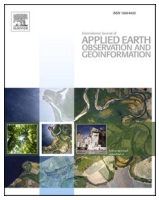
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

International Journal of Applied Earth Observations and Geoinformation 113 (2022) 102967

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Social media mining under the COVID-19 context: Progress, challenges, and opportunities

Xiao Huang [a,*], Siqin Wang [b], Mengxi Zhang [c], Tao Hu [d,*], Alexander Hohl [e], Bing She [f], Xi Gong [g], Jianxin Li [h], Xiao Liu [h], Oliver Gruebner [i], Regina Liu [j], Xiao Li [k], Zhewei Liu [l], Xinyue Ye [m], Zhenlong Li [n]

[a] Department of Geosciences, University of Arkansas, Fayetteville, AR 72701, USA
[b] School of Earth Environmental Sciences, University of Queensland, Brisbane, Queensland 4076, Australia
[c] Department of Nutrition and Health Science, Ball State University, Muncie, IN 47304, USA
[d] Department of Geography, Oklahoma State University, Stillwater, OK 74078, USA
[e] Department of Geography, The University of Utah, Salt Lake City, UT 84112, USA
[f] Institute for social research, University of Michigan, Ann Arbor, MI 48109, USA
[g] Department of Geography & Environmental Studies, University of New Mexico, Albuquerque, NM 87131, USA
[h] School of Information Technology, Deakin University, Geelong, Victoria 3220, Australia
[i] Department of Geography, University of Zurich, Zürich CH-8006, Switzerland
[j] Department of Biology, Mercer University, Macon, GA 31207, USA
[k] Texas A&M Transportation Institute, Bryan, TX 77807, USA
[l] Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[m] Department of Landscape Architecture and Urban Planning, Texas A&M University, College Station, TX 77840, USA
[n] Geoinformation and Big Data Research Lab, Department of Geography, University of South Carolina, Columbia, SC 29208, USA

ARTICLE INFO

Keywords:
COVID-19
Pandemic
Social media
Big data
Data mining

ABSTRACT

Social media platforms allow users worldwide to create and share information, forging vast sensing networks that allow information on certain topics to be collected, stored, mined, and analyzed in a rapid manner. During the COVID-19 pandemic, extensive social media mining efforts have been undertaken to tackle COVID-19 challenges from various perspectives. This review summarizes the progress of social media data mining studies in the COVID-19 contexts and categorizes them into six major domains, including early warning and detection, human mobility monitoring, communication and information conveying, public attitudes and emotions, infodemic and misinformation, and hatred and violence. We further document essential features of publicly available COVID-19 related social media data archives that will benefit research communities in conducting replicable and reproducible studies. In addition, we discuss seven challenges in social media analytics associated with their potential impacts on derived COVID-19 findings, followed by our visions for the possible paths forward in regard to social media-based COVID-19 investigations. This review serves as a valuable reference that recaps social media mining efforts in COVID-19 related studies and provides future directions along which the information harnessed from social media can be used to address public health emergencies.

1. Introduction

The COVID-19 pandemic has posed a global crisis, causing serious social, economic, and health challenges. Due to social media's interactive nature and popularity amongst users throughout the pandemic, there is rising democratization of health communications, which poses a sharp contrast to decades ago, when communications were predominantly controlled by individuals and entities endowed with the power, money, public trust, or platforms required to drive the conversation (Schillinger et al., 2020). The emerging concepts of "Web 2.0"

* Corresponding authors.
E-mail addresses: xh010@uark.edu (X. Huang), s.wang6@uq.edu.au (S. Wang), mzhang2@bsu.edu (M. Zhang), tao.hu@okstate.edu (T. Hu), alexander.hohl@geog.utah.edu (A. Hohl), bingshe@umich.edu (B. She), xigong@umn.edu (X. Gong), jianxin.li@deakin.edu.au (J. Li), xiao.liu@deakin.edu.au (X. Liu), oliver.gruebner@geo.uzh.ch (O. Gruebner), Regina.liu@live.mercer.edu (R. Liu), xiao.li@tamu.edu (X. Li), jackie.zw.liu@connect.polyu.hk (Z. Liu), xinyue.ye@tamu.edu (X. Ye), zhenlong@mailbox.sc.edu (Z. Li).

https://doi.org/10.1016/j.jag.2022.102967
Received 26 March 2022; Received in revised form 17 June 2022; Accepted 5 August 2022
Available online 19 August 2022

(Murugesan, 2007), "Big Data" (Yang et al., 2017), and "Citizen as Sensors" (Goodchild, 2007) have greatly promoted social media as the platforms and virtual communities where users worldwide can create and share information, forming vast sensing networks that allow information in certain topics to be collected, stored, mined, and analyzed in a rapid manner (Li et al., 2021a; Ye et al., 2021; Gong and Yang, 2020).

