

INTERAÇÕES NAS REDES SOCIAIS E A CRIAÇÃO DE ALERTAS PARA SAÚDE PÚBLICA: UMA REVISÃO SISTEMÁTICA DE LITERATURA.

RESUMO

Este artigo desenvolve uma revisão sistemática da literatura, com objetivo de identificar como o uso de dados de rede social é abordado pela literatura no mapeamento de potenciais ameaças sanitárias.

Com a popularização do acesso à internet, as redes sociais estão entre as principais plataformas em número de usuários. Dados estimados de julho de 2021 indicam que a principal rede social, o Facebook, tem cerca de 2,5 bilhões de usuários. Estima-se que o Twitter tenha cerca de 400 milhões de usuários ativos, com quase 17,2 milhões de usuários no Brasil (Statista, 2021). Dentre os aplicativos disponíveis, o Youtube lidera a lista dos mais usados entre os internautas brasileiros, seguido por Whatsapp, Facebook, Instagram, Facebook Messenger, Twitter. Outros aplicativos estão também presentes, porém com taxa de uso menor do que 50% dos usuários, segundo dados do relatório Digital 2021 (We_are_Social, 2021).

No que se refere à saúde pública, a compreensão do uso de redes sociais para divulgação de informações sobre saúde é relevante sob os seguintes aspectos: fontes oficiais podem divulgar alertas e orientações à população em seus perfis sobre temas pertinentes, porém o usuário comum pode também interagir e expor informações importantes, não apenas compartilhando e comentando sobre publicações oficiais, mas expondo condições clínicas, que podem ajudar as autoridades a identificarem de forma precoce o início de focos de problemas sanitários, sejam em nível localizado ou mesmo epidemias e pandemias.

Neste contexto, a Revisão Sistemática da Literatura (RSL) apresentada neste artigo teve como objetivo compreender como o uso de dados da rede social Twitter vem sendo abordado pela literatura no mapeamento de potenciais ameaças sanitárias. Assim, a revisão buscou responder a seguinte pergunta: Como está sendo tratado na literatura o uso de dados das redes sociais Twitter no contexto de monitoramento e detecção de eventos na saúde coletiva?

METODOLOGIA

Para responder à questão de pesquisa, foi elaborada uma revisão sistemática da literatura (RSL) baseando-se no processo descrito por Kitchenham (2004) dividido em cinco passos: Definição dos Objetivos da Pesquisa, Estratégia da Pesquisa, Triagem dos Documentos, Extração dos dados e Classificação.

A pesquisa teórica foi realizada nas bases de dados da Scopus, Periódicos Capes, Science Direct, Web of Science e Google Scholar. As palavras-chave utilizadas na busca foram: (*“Health Surveillance”*) AND (*“Social Network”* OR *“Social Media”*) AND (*“Machine Learning”* OR *“Big Data”* OR *“Artificial Intelligence”* OR *“predictive model”*) AND (*“Twitter”*). A rede social Twitter foi escolhida como recorte da pesquisa, por ser a única que disponibiliza acesso aos dados de interação entre seus usuários para fins de pesquisa, por meio de uma API oferecida pela empresa. As demais empresas controladoras das redes sociais não oferecem acesso aos dados. Foram filtrados documentos escritos nos idiomas inglês, português e espanhol. A busca foi realizada em 17 de maio de 2021 resultando em 3.212 documentos (n= 3.212, Fase 1).

O foco desta revisão sistemática é obter um mapa com os principais problemas e conceitos abordados no mapeamento de ameaças sanitárias utilizando a rede social Twitter, bem como as técnicas de coleta e análise de resultados utilizadas. Para atingir este objetivo utilizou-se uma planilha em Excel para consolidar as informações de cada artigo desta revisão. Foram tabuladas as informações contendo o nome do artigo, autores, doença estudada, data do estudo, abordagem de dados georreferenciados, país do estudo, métodos estatísticos ou de aprendizado de máquinas utilizados e o principal objetivo do estudo.

