicionarem sobre qualquer assunto, como a política. A análise das postagens neste estudo revelou uma grande politização do tema Covid-19. A título de exemplo, das mais de 7 milhões de postagens analisadas, foram feitas 365.331 menções ao presidente do Brasil (cerca de 4,73% do total de postagens). Logo, qualquer estudo epidemiológico a partir de dados de redes sociais poderia considerar que esses dados contêm vieses relacionados à questões ideológicas e ter métodos para identificar postagens com esse teor. Usar ou não essas postagens depende, obviamente, dos objetivos de cada estudo.

Embora existam diversos desafios para análise dos dados de redes sociais para vigilância em saúde, pode-se afirmar que esses desafios representam oportunidades para o desenvolvimento de pesquisas, tanto para desenvolvimento de novos algoritmos mais adequados a esse contexto quanto para a proposição de novas abordagens para análise desses tipos de dados.

Considerações finais

Este estudo teve como objetivo discutir possibilidades de uso de dados de redes sociais como apoio às atividades de vigilância em saúde. Para isso, foi realizado um estudo de diversas aplicações existentes bem como foram conduzidos quatro experimentos para analisar os dados de postagens no Twitter a respeito da Covid-19.

Os resultados mostraram que muitas informações úteis podem ser extraídas de maneira eficiente a partir de métodos computacionais, fornecendo visão em tempo real que pode ser útil nos processos de tomada de decisão. A velocidade e qualidade da decisão podem ser fatores decisivos para o sucesso no combate às doenças. Discutiu-se, também, desafios na análise de dados de redes sociais. No entanto, ao mesmo tempo que podem representar barreiras para inclusão dessa abordagem no cotidiano da vigilância em saúde, esses desafios constituem grandes oportunidades para pesquisa na área.

O caminho para transformar essas oportunidades em resultados com curto tempo de implementação passa necessariamente pelo desenvolvimento de mais pesquisas multidisciplinares, aproximando quem domina as técnicas de análise de quem, de fato, pode falar sobre os dados e problemas da área da saúde.

Nota

1 Este trabalho contou com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (Capes) – Código de Financiamento 001

Referências

ABD-ALRAZAQ, A.A. et al. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, v.132, p.1-7, 2019.

ABDOLAHI, M.; ZAHEDH, M. Sentence matrix normalization using most likely n--grams vector. In: 4TH INTERNATIONAL CONFERENCE ON KNOWLEDGE--BASED ENGINEERING AND INNOVATION (KBEI). Tehran. 2017. 6p.

ACHREKAR, H. et al., Predicting Flu Trends using Twitter Data, In: 2011 IEEE CONFERENCE ON COMPUTER COMMUNICATIONS WORKSHOPS (INFOCOM WKSHPS). Shangai. 2011. 6p.

ARREAZA, A. L. V.; MORAES, J. C. D. Vigilância da saúde: fundamentos, interfaces e tendências. *Ciência & Saúde Coletiva*, v.15, p.2215-28, 2010.

BYRD, K.; MANSUROV, A.; BAYSAL, O. Mining Twitter data for influenza detection and surveillance. In: PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE ENGINEERING IN HEALTHCARE SYSTEMS (SEHS '16). Austin. 2016. 7p.

CHOWDHURY, G. G. Natural language processing. *Annual Review of Information Science and Technology*, v.37, p.51-89, 2003.

CONCOLATO, C. E.; CHEN, L. M. Data science: A new paradigm in the age of big--data science and analytics. *New Mathematics and Natural Computation*, v.13, p.119-43, 2017.

DAI, X.; BIKDASH, M.; MEYER, B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In: SOUTHEASTCON 2017. Charlotte. 2017. 7p.

DHAR, V. Data science and prediction. *Communications of the ACM*, v.56, p.64-73, 2013.

FREITAS, A. S.; SILVA, L. S.; SANDES, S. S. L. *New SIR model used in the projection of Covid 19 cases in Brazil*. Disponível em: <https://www.medrxiv.org/content/10.1101/2020.04.26.20080218v1,2020>.

GARZÓN-ALFONSO, C. C.; RODRÍGUEZ-MARTÍNEZ, M. Twitter Health Surveillance (THS) System. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA. Seattle. 2018. 8p.

JORDAN, S. E. et al. Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*, v.4, p.6, 2019.

MALTA, D. C. et al. Inquéritos Nacionais de Saúde: experiência acumulada e proposta para o inquérito de saúde brasileiro. *Revista Brasileira de Epidemiologia*, v.11, p.159-67, 2008.

MEDFORD, R. J. et al. An "Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the Covid-19 Outbreak. Disponível em: <https://www.medrxiv.org/content/10.1101/2020.04.03.20052936v1>. 2020.

MINISTÉRIO DA SAÚDE. Painel Coronavírus. Disponível em: <https://covid.saude.gov.br/> . Acesso em: 16 maio 2020.

PONS, E. et al. Natural language processing in radiology: a systematic review. *Radiology*, v.279, p.329-43, 2016.