Since the early stage of the COVID-19 pandemic, governments, local authorities/agencies, and organizations have started to disseminate crucial information to the public via social media platforms. In addition, we have seen a massive influx of opinions, perceptions, and attitudes towards COVID-19 related events and/or public health policies from regular users on social media platforms. If appropriately utilized, the vast amount of minable information in the social media space would allow scholars to address various aspects of COVID-19 challenges. Extensive social media mining efforts have been made to tackle COVID-19 issues from various perspectives, including but not limited to case hotspot prediction (Li et al., 2020a), policy compliance monitoring (Huang et al., 2020), misinformation modeling (Cinelli et al., 2020), and sentimental analysis (Nemes and Kiss, 2021). Despite the existing studies, there is a lack of review work that cohesively summarizes the current findings in the COVID-19 context. Tsao et al. (2021) examined 81 peer-reviewed empirical studies relating to COVID-19 and social media published between November 2019 and November 2020. Their review predominantly targeted the early stage of the pandemic; therefore, it did not capture the major milestones amidst the middle and later stages of the pandemic after the mass vaccination. Other reviews related to social media and public health more broadly merely describe the general functionality and utility of social media in public health applications but lack the focus on social media analytics that are derived via data mining efforts (Giustini et al., 2018; Grajales III et al., 2014; Moorhead et al., 2013). Additionally, these reviews gave scarce attention to the challenges present in different domains of COVID-19 related studies (e.g., Asian hate, lockdown debate, and vaccination preferences)—prevalently discussed on social media platforms.

To address these knowledge deficits, we review existing social media mining efforts related to the COVID-19 crisis, document publicly available data archives, summarize social media mining challenges as well as their potential impacts on derived findings, and envision the future directions of social media-based COVID-19 and public health investigations. Due to the strong interdisciplinarity and the fact that we specifically target data mining efforts, a systematic reviewing workflow using keywords and databases for queries fails to provide a satisfactory article pool without intensive post-selection trimming. Thus, we organize this review in a narrative manner. The narrative review has been widely used to obtain a broad perspective on topics of interest. Instead of systematically searching for all relevant literature, it specifically focuses on pivotal papers known to the subject expert. The articles reviewed in this effort are purposively selected by the authors with rich experience in social media mining and who have conducted interdisciplinary COVID-19 investigations using social media data.

In the following sections, we group the progress of social media data mining studies to address COVID-19 challenges into six major categories and summarize notable efforts in each category (Section 2). These six categories include 1) early warning and detection; 2) human mobility monitoring; 3) communication and information conveying; 4) public attitudes and emotions; 5) infodemic and misinformation; 6) hatred and violence. Note that the authors' expertise and research experience well cover the identified six categories. We further document essential features of publicly available COVID-19 social media data archives that will benefit research communities in conducting replicable and reproducible studies (Section 3). In addition, we discuss seven challenges in social media analytics associated with their potential impacts on derived COVID-19 findings, followed by our visions for the possible paths forward in regard to social media-based COVID-19 investigations (Section 4). These challenges include 1) biased population spectrum; 2) multilingual investigations; 3) posting incentives; 4) positioning accuracy; 5) uncertainties in sentiment and emotion acquisition; 6) bots, retweets, and skewed posting behaviors; 7) data sharing. The structure of this review is presented in Fig. 1. We believe that this review can serve as a valuable reference that recaps COVID-19 related social media mining efforts and provide potential future directions for better employing the information harnessed from social media to address future public health emergencies.

2. Current progress

2.1. Early warning and detection

Public health surveillance is critical for monitoring the spread of infectious diseases, rapidly detecting outbreaks, and proposing effective countermeasures. With the support of early warning signs, governments are able to better prepare for public health emergencies such as the COVID-19 pandemic. The initial hotspot of COVID-19 was reported in China before cases were reported in European countries and in the United States (U.S.), which became the new epicenter of the disease as its number of confirmed cases surpassed that of Italy's on March 26, 2020. The rapid viral spread on a global scale demands public health authorities in many countries to develop mitigation strategies within a rapid timespan. Social media has played a crucial role in supporting traditional surveillance systems for tracking the progress of the COVID-19 pandemic and informing the judgments and decisions of public health officials and experts (Samaras et al., 2020). The real-time information from a massive sensor network consisting of millions of social media users provides timely situational awareness that uncovers early warning signs of an upcoming hotspot of cases, greatly facilitating the estimation of disease prevalence in (near) real time.

For example, Kogan et al. (2021) found that digital data sources may provide an earlier indication of the epidemic spread than traditional COVID-19 metrics, such as confirmed cases or deaths. By proposing a metric that combines six digital sources, including COVID-19 related Twitter activity, into a multiproxy estimator, their study demonstrated the potential of situational awareness that is derived from digital sources in estimating the probability of an impending COVID-19 outbreak (Kogan et al., 2021). By analyzing a multilingual dataset of tweets (i.e., English, German, French, Italian, Spanish, Polish, and Dutch posts that contain the keyword "pneumonia"), Lopreite et al. (2021) uncovered early-warning signals of the COVID-19 outbreaks in Europe during the winter season 2019–2020, before receiving the first public announcements of local sources of infection. This evidence suggests that European countries saw unexpected levels of concerns regarding COVID-19 cases, and whistleblowing came primarily from the geographical regions that eventually turned out to be the new COVID-19 hotspot (Lopreite et al., 2021). Qin et al. (2020) predicted the number of newly suspected or confirmed COVID-19 cases by analyzing social media search indexes for symptoms, coronavirus, and pneumonia. Via a series of analytical approaches, e.g., lasso regression, ridge regression, and elastic net, their study proved the feasibility of social media search indexes in predicting new suspected COVID-19 cases 6–9 days in advance (Qin et al., 2020). Similarly, Li et al. (2020a) analyzed COVID-19 related internet searches (Google and Baidu) and social media data (i.e., Sina Weibo) and demonstrated that for trend data and the number of cases, the highest correlation between these two variables occurred 8–12 days before an increase in confirmed COVID-19 cases, and the highest correlation between trend data and the newly suspected cases occurred 6–8 days before the increase in newly suspected cases. The above studies, as well as other social media based early warning and detection efforts (Lu and Zhang, 2020; Mackey et al., 2020), highlight the necessity of establishing social media surveillance systems that facilitate the identification of disease communication.