Como a pesquisa tem por objetivo abordar acontecimentos recentes como a pandemia da COVID-19, foi estabelecido um filtro de data com artigos a partir de 2018 permitindo também analisar como este tema estava sendo tratado antes e durante essa pandemia. Esta aplicação de filtro de datas resultou em 1.845 documentos (n= 1.845, Fase 2). A partir da aplicação dos filtros por palavras-chave e por data de publicação, realizou-se a leitura dos títulos dos documentos avaliando a aderência ao objetivo desta RSL, o que resultou em 457 documentos (Fase 3).

Como o objetivo desta revisão sistemática é saber como a rede social Twitter vem sendo utilizada publicar informações e dados referentes à saúde coletiva, foram definidos alguns atributos que deveriam ser abordados nos artigos selecionados (critérios de qualidade). A identificação geográfica (geolocalização) é um fator relevante para definir áreas de alto risco de contaminação, a identificação por regiões e mapas foi utilizado como critério de inclusão.

Foram excluídos artigos que não incluíam uma visualização espacial e temporal das informações. Com o objetivo era entender os principais modelos de análise dos dados utilizados, foram excluídos artigos que não descreviam os algoritmos de Machine Learning utilizados. Os estudos deveriam utilizar dados da rede social Twitter e descrever detalhadamente como os dados foram obtidos e processados.

Após a seleção pela leitura dos títulos, foi avaliada a partir da leitura do resumo (*abstract*) e da conclusão de cada artigo a aderência aos objetivos da RSL, restando na triagem 60 documentos (Fase 4). Realizou-se, então, nova triagem com a leitura atenta e detalhada de cada artigo, resultando em 23 artigos (Fase 5). A Tabela 1 apresenta as fases da revisão.

Tabela 1 – Fases da RSL

Fase	Critério	Número Artigos
1	Bases = Scopus, Periódicos Capes, Science Direct, Web of Science e Google Scholar Palavras chaves = (“Health Surveillance”) AND ("Social Network" OR "Social Media") AND ("Machine Learning" OR "Big Data" OR "Artificial Intelligence" OR "predictive model") AND (“Twitter”) Idiomas = Inglês, Português e Espanhol	N = 3.212
2	Data publicação: janeiro 2018 a junho 2021	N = 1.845
3	Leitura dos títulos dos documentos relacionados ao objetivo de pesquisa. Foram excluídos artigos que não incluíam visualização espacial e temporal das informações, não descreviam os algoritmos de Machine Learning utilizados. Foram considerados somente os artigos que utilizaram a rede social Twitter e que descreviam como os dados foram obtidos e processados.	N = 457
4	Leitura dos resumos e conclusões dos documentos relacionados ao objetivo de pesquisa	N = 60
5	Leitura detalhada dos documentos e seleção dos artigos para referências à pesquisa	N = 23

Fonte: Elaborada pelos autores.

RESULTADOS

A triagem dos documentos resultou em 23 artigos aderentes ao objetivo desta revisão sistemática. Uma abordagem que pode ser adotada para análise dos artigos 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 palavras, que é uma

representação visual da ocorrência das palavras nos resumos (abstract) dos artigos selecionados. Na Figura 1, são representadas as 50 palavras mais citadas, quanto maior o tamanho da palavra maior é a frequência de uso.



Figura 1 – Nuvem de palavras dos resumos dos artigos da RSL

Fonte: Dos autores

A Tabela 2 traz as principais informações obtidas durante a leitura dos textos.

Tabela 2 - Objetivo dos artigos selecionados na RSL

# Artigo	Primeiro Autor	Ano	Objetivo
1	Edo-Osagie et al.	2019	Propõe modelo supervisionado para classificação de <i>tweets</i> sintomáticos e filtra a relevância.
2	Pruss et al.	2019	Examina com modelo LDA as discussões no Twitter sobre a epidemia de Zikavirus em 2015.
3	Nguyen, H	2021	Desenvolve índice de saúde da população usando Twitter e modelo de redes neurais convolucionais.
4	Hassan Zadeh, Amir	2019	Apresenta modelo de acompanhamento de Influenza utilizando redes sociais e faz correlação temporal.
5	Rashid & Wang	2021	Introdução da ferramenta CovidSens que faz uso das redes sociais para alertas sobre COVID.
6	Souza et al.	2019	Analisa e detecta pontos de risco de Dengue em 2 cidades do Brasil.
7	Martínez, Pascual	2020	Apresenta forma de geolocalização dos tweets com Processamento de Linguagem Natural.
8	Heth et al.	2018	Análise correlação (Pearson) da evolução dos casos semanais de influenza com <i>tweets</i> .