PRYSS, R. et al. Using Chatbots to Support Medical and Psychological Treatment Procedures: Challenges, Opportunities, Technologies, Reference Architecture. In: BAUMEISTER, H.; MONTAG, C. *Digital Phenotyping and Mobile Sensing*. S.l.: s.n., Springer 2019. p.249-60.

RIBEIRO, L. C.; BERNARDES, A.T. Estimate of underreporting of Covid-19 in Brazil by Acute Respiratory Syndrome hospitalization reports. Nota Técnica. Cedeplar. Belo Horizonte: UFMG, 2020.

SHCHERBAKOV, M. et al. Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development. *Knowledge-Based Software Engineering*, v.466, p.708-16, 2014.

SMUTNY, P.; SCHREIBEROVA, P. Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, v.151, p.1-11 , 2020.

STATISTA. Twitter - Statistics & Facts. Disponível em: <https://www.statista.com/topics/737/twitter/>. Acesso em: 18 mai. 2020.

TA, V. et al. User Experiences of Social Support from Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of Medical Internet Research*, v.22, p.e16235, 2020.

TEIXEIRA, C. F.; PAIM, J. S.; VILASBÔAS, A. L. SUS, modelos assistenciais e vigilância da saúde. In: ROZENFELD, S. *Fundamentos da vigilância sanitária*. Rio de Janeiro: Fiocruz, 2000. p.49-60.

TWITTER. Q1' 2020 Shareholder Letter. Disponível em: <https://investor.twitterinc.com/home/default.aspx>. Acesso em: 18 mai. 2020.

VAIDYAM, A. N. et al. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, v.64, p.456-64, 2019.

YANG, K. C.; TORRES-LUGO, C.; MENCZER, F. *Prevalence of Low-Credibility Information on Twitter During the Covid-19 Outbreak*. 2020. Disponível em: <https://arxiv.org/abs/2004.14484>.

RESUMO – O grande volume de dados gerados em redes sociais é usado por empresas para monitoramento das opiniões do público sobre seus produtos e serviços. Na área da Saúde, esses dados podem conter informações também aplicáveis na vigilância, como na avaliação do impacto de políticas públicas ou na identificação de *fake news*. Este trabalho apresenta resultados de estudos demonstrando como a análise de dados de redes sociais pode ser utilizada nas atividades de vigilância, tendo como estudo de caso a pandemia da Covid-19. Foi utilizada uma abordagem baseada em Ciência de Dados, com extração de informações através de algoritmos de aprendizado de máquina. Os resultados indicam que essa abordagem pode revelar importantes informações para as atividades de vigilância, trazendo uma visão em tempo real de aspectos relacionados à pandemia.

PALAVRAS-CHAVE: Vigilância em saúde, Redes sociais, Aprendizado de máquina, Covid-19.

ABSTRACT – The large volume of data generated on social networks is used by companies to monitor public opinion about their products and services. These data may contain useful information for health surveillance, such as in assessing the impact of public poli-

cies or identifying fake news. This work presents results of studies that demonstrate how analysis of data from social networks may be applied to surveillance activities, using the covid-19 pandemic as a case study. An approach based on data science was used, with information extracted through machine learning algorithms. Results indicate that this approach can reveal useful information for surveillance activities, providing a real-time view of aspects related to the pandemic.

KEYWORDS: Health surveillance, Social networks, Machine learning, Covid-19.

Fernando Xavier é doutorando em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo, pesquisador do Grupo de Estudos em Saúde Planetária do Instituto de Estudos Avançados da USP. @ – fxavier@usp.br / https://orcid.org/0000-0001-5797-7339.

João Rodrigo W. Olenscki é aluno de graduação em Engenharia Mecatrônica pela Escola Politécnica da Universidade de São Paulo (USP). @ – joao.olenscki@usp.br / https://orcid.org/0000-0002-9948-541X.

Andre Luis Acosta é doutor em Ecologia pela Universidade de São Paulo, post-doc na Faculdade de Saúde Pública (USP), pesquisador no Centro Brasil-Reino Unido de Descoberta, Diagnóstico, Genômica e Epidemiologia de Arbovírus. Membro do Grupo de Estudos em Saúde Planetária do Instituto de Estudos Avançados da USP. @ – andreluisacosta@gmail.com / https://orcid.org/0000-0002-4244-9637.

Maria Anice Mureb Sallum é doutora em Saúde Pública e professora do Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo. Participa do projeto Centro Brasil-Reino Unido de Descoberta, Diagnóstico, Genômica e Epidemiologia de Arbovírus. Membro do Grupo de Estudos em Saúde Planetária do Instituto de Estudos Avançados da USP. @ – masallum@usp.br / http://orcid.org/0000-0002-7051-2891.

Antonio Mauro Saraiva é doutor em Engenharia de Computação e Professor da Escola Politécnica da Universidade de São Paulo. Coordenador do Grupo de Estudos em Saúde Planetária do Instituto de Estudos Avançados da USP. @ – saraiva@usp.br / https://orcid.org/0000-0003-2283-1123.

Recebido em 25.5.2020 e aceito em 19.6.2020.

[I, V] Escola Politécnica, Universidade São Paulo, São Paulo, Brasil.

[II] Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brasil.

[III, IV] Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, Brasil.