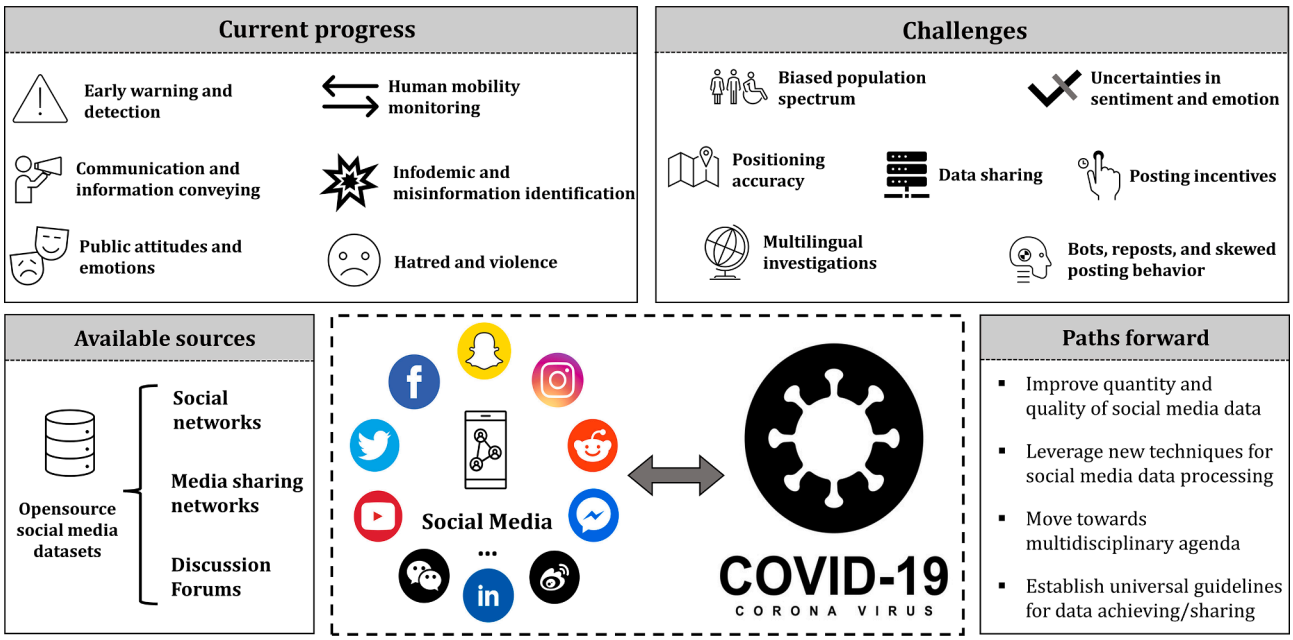


Fig. 1. The structure of this review.

2.2. Human mobility monitoring

The COVID-19 pandemic highlights the importance of rapid human mobility monitoring. User-generated information from social media platforms (e.g., Twitter, Facebook, Sina Weibo, and Instagram), when coupled with geo-information (i.e., geograohic coordinates and information on place names), allows human–human, human-place, and place-place interactions to be monitored in an active and less privacy-concerning manner (Huang et al., 2020; Li et al., 2021a), thus serving as an important venue where timely human mobility dynamics can be collected and analyzed to assist with decision making. Despite the existence of many social media platforms, only a small proportion of them permit information mining or open-source aggregated mobility records for researchers and the public, while for some social media platforms (e.g., Facebook and Sina Weibo), certain agreements have to be met to access to the records. Below, we review notable efforts that address COVID-19 challenges by monitoring human mobility dynamics via geotagged social media data.

With several categories of publicly available application programming interfaces (APIs), Twitter has become the most popular social media platform that allows geographic data mining. These APIs return certain percentages of their total content, with some of them containing geo-information at various levels. Studies have found that the Twitter-derived mobility patterns can approximate commuting patterns (Petutschnig et al., 2021) as well as mobility records released by Apple, Google, and Descartes Labs (Huang et al., 2021). Using 580 million geotagged tweets collected worldwide, Huang et al. (2020) measured human mobility by proposing the concept of single-day distance and cross-day distance, which highlight the users' daily travel behavior and the users' displacement between two consecutive days, respectively. Their investigations, conducted at various scales (i.e., global, country, and U.S. states), suggest that Twitter-derived mobility dynamics are amenable to reflect the geographical differences in policy implementations and discrepancies with policy compliance. Notably, Xu et al. (2020) proposed and utilized a Twitter Social Mobility Index, which measures social distancing and travel derived from geotagged Twitter posts, to analyze U.S. weekly travel patterns. Similar efforts were made to monitor global human mobility dynamics (Bisanzio et al., 2020; Lai et al., 2021; Li et al., 2021c; Li et al., 2021d) as well as country/region-specific dynamics where Twitter is widely used, such as the U.S. (Jiang et al., 2021; Zeng et al., 2021) and Australia (Nguyen et al., 2020a).