9	Li, Lingyao	2021	Desenvolve criação de um sinal de alerta por estado dos EUA usando PLN e ML.
10	AlAgha	2021	Utiliza modelo não supervisionado para análise de sentimento e classificação por tema dos tweets sobre COVID-19.
11	Adriani, M	2020	Desenvolve <i>framework</i> para identificar emergências sanitárias pelo Twitter e gera informações em mapas na Indonésia.
12	Mackey T,	2020	Usa ML não supervisionado (BTM) e desenvolve sinais de casos positivos de Covid durante o início da pandemia.
13	Cuomo et al.	2021	Desenvolve modelo com SVM para mapa de risco de contaminação de COVID-19.
14	Marsri, S	2019	Desenvolve 2 modelos calibrados por auto-regressão usando Twitter para estimar casos com 1 semana de antecedência.
15	Fazeli et al., 2021	2021	Apresenta um <i>framework</i> para coleta, análise e mapeamento de conteúdos sobre a pandemia de COVID-19.
16	Euzebio, C	2020	Desenvolve algoritmo que analisa uma pequena amostra do Twitter para monitorar Dengue em Ribeirão Preto.
17	Spurlock & Elgazzar	2020	Desenvolve aplicação que agrupa usuários por propensão a contaminação de COVID-19 de acordo com sua rede de contatos nas redes sociais.
18	Jain & Cherikkallil	2018	Apresenta ferramenta Mendinsights que transforma <i>tweets</i> em informações para ações na saúde.
19	Sidana et al.	2018	Desenvolve 2 modelos para análise temporal de possíveis doenças analisando informações das redes sociais.
20	Şerban et al.	2019	Descreve <i>software</i> criado para monitorar propagação de doenças pelo Twitter.
21	Z. Li et al.	2020	Desenvolve modelo usando redes neurais com dados das redes sociais e dados socioeconômicos.
22	Tufts et al.	2018	Análise a correlação da incidência de 14 doenças na Pensilvânia com menções no Twitter no período de 2012 a 2015.
23	Xavier et al.	2020	Faz análises exploratórias de dados do Twitter e demonstra o potencial das Redes Sociais como ferramenta de vigilância sanitária.

Fonte: Elaborado pelos autores.

Nota-se que existem estudos com dados de redes sociais para várias doenças. Na maioria (n=16) é analisada uma única doença por artigo, sendo 9 artigos sobre a COVID-19, seguidos por Dengue e Influenza com 3 estudos cada. Os demais artigos (n=7) são de estudos incluindo mais de uma doença. A maior parte dos artigos analisados foi publicada a partir de 2020, o que coincide com o início da Pandemia da COVID-19 que trouxe discussões sobre a necessidade de se criar mecanismos de identificação precoce de doenças.

Para a apresentação dos dados, os artigos utilizaram ferramentas capazes de apresentar os dados em mapas, com variações de cores para diferenciar e evidenciar regiões com maiores riscos, demonstrando as possibilidades de entender os dados de forma rápida.

Com relação aos modelos utilizados para análise dos dados obtidos do Twitter há uma divisão bem próxima entre modelos supervisionados e não supervisionados. É importante esclarecer que os dados advindos da rede social são classificados como não estruturados, ou seja, contemplam uma mensagem de texto, identificação do usuário e informações complementares (curtidas, compartilhamentos e réplicas). Pela natureza dos dados, para sua manipulação e processamento, em geral são adequadas técnicas de mineração de texto e machine learning, por isso, a presença desse tipo de algoritmo nos artigos analisados. O exemplo apresentado na Figura 2, do trabalho de Rashid & Wang, 2021 traz o tipo de mensagem postada pelos usuários, que após processamento gera alertas de saúde.



Figura 2. Exemplos de postagem.