Facebook is another popular social media platform with a large global user base. Beginning in the initial phases of the COVID-19 pandemic, Facebook Data for Good began to provide human mobility information to assist with pandemic mitigation. For example, Chang et al. (2021) explored Facebook-derived movement patterns and used *meta*-population models to assess the potential effects of local travel restrictions imposed within Taiwan. Zachreson et al. (2021) used Facebook mobility data to estimate future spatial patterns of relative transmission risk and examine the degree to which these estimates correlate with observed cases in Australia. Besides these two efforts, Facebook mobility records were employed for mobility monitoring at a continental scale, as well as at a country/sub-country scale, e.g., the U.S. (Holtz et al., 2020; Ilin et al., 2021), the U.K. (Shepherd et al., 2021), Italy (Beria and Lunkar, 2020; Bonaccorsi et al., 2020), Spain (Pérez-Arnal et al., 2021), Japan (Fraser and Aldrich, 2020), Germany (Fritz and Kauermann, 2020), Demark (Edsberg Møllgaard et al., 2022), and Australia (Zachreson et al., 2021).

Several other social media platforms were harnessed to address COVID-19 challenges as well, such as U.S.' Instagram and China's Tencent and Sina Weibo. Zarei et al. (2020) constructed the first Instagram dataset, which featured COVID-19 related posts with locational information. Using Tencent's mobility data derived from Tencent's media various platforms, Li et al. (2020d) revealed daily human movement patterns in Sichuan, China (which covers the mobility of 90% of Sichuan citizens) during the initial stages of the COVID-19 outbreak, and Wei et al. (2021) evaluated how people in Wuhan, China reduced their mobility in response to city lockdowns. Another Chinese social media platform, Sina Weibo, also renders geotagged posts that allow researchers to mine the spatiotemporal patterns of human interactions and place visitations (Peng et al., 2020).

Social media platforms have proven to be one of the most vital sources of mobility data, enabling researchers to obtain critical insight into human mobility amidst COVID-19. Due to their active sharing characteristics, social media mobility records are less abundant compared to other passively collected records (e.g., mobile phone data, smart cards, or wireless networks), though they are less intrusive, more accessible, and more harmonized (Li et al., 2021c). However, limitations such as the necessity for users needing pre-existing incentives to make posts and varying positioning accuracy need to be recognized (discussed

in Section 4.1).

2.3. Communication and information conveying

Social media platforms are not only popular among individual users for user/news following, microblogging, and content sharing (Gong and Yang, 2020; Kietzmann et al., 2011) but have also become crucial tools for institutions (such as governments, organizations, and universities, etc.) to disseminate information, foster connections, and even manage crises (Gong and Lane, 2020; Kelly, 2013; Kostkova et al., 2014). Crisis communication refers to the sharing of information among individuals and institutions to improve crisis management and understanding (National Research Council, 1989). Crisis communication has been reshaped by social media in numerous ways, including raising public awareness through collaboration and participation, distributing information and instructions in real time, and monitoring and managing risks with greater efficiency (Olteanu et al., 2015; Reuter et al., 2016; Yoo, 2019). In spite of the virtual nature of social media interactions, the spatial social networks they have formed still reflect the geography of communication (Ye and Andris, 2021). Human interactions in real life and in cyberspace are similar in terms of their social, economic, cultural, and linguistic constraints; thus, spatial social networks tend to mimic real-life patterns (Bild et al., 2015; Stephens and Poorthuis, 2015). Many studies have used social media data to examine crisis communication under the COVID-19 context from a geographic and social network perspective. The majority of the COVID-19 crisis communication research focused on governmental agencies, but some also examined other public health stakeholders, such as non-governmental organizations (NGOs), educational organizations, and the public.