Fonte: (Rashid & Wang, 2021)

Os modelos supervisionados, também conhecidos como modelos de aprendizado via exemplos, são modelos que através e uma amostra dos dados contendo os corretos dados de saída são usados como base para o treino do modelo, com estas amostras o algoritmo aprende comparando as saídas oferecidas com os dados de entrada. Uma vez que o modelo passe pela

etapa de treino, são oferecidos novos dados de entrada que o modelo automaticamente gera a variável de saída (Alzubi et al., 2018)

Nos modelos não supervisionados, os dados de treino não possuem identificação prévia do resultado. O modelo reconhece os padrões que não são facilmente identificados nos dados e definem regras para estes padrões. Esta técnica é bastante utilizada para classificar dados por categoria quando estas classificações não são conhecidas (Alzubi et al., 2018).

Dentre os 23 artigos selecionados nesta revisão sistemática, os modelos supervisionados foram utilizados em 8 artigos, sendo que o algoritmo *Support Vector Machine* (SVM) foi utilizado em todos os casos, e em 2 estudos houve a utilização também de outros modelos para fazer comparações de performance. Já os modelos de classificação não supervisionados foram utilizados em 10 estudos, sendo que a maioria utilizou *Latent Dirichlet allocation* (LDA) (n=5) seguido por *K-means* e *Biterm Topic Models for Short Text* (BTM), ambos com 2 artigos. A Tabela 3 traz as informações de análise observadas nos artigos.

Tabela 3 – Matriz de visualização dos artigos

Artigo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Doença: Covid					x				x	x		x	x		x		x				x		x
Geral			x															x	x			x	
Dengue						x					x					x							
Zika		x												x									
Influenza				x				x			x												
Malária											x												
Asma	x																						
Diarreia											x												
Elefantíase											x												
Região Geográfica																							
País		x								x							x		x				
Estado			x						x					x								x	
Cidade	x			x	x	x					x	x	x		x	x		x			x		
Fonte Comparação																							
Google Trends									x														
Governo	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x			x			x	x
Visualização no Mapa																							
Ponto		x	x						x			x		x							x	x	
Polígono				x	x	x					x		x		x			x	x				
Modelo Estatístico																							
TF-IDF									x											x			x
Supervisionado																							
LR	x								x														
SVM	x			x					x		x		x					x		x			x
Decision Tree	x								x														
RandomForest																		x					
NaiveBayes	x								x									x		x			
StochasticGradiente Descent																		x					

Artigo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
GradientBoosting																		x					
FFNN																		x					
Não Supervisionado																							
LDA		x								x					x				x			x	
K-Means											x						x						
ATAM																			x				
BTM												x	x										
Redes Neurais			x																				
CNN																				x	x		
LSTM																				x	x		
Contagem palavras							x	x								x							
LCA						x																	
K-S				x																			
Moran (Correlação)				x																			
Regressão Linear														x									
Pearson R				x				x	x			x		x		x							

Fonte: Elaborada pelos autores.

Nota: Term Frequency Inverse Document Frequency (TF-IDF); Logistic Regression (LR); Feed Forward Neural Network (FFNN); Architecture tradeoff analysis method (ATAM); Convolutional Neural Network (CNN); Long short-term memory (LSTM); Life Cycle Assessment (LCA); Kolmogorov-Smirnov (K-S)

Redes Sociais e situações de calamidade e epidemias

As informações disponíveis nas redes sociais são terreno fértil para a criação de novas ideias para gerar sistemas de alertas precoces em vários cenários. O entendimento das informações *online* criadas em um determinado lugar e horário são subsídios que podem ser incorporados em sistemas de alertas precoces (L. Li et al., 2021).

Na literatura é possível encontrar alguns trabalhos relacionados com a utilização de redes sociais em situações de calamidade e epidemia, como os apresentados a seguir.

Por meio de análise espaço-temporal das informações georreferenciadas, o estudo apresentado por Hassan Zadeh et al. (2019), sobre o uso de dados georreferenciados das mídias sociais monitora as informações sobre o vírus influenza. Esse estudo mostrou uma correlação relevante entre a frequência das postagens relacionadas ao vírus e a quantidade de pacientes reportados oficialmente nos Estados Unidos.

Masri et al. (2019) analisam informações do Twitter durante o surto de Zika Vírus em 2016, apresentando dois modelos de regressão que foram calibrados para estimar a quantidade de casos na semana seguinte.

No Brasil, Euzebio et al. (2020) apresentaram um estudo sobre o monitoramento da dengue por meio de postagens no Twitter na cidade de Ribeirão Preto. Nesse estudo, observou-se que o número de comentários no Twitter sobre a dengue acompanha o crescimento de números de

casos oficiais da doença indicando a existência de uma significativa correlação entre os rumores sobre a dengue e o aumento do número de casos notificados.

Adriani et al. (2015) estudaram a possibilidade de detectar a ocorrência das principais doenças na Indonésia a partir da classificação de *posts* das redes Sociais. Nesse estudo abordam-se formas de identificar o tipo da doença a partir dos *tweets* e como pode ser visualizada em mapas a disseminação das doenças.

Li et al. (2021) exploram o potencial do uso dos dados da rede social Twitter para melhorar os alertas sobre a pandemia do COVID-19, o trabalho conduz um estudo investigando mais de 14 milhões de *tweets* postados de 01 janeiro de 2020 a 10 março de 2020. Com o uso de processamento de linguagem natural e técnicas de *machine learning* é criado um sinal de alerta que, segundo os autores, poderia antecipar a detecção do risco em 16 dias.

Dados Georreferenciados no Twitter

Análises automáticas dos dados gerados pelos usuários da redes sociais envolvem a extração das coordenadas geográficas dos posts para identificação de local com a possibilidade de plotar em um sistema de geolocalização em mapas (Martínez & Pascual, 2020)

Diversos estudos utilizando o Twitter tem enfatizado o uso de posts georreferenciados com a localização via GPS fornecida pelos *smartphones* (Dredze et al., 2013). Associar os comentários ou tweets sobre diversos assuntos ao local de origem dos usuários do Twitter possibilita efetuar análises espaciais interessantes.

No Twitter, a informação da localidade pode ser inserida manualmente no perfil do usuário ou por meio da habilitação do dispositivo GPS, o qual fornece as coordenadas geográficas latitude e longitude. Entretanto, Dredze et al. (2013) relatam que aproximadamente 1,2% dos *tweets* possuem coordenadas de localização. No mesmo trabalho os autores apresentam outras formas primárias de se obter informações de localização, utilizando o campo de descrição da localidade dos usuários (cidade ou país), sendo possível incrementar a informação de localização para 22,3% dos *posts*.

Na literatura, diferentes abordagens vêm sendo utilizadas com o objetivo de identificar a localidade do usuário do Twitter. Martínez e Pascual (2020) fazem uso de modelos baseados no conteúdo dos comentários que permite extrair nomes de localidades ou frases associadas às localidades. No mesmo estudo é extraída a localidade a partir dos comentários e por meio da utilização de processamento de linguagem natural, em que as palavras são classificadas em

sujeito, predicado e objeto. Assim, o conteúdo do objeto permite extrair nomes de localidades e a localidade também pode ser extraída do texto do comentário utilizando-se como fonte de referência a base de dados geográfica GeoNames.

Processamento de Linguagem Natural (PLN)

O advento das redes sociais proporcionou às pessoas compartilhar seus próprios conteúdos, ideias, opiniões em tempo real, virtualmente para milhões de pessoas conectadas. Entretanto, este volume enorme e crescente de informações, em sua maioria, é não estruturado devido à especificidade de ser produzida para o consumo humano e por consequência não é diretamente processada por máquinas. O processamento de linguagem natural (PLN) é área de técnicas computacionais que estuda as análises automáticas e representação da linguagem humana pelos computadores (Cambria & White, 2014).

No contexto relacionado à COVID-19, Li et al. (2021) analisaram manualmente 5.000 comentários do Twitter classificando os *tweets* contendo sinais de propagação da doença. Verificou-se que 13,7% dos *posts* tinham informações contendo tais sinais. Utilizando a classificação desses 5.000 *tweets* como dados de treinamento, aliados a técnicas de *Machine Learning* e Processamento de Linguagem Natural os pesquisadores desenvolveram um algoritmo que automaticamente analisou 14,7 milhões de *tweets* coletados de 20 janeiro a 10 de março de 2020 demonstrando o potencial que as opiniões públicas via redes sociais têm de gerar alertas precoces.