As the COVID-19 crisis unfolds, government organizations at different levels must act quickly to communicate crisis information to the public in an efficient and effective manner; failure to do so could lead to an increase in fear, uncertainty, and anxiety among the public (Chen et al., 2020b). Based on spatial–temporal analyses, network analyses, and text mining of the U.S. state governors' crisis communication on Twitter during the pandemic, Gong and Ye (2021) found that the current usage patterns are generally consistent with effective crisis communication principles (listening, informing, providing feedback, and establishing connections) and provided some concrete recommendations for improving the process. One qualitative analysis of how world leaders of the Group of Seven (G7) communicated about the COVID-19 pandemic indicated that 82% of their tweets were informative; many of them dealt with government resources, morale boosting, and political issues (Rufai and Bunce, 2020). According to Zhu et al. (2020), the analysis of Sina Weibo posts related to COVID-19 confirmed that early warnings of crises are vital because public attention to COVID-19 was relatively limited until the Chinese government acknowledged that the novel coronavirus could be transmitted between humans and designated control of the outbreak as a high priority on January 20, 2020. Through analyzing tweets from 292 federal members of the Canadian parliament, Merkley et al. (2020) reported a moment of cross-party consensus on COVID-19 communication. No matter which party the members were from, they emphasized social distancing and proper hand hygiene as a necessity for combatting the COVID-19 pandemic (Merkley et al., 2020). Wang, Hao, and Platt (2021) analyzed 13,598 COVID-19-relevant tweets from 67 U.S. federal and state-level government agencies from January to April 2020. They identified inconsistencies and incongruities in four crucial prevention topics and found that communications coordination increased over time. Using tweets from Texas-based public health agencies, Liu, Xu, and John (2021) examined interagency coordination at different stages of the pandemic. In addition to stage-specific variations in peer-to-peer and federal-to-local coordination, they also observed consistency in content across stages, i.e., state and federal agencies acting as agenda setters (Liu et al., 2021). Studying 138,546 tweets from 696 public health agency accounts from February 1 to March 31, 2020, Sutton, Renshaw, and Butts (2020) observed that longitudinal COVID-19 risk communication shifted as secondary threats emerged. In addition, there are studies addressing the best practices in COVID-19 crisis communication on social media. Government agencies can improve public engagement and crisis communication efficiency on social media by leveraging narrative evidence (Gesser-Edelsburg, 2021; Ngai et al., 2020), adopting an empathic communication style (Liao et al., 2020), actively using the dialogic loop rather than media richness (Chen et al., 2020b), and joining forces with leading scientists from various domains (Tsoy et al., 2021) to generate persuasive and potent content. These findings may help government agencies to create communication plans for future crises and assist the public in understanding, preparing for, and predicting governments' response strategies.

The pandemic has spawned a wealth of complex problems, such as healthcare resource shortages, economic recession, mental health issues, and other social problems, all of which are difficult to resolve by governments alone. Therefore, it is imperative for public health stakeholders, NGOs, education institutions, and the general public to collaborate with government agencies within and across boundaries to address problems collectively (Head and Alford, 2015; Li et al., 2021b; Roberts, 2000; Weber and Khademian, 2008). After examining the evolution of Twitter-based networks and discourse across 2,588 U.S. NGOs in the first five months of the COVID-19 outbreak, Li et al. (2021b) discovered that social media usage helped NGOs to connect with each other by removing geographical barriers and specialty constraints. Over time, distinct organizational communities emerged around different topics, mostly reflecting theoretical predictions based on Issue Niche Theory (Yang, 2020). The interactions and connections among NGOs and government agencies during the COVID-19 pandemic are well reflected on social media platforms, reflecting their goals to share information, build communities, and take action for disaster response. The government agencies played a leading role in the NGO-government collaborations, while NGOs from the Human Services, International and Foreign Affairs, and Public and Societal Benefit sectors, especially the American Red Cross, played a more central role in the NGO collaboration network. The study of social media usage by 189 Greek libraries during the pandemic revealed that although libraries embraced social media quickly as a channel for communication, only a few highlighted their roles in the promotion of public health by providing timely and reliable information (Koulouris et al., 2020). The COVID-19 pandemic has forced all educational institutions to move from face-to-face to online instruction. Students in higher education use social media primarily to build an online community and to support one another, whereas faculty members use it exclusively for teaching and learning (Sobaih et al., 2020). Based on analyses of tweets from 492 U.S. K-12 school districts in March-April 2020, Michela et al. (2022) found that these districts followed recommendations for social media crisis communication by posting more announcements and engaging more collaboratively during the early pandemic phases, and by sharing more community-building contents later. Crisis communication among the general public is also crucial to disaster response. Yu et al. (2021) analyzed 10,132 COVID-19 related online comments on TripAdvisor and discovered a dynamic shift in risk perceptions and communication intensities among the general public as a result of the pandemic's rapid and unpredictable spread. During the COVID-19 pandemic, many instances of stereotyping and discrimination toward Asian Americans and the elderly population have been posted on social media, many of which are associated with stigmatizing and blaming these populations (Croucher et al., 2020; Meisner, 2021). Meisner (2021) urged the public to be aware of and to resist ageism that devalues later life in crisis communication. All of these findings provide unprecedented insight into how different public health stakeholders are working collaboratively to combat the pandemic, which can help the entire society prepare for the implantation of crisis communication strategies in anticipation of future global hazards.

2.4. Public attitudes and emotions

The COVID-19 pandemic has led to a major uprise in studies that apply sentiment analysis towards social media platforms' text-based content in order to gauge the public's attitude and sentiment revolving both the pandemic and related categories, such as public health policies (e.g., mask-wearing) and/or events (e.g., vaccination) (Ewing and Vu, 2021; Kwok et al., 2021; Manguri et al., 2020). Sentiment analysis is thought to enable the derivation of the users' emotional response to a particular event or phenomena via the text-based contents that they post (e.g., words, expressions, languages, and syntaxes) (Agarwal et al., 2011; Kouloumpis et al., 2011). Moreover, social media-based sentiment studies have also been used to indicate the public's awareness, opinions, or mental health signals based on the quantification and intensity of sentiments (e.g., positive V.S. negative, or optimistic V.S. pessimistic) and the type of emotions (e.g., fear, sadness, joy, and surprise) (Coppersmith et al., 2014). Such studies are further able to supplement survey-based mental health assessments, enabling researchers to mitigate issues such as a limited data pool (e.g., limited spatial and temporal data coverage, data under-representativeness) (Balcombe and De Leo, 2020). Sentiment research that seeks to quantify sentiments via social media data typically relies on advanced measuring techniques, including artificial intelligence (AI) models and machine and/or deep learning algorithms (Ewing and Vu, 2021; Hu et al., 2021; Kwok et al., 2021; Wang et al., 2020a; Wang et al., 2022). The advanced AI models include the Valence Aware Dictionary for sEntiment Reasoning (VADER) (Wang et al., 2022) and the National Research Council Canada Lexicon model (NRCLex) (Hu et al., 2021), both of which target English-based contents, as well as the XLM-R or XLM-T model (Conneau et al., 2019; Imran et al., 2022) and the Hugging Face (Barbieri et al., 2021), which targets multilingual contents. More nuanced reviews and surveys of the models and algorithms used in sentiment analysis can be found in Alsaeedi and Khan (Alsaeedi and Khan, 2019) and Medhat et al. (2014).

Empirically, sentiment-based studies that employ social media data were typically used to evaluate the public's attitudes and sentiments towards COVID-19 related policy implementations, including mask-wearing, economic support, and school closure (Ewing and Vu, 2021; Kwok et al., 2021; Manguri et al., 2020; Niu et al., 2021). Others have investigated the public's opinion and awareness of COVID-19 related events (e.g., protests against lockdown, vaccination, and university reopening) and speeches/comments of political leaders (e.g., Donald Trump) (Hu et al., 2021; Jang et al., 2021). Current studies have been conducted in several countries such as the U.S. (Jang et al., 2021; Lyu et al., 2021), the U.K. (Cheng et al., 2021; Rahman and Islam, 2022), Australia (Ewing and Vu, 2021; Wang et al., 2022), India (Barkur and Vibha, 2020), China (Li et al., 2020a; Wang et al., 2020a), Europe (Kruspe et al., 2020), as well as across multiple countries (Boon-Itt and Skunkan, 2020; Matošević and Bevanda, 2020; Rowe et al., 2021). Existing studies focus predominantly on solely English-based content, while a smaller proportion uses either content that is in Chinese and retrieved from Weibo (the largest social media platform in China) (Li et al., 2020a; Wang et al., 2020a) or non-verbal content (e.g., emoticons) (Yamamoto et al., 2014); scarce attention has been allotted to sentiment analysis involving multilingual content (discussed in Section 4.1.2).

2.5. Infodemic and misinformation

With the rapid dispersion of COVID-19, a tsunami of related information rushed across the internet. Yet, such information remains unfiltered, and many contain misinformation, rumors, and conspiracy theories. On this influx of information, the World Health Organization's (WHO) Director, General Tedros proclaimed, "We're not just fighting an epidemic; we're fighting an infodemic". Thus, on February 15, 2020, at the Munich Security Conference, Tedros officially coined this phenomenon as the "Infodemic", which describes a situation where, during a period of disease outbreak, there exists a vast amount of information that is false or misleading in nature and is present in physical and digital environments. During the COVID-19 crisis, misinformation can spread faster and farther on social media platforms than the virus itself. According to a Reuters report, the number of English-language fact-checks rose more than 900% from January to March 2020 (Brennen et al., 2020). This information covers a wide range of topics, e.g., "5G virus is true", "eating garlic can prevent coronavirus", and "Bill Gates is planning to microchip the world through a COVID-19 vaccine". Such misinformation and the resulting risk-taking behaviors can lead to mistrust in health authorities and undermine public health response. In light of this context, scholars around the world have started to investigate misinformation spread on social media platforms.

Social media generates a massive amount of information related to the COVID-19 pandemic every day, and manual identification of misinformation is time- and labor-consuming. As an alternative solution, advanced machine learning techniques have been deployed to detect misinformation automatically. The accuracy of misinformation detection models relies on sufficient and reliable datasets. Thus, many efforts have been made to provide high-quality social media misinformation datasets. Researchers collected ground-truth data from fact-checking websites (Ceron et al., 2021; Saakyan et al., 2021; Shahi and Nandini, 2020) and reliable websites (Cui and Lee, 2020; Zhou et al., 2020) for the misinformation detection task. Specifically, FakeCovid (Shahi and Nandini, 2020) is a multilingual cross-domain fact-check news dataset that contains 5,182 articles circulated in 105 countries (40 languages) from 92 fact-checkers. CoAid is a healthcare domain dataset containing 4,521 true news articles and claims from reliable media outlets (e.g., Healthline, ScienceDaily, and WHO) (Cui and Lee, 2020). CoAid collects fake news by retrieving URLs from multiple fact-checking websites such as LeadStories, PolitiFact, and FactCheck. ReCOVery is a multimodal repository for COVID-19 news credibility research, which contains 1,364 news articles from 22 reliable websites (e.g., National Public Radio and Reuters) and 665 news articles from 38 unreliable websites (e.g., Human Are Free and Natural News).

Another research direction is to collect misinformation-related posts from social media users to explore user engagement (Cui and Lee, 2020; Kim et al., 2021; Li et al., 2020c) and public opinion (Gupta et al., 2020; Wang et al., 2020b; Xue et al., 2020; Yin et al., 2020). Misinformation detection is essentially a classification task. The common workflow is to develop a dataset with true and false labels for model training, adjust the model based on the results of the test set, and apply it to unknown data in order to generate predictions. Machine learning (ML) and deep learning (DL) models have been widely used for misinformation detection (Alenezi and Alqenaei, 2021; Elhadad et al., 2020; Gundapu and Mamidi, 2021; Kar et al., 2020; Koirala, 2020). Traditional ML models, such as decision trees, support vector machines, and logistic regression, usually serve as baseline models in fake news detection model experiments. Al-Rakhami and Al-Amri (2020) proposed an ensemble framework for misinformation detection by using traditional ML and conducting extensive experiments on a self-collected Twitter dataset. Their work demonstrates that a combination of models outperforms a single model. For DL models, a variety of models have been used to address the COVID-19 misinformation identification challenge, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Bidirectional Encoder Representations from Transformers (BERT) (Al-Rakhami and Al-Amri, 2020), to list a few. Alkhalifa et al. (2020) introduced a CNN-based classification system with different preprocessing and embedding methods to classify COVID-19 rumors. An ensemble deep learning technique that was implemented to detect misleading information for COVID-19 had achieved satisfactory performance (Elhadad et al., 2020). Other two advanced models, i.e., BiLSTM (Boukouvalas et al., 2020; Dharawat et al., 2020; Hossain et al., 2020; Kumar et al., 2021) and BiGRU (Cui and Lee, 2020; Elhadad et al., 2020), have also been widely adopted in recognizing misinformation on social media. The recent development of BERT has pushed

natural language processing to a new level, thanks to its capability in capturing both left and right contexts, given its bidirectional design. Some BERT variants were adopted for COVID-19 misinformation detection (Alkhalifa et al., 2020; Glazkova et al., 2021; Heidari et al., 2021; Perrio and Madabushi, 2020; Tziafas et al., 2021). The multilingual BERT (mBERT) is a notable variant trained on Wikipedia andconsiders a total of 104 languages. COVID-Twitter-BERT was trained on the 160 million COVID-19 related corpus on the Crowdbreaks platform and performed very well on many textual representations related to COVID-19 (Müller et al., 2020).

2.6. Hatred and violence

Since the first confirmed case of COVID-19 in the U.S. on January 19, 2020 (Hossain et al., 2020), hateful and xenophobic language has surged on social media. This was quickly followed by prejudice and discriminatory acts against minorities, particularly the Asian and Asian American population (Croucher et al., 2020; Fan et al., 2020). Notably, during the period of March 19, 2020, to September 30, 2021, the Stop AAPI (Asian American and Pacific Islander) Hate reporting center recorded a total of 10,370 hate incidents against Asians and Asian Americans (Horse et al., 2021). Today, racism is recognized as a public health threat by the American Medical Association, as the connection between hateful social media posts and offline racially and religiously aggravated crime has previously been documented (Williams et al., 2020). Moreover, for both traditional media and social media, the spread of hate during COVID-19 on such platforms results in potentially negative effects on population health, and such an observation has also been previously recorded (Gao et al., 2020a; Quintero Johnson et al., 2021). Malicious content like racism and disinformation is spreading quickly beyond the control of individual social media platforms, thereby subverting their efforts to moderate content (Velasquez et al., 2021).

In response to the rapid dissemination of such malicious content, the scientific community has responded with a series of actions that are of similar fervor: for instance, annotated tweet datasets for the detection of racism and sexism were readily available before the pandemic (Davidson et al., 2017; Waseem and Hovy, 2016) in many languages other than English, i.e., a dataset of Spanish tweets for misogyny detection (Fersini et al., 2018), an Arabic dataset for detection of hate speech and fake news (Ameur and Aliane, 2021), and an annotated dataset of abusive language in German (Wich et al., 2021). Early work on social media mining during COVID-19 established the theoretical groundwork for detecting hate speech on social media by using keyword-based classifiers. Nguyen et al. (2020b) found evidence of increased negative sentiment towards Asians associated with the "#chinesevirus" hashtag in early 2020 (Nguyen et al., 2020b). Anti-Chinese and anti-Asian attacks on social media platforms were mainly targeting eating habits, hygiene, and in general, culture (Stechemesser et al., 2020). Lastly, exploratory work during the early stages of the pandemic includes the application of space–time scan statistics (Kulldorff, 1997) to assess the spatiotemporal distribution of geotagged tweets regarding Asian hate (Hohl et al., 2022).

Other studies, though aspatial, have focused on identifying anti-Asian hate and counterspeech on social media via using BERT (He et al., 2021). BERT was used to identify hate-related keywords that targeted older people during the pandemic (Vishwamitra et al., 2020) and fine-tuned to analyze COVID-19 content on Twitter (Müller et al., 2020). This approach utilizes word embeddings in conjunction with machine learning to classify tweets, therefore providing an advantage over the keyword-based classifiers' method of incorporating word context. Such a method was used for analyzing the dehumanization towards LGBTQ people in articles in the New York Times (Mendelsohn et al., 2020). Further, crowdsourcing and ensemble learning algorithms were utilized to detect hate on social media in Germany (Garland et al., 2020, 2022). Amidst sudden changes during the early stages of the pandemic, issues with obtaining costly training data regarding hate speech detection (e.g., labeled social media posts) were circumvented through the usage of unsupervised progressive domain adaptation based on a deep-learning language model (Bashar et al., 2021). Lastly, efforts to analyze the effects of content moderation policies on the propagation of malicious posts (within social media platforms) using mathematical models produced encouraging results, accompanied by actionable suggestions towards slowing the spread of online hate (Velasquez et al., 2021).

3. Available sources and application examples

In the past two years, we have witnessed a tremendous surge in the use of social media platforms during the COVID-19 pandemic to study misinformation, public opinion, human behavior, infodemic, and more. Social media platforms can be categorized as social networks (e.g., Twitter, Sina Weibo, and Facebook), media sharing networks (e.g., YouTube), and discussion forums (e.g., Reddit). However, limited social media datasets are shared with the public, hindering collaborative research and increasing the crisis of research reproducibility and replicability. As a result, this section summarizes and compares the most popular and publicly available social media datasets in terms of geolocation, content, advanced data analytics, geographic coverage, time coverage, and selected citations, as shown in Table 1.

3.1. Social networks

Twitter plays a significant role in featuring COVID-19 related research, and its original data includes user information, content, post time, and more. However, due to privacy concerns, the publication of personal Twitter data is not permissible. Therefore, after collecting COVID-19 related tweets, many researchers publicly share Twitter IDs to allow ease of access. With these Twitter IDs, users can hydrate tweets and access original information via the Twitter API. For example, Chen et al. (2020a) used Twitter's search API to gather global wide historical COVID-19 related Tweets based on the keywords (i.e., Coronavirus, Koronavirus, Corona, covid-19, and N95) dating back to January 21, 2020. Their team has shared their repository, which contains an ongoing collection of tweet IDs. So far, it is the most popular Twitter dataset cited by researchers across the world.

Although the act of sharing raw Twitter data is restricted, many researchers share their findings with the public via advanced approaches, such as releasing their findings regarding the sentiment, emotions, and topics of users' tweets. For example, Lopez and Caleb (2021) collected over 2.2 billion tweets across the globe in multiple languages. Additionally, they employed state-of-art algorithms to analyze sentiment and recognize named entities in Twitter content. Such aggregated information facilitates the researchers' exploration and hypothesis testing on social discourse regarding the COVID-19 pandemic (Lopez and Gallemore, 2021).

Locations and medical emergencies are intrinsically linked. Geotagged tweets are able to provide real-time information about human activities at a low cost and high spatial and temporal resolutions. They also enable researchers to, using geography as a common variable, join attributes across various datasets (e.g., sociodemographic) (Hu and Wang, 2020). Due to this advantage, Qazi et al. (2020) released the GeoCoV19 dataset, which contains around 378,000 geotagged tweets and 5.4 million tweets with locational information at the country, state, and city levels. Further, Lamsal (2021)'s publication of tweet IDs enabled individuals who hydrate these IDs access to the geotagged datasets.

Sina Weibo, commonly referred to as the "Chinese Twitter", is the leading social media platform in China, with 497 million active monthly users in 2019 (Fu and Zhu, 2020). Given that China was the earliest country to report COVID-19 outbreaks, many researchers have shared and utilized datasets from Sina Weibo to analyze misinformation. For example, Leng et al. (2020) crawled Sina Weibo posts via Weibo API

Table 1
Public available COVID-19 related social media datasets.

Data Provider	Dataset Name	Geolocation Included	ID Only	Text	Advanced Analysis/ secondary data	Geographic Coverage	Temporal Coverage	Publication
Twitter	COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions	No	No	No	Sentiment and emotion	Global/country	1/28/2020 – 9/1/2021	(Gupta et al., 2020)
	COVID-19-TweetIDs	No	Yes	No	No	Global	01/21/ 2020–02/11/ 2022	(Chen et al., 2020a)
	Coronavirus (COVID-19) tweets dataset	No	No	No	Sentiment	Global	03/20/ 2020–02/12/ 2022	(Lamsal, 2021)
	Coronavirus geo-tagged tweets datasets	Yes	No	No	Sentiment	Global	03/20/ 2020–02/12/ 2022	(Lamsal, 2021)
	Covid-19 Twitter chatter dataset for scientific use	No	Yes	No	No	Global/country	03/22/ 2020–02/12/ 2022	(Banda et al., 2021)
	CoronaVis: A Real-time COVID-19 Tweets Analyzer	No	Yes	No	No	Global	03/05/ 2020–12/31/ 2020	(Kabir and Madria, 2020)
	An Augmented Multilingual Twitter dataset for studying the COVID-19 infodemic	Yes	No	No	Sentiment, entity recognition, mentions, and hashtags