Em outro estudo relacionado à COVID-19, Mackey et al. (2020) coletaram 4,5 milhões de *tweets* durante 17 dias em março de 2020. Neste estudo foi possível também identificar comentários relacionados a sintomas, resultados de testes laboratoriais ou outros temas relacionados ao contato com o vírus pelos usuários.

Analisando diversos estudos, há evidências de que existem oportunidades de pesquisa na análise de dados do Twitter durante a pandemia de COVID-19 no território brasileiro com aplicação de Processamento de Linguagem Natural. Além disso, por meio de métodos computacionais, essas informações podem ser acessadas em tempo real pelos gestores e profissionais da saúde (Xavier et al., 2020). 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. Algumas possíveis aplicações com Processamento de Linguagem aplicada a redes sociais podem incluir:

- Análise da opinião sobre medidas adotadas (análise de sentimentos);

- Avaliação do impacto das estratégias de comunicação;
- Identificação de possíveis sintomas relacionados às doenças e antecipação de surtos;
- Identificação e avaliação do impacto de *fake news*.

DISCUSSÕES

Embora ainda relativamente recente, o uso de dados de interações em redes sociais tem sido crescentemente explorado nos estudos acadêmicos e nas organizações, reconhecendo-se a riqueza desses dados gerados de forma espontânea pelo usuário, em tempo real e grande volume. A criação de artefatos que consigam coletar, manipular e gerar relatórios gerenciais, tem grande potencial no apoio à gestão pública e privada, não apenas na identificação de problemas em curso, mas na identificação de condições favoráveis ao surgimento de problemas, no caso abordado por esta revisão, surtos de doenças, epidemias e pandemias.

As redes sociais podem ser encaradas, por empresas e governos, como ferramentas de colaboração e co-criação, porém, para agregar valor, é necessário criar uma estratégia de atuação nas redes, monitoramento de interações e captura e tratamento desses dados, que requer uma mudança de paradigma dessas organizações, especialmente aquelas não-nativas digitais (caso das organizações em setores tradicionais da economia e gestões públicas).

Os estudos que compuseram esta revisão evidenciaram diferentes algoritmos analíticos possíveis para a modelagem de dados não estruturados advindos de interações em redes sociais, baseando-se todos eles em postagens textuais. Demonstraram ser possível criar alertas para riscos em saúde, em diferentes doenças. Os algoritmos utilizados contemplam, em geral aprendizado de máquina, com processamento opaco (black-box), o que não permite identificar prontamente como foram criadas as regras de classificação dos dados.

Nesse sentido, algumas lacunas puderam ser identificadas, carecendo de investigação: como atestar a veracidade do conteúdo da mensagem ou como identificar mensagens falsas ou postadas por robôs? Como se certificar de que os dados, não-estruturados por natureza, não carregam padrões discriminatórios que são propagados pelos modelos opacos? Como incluir dados de diferentes naturezas como imagens, áudio e vídeo, comuns nas redes sociais, no pool de dados a serem analisados? Como identificar elementos da emoção humana nas postagens, por exemplo, ironia e sarcasmo? Os dados gerados em redes sociais são de propriedades das organizações que as controlam, podendo permitir ou não acesso a eles por empresas e governos; como mitigar esse risco? Em específico, na gestão pública, quais as barreiras legais, culturais e

de infra-estrutura para que as redes sociais possam ser incorporadas como fonte de informação relevante à tomada de decisão?

Desta forma, constata-se que os estudos apresentam abordagens interessantes sobre o uso de dados de redes sociais para a criação de alertas em saúde, sendo esses métodos, passíveis de replicação em outras áreas, contudo, há ainda muitas perguntas que precisam ser respondidas, tanto no nível técnico, quanto no nível ético de sua utilização.

CONCLUSÕES

Esta revisão sistemática traz uma importante contribuição na análise da utilização dos dados da rede social Twitter no contexto de monitoramento e alertas precoces de emergências sanitárias. Neste trabalho é apresentado uma visão analítica dos principais achados de como o tema vem sendo tratado na literatura, incluindo as técnicas de análise de dados utilizadas, e as principais aplicações. Também são analisados como as aplicações podem trazer informações úteis por localidade (georreferenciadas). Estas informações podem ser úteis para o desenvolvimento de ferramentas governamentais e particulares para monitoramento e prevenções alertas precoces. O uso dos dados do Twitter em conjunto com técnicas de Processamento de Linguagem Natural (PLN) vem sendo explorados em diversos estudos e pesquisas. A pandemia do COVID-19 e maior interesse nestas técnicas podem gerar maiores interesses dos estudiosos e ampliar os conhecimentos neste campo de estudo.

Em futuros estudos é indicado a criação de um “framework” robusto incluindo a utilização ética das informações para possibilitar o uso de informações sensíveis no desenvolvimento de sistemas de utilização massiva tanto pelos órgãos públicos como para organizações privadas. Este estudo também pode ser ampliado analisando outras plataformas de redes sociais com maior crescimento possibilitando uma maior generalização dos conceitos explorados. As metodologias aplicadas na extração e análise de dados do Twitter ou outra rede social podem ser extrapoladas a outras situações, por exemplo, para a gestão de marcas de empresas privadas ou dentro da gestão pública, para monitoramento de segurança, zeladoria, entre outros.

REFERÊNCIAS

- Adriani, M., Azzahro, F., & Hidayanto, A. (2015). Disease surveillance in Indonesia through Twitter posts. *Journal of Applied Research and Technology*, 13, 374–381.
- AlAgha, I. (2021). Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives. *Journal of Information Science Theory and*

- Practice*, 9(1), 35–53. <https://doi.org/10.1633/JISTaP.2021.9.1.3>
- Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142(1). <https://doi.org/10.1088/1742-6596/1142/1/012012>
- Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2), 48–57. <https://doi.org/10.1109/MCI.2014.2307227>
- Cuomo, R. E., Purushothaman, V., Li, J., Cai, M., & Mackey, T. K. (2021). A longitudinal and geospatial analysis of COVID-19 tweets during the early outbreak period in the United States. *BMC Public Health*, 21(1), 793. <https://doi.org/10.1186/s12889-021-10827-4>
- Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013). Carmen: A twitter geolocation system with applications to public health. *AAAI Workshop - Technical Report, WS-13-09*, 20–24.
- Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., & De La Iglesia, B. (2019). Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PLoS ONE*, 14(7), 1–29. <https://doi.org/10.1371/journal.pone.0210689>
- Euzebio, C., Agy, S., Boldorini Jr., C., Porto, L., Alcarás, J. R., Martinez, A., & Ruiz, E. (2020). *Statistical analysis of small twitter data collection to identify dengue outbreaks*. 17–24. <https://doi.org/10.5753/kdmile.2020.11954>
- Fazeli, S., Zamanzadeh, D., Ovalle, A., Nguyen, T., Gee, G., & Sarrafzadeh, M. (2021). *COVID-19 and Big Data: Multi-faceted Analysis for Spatio-temporal Understanding of the Pandemic with Social Media Conversations*. <http://arxiv.org/abs/2104.10807>
- Hassan Zadeh, A., Zolbanin, H. M., Sharda, R., & Delen, D. (2019). Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis. *Information Systems Frontiers*, 21(4), 743–760. <https://doi.org/10.1007/s10796-018-9893-0>
- Heth, Z., Bemis, K., & Christiansen, D. (2018). Correlation of Tweets Mentioning Influenza Illness and Traditional Surveillance Data. *Online Journal of Public Health Informatics*, 10(1), 2579. <https://doi.org/10.5210/ojphi.v10i1.8773>
- Jain, A., & Cherikkallil, S. (2018). Medinsights: Twitter Based Platform for Health Care Analytics. *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018, Icirca*, 1104–1109. <https://doi.org/10.1109/ICIRCA.2018.8597360>
- Kitchenham, B. (2004). Procedures for Performing Systematic Reviews, Version 1.0. *Empirical Software Engineering*, 33(2004), 1–26.

- Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, 591–600. <https://doi.org/10.1145/1772690.1772751>
- Li, L., Gao, L., Zhou, J., Ma, Z., Choy, D. F., & Ha, M. A. (2021). Can Social Media Data Be Utilized to Enhance Early Warning: Retrospective Analysis of the U.S. Covid-19 Pandemic. *MedRxiv*, XX, 1–13. http://medrxiv.org/cgi/content/short/2021.04.11.21255285v1?rss=1&utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound
- Li, Z., Li, X., Porter, D., Zhang, J., Jiang, Y., Olatosi, B., & Weissman, S. (2020). Monitoring the spatial spread of covid-19 and effectiveness of control measures through human movement data: Proposal for a predictive model using big data analytics. *JMIR Research Protocols*, 9(12), 1–10. <https://doi.org/10.2196/24432>
- Mackey, T., Purushothaman, V., Li, J., Shah, N., Nali, M., Bardier, C., Liang, B., Cai, M., & Cuomo, R. (2020). Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. *JMIR Public Health and Surveillance*, 6(2), e19509. <https://doi.org/10.2196/19509>
- Martínez, N. J. F., & Pascual, C. P. (2020). Knowledge-based rules for the extraction of complex, fine-grained locative references from tweets. *Revista Electronica de Linguistica Aplicada*, 19(1), 136–163.
- Masri, S., Jia, J., Li, C., Zhou, G., Lee, M.-C., Yan, G., & Wu, J. (2019). Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC Public Health*, 19(1), 761. <https://doi.org/10.1186/s12889-019-7103-8>
- Pruss, D., Fujinuma, Y., Daughton, A. R., Paul, M. J., Arnot, B., Szafir, D. A., & Boyd-Graber, J. (2019). Zika discourse in the Americas: A multilingual topic analysis of Twitter. *PLoS ONE*, 14(5), 1–23. <https://doi.org/10.1371/journal.pone.0216922>
- Rashid, M. T., & Wang, D. (2021). CovidSens: a vision on reliable social sensing for COVID-19. *Artificial Intelligence Review*, 54(1), 1–25. <https://doi.org/10.1007/s10462-020-09852-3>
- Şerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management*, 56(3), 1166–1184. <https://doi.org/10.1016/j.ipm.2018.04.011>
- Sidana, S., Amer-Yahia, S., Clausel, M., Rebai, M., Mai, S. T., & Amini, M.-R. (2018). Health

- Monitoring on Social Media over Time. *IEEE Transactions on Knowledge and Data Engineering*, 30(8), 1467–1480. <https://doi.org/10.1109/TKDE.2018.2795606>
- Souza, R. C. S. N. P., Assunção, R. M., Oliveira, D. M., Neill, D. B., & Meira, W. (2019). Where did I get dengue? Detecting spatial clusters of infection risk with social network data. *Spatial and Spatio-Temporal Epidemiology*, 29, 163–175. <https://doi.org/10.1016/j.sste.2018.11.005>
- Spurlock, K., & Elgazzar, H. (2020). Predicting COVID-19 Infection Groups using Social Networks and Machine Learning Algorithms. *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2020*, 0245–0251. <https://doi.org/10.1109/UEMCON51285.2020.9298093>
- Statista. (2021). *Statista*. <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>
- Tufts, Polsky, D., Volpp, K. G., Groeneveld, P. W., Ungar, L., Merchant, R. M., & Pelullo, A. P. (2018). Characterizing tweet volume and content about common health conditions across Pennsylvania: Retrospective analysis. *JMIR Public Health and Surveillance*, 4(4), 1–9. <https://doi.org/10.2196/10834>
- We_are_Social. (2021). *We are Social*. <https://wearesocial.com/>
- Xavier, F., Olenski, J. R. W., Acosta, A. L., Sallum, M. A. M., & Saraiva, A. M. (2020). Analise de redes sociais como estrategia de apoio a vigilancia em saude durante a Covid-19. *Estudos Avancados*, 34(99), 261–282. <https://doi.org/10.1590/S0103-4014.2020.3499.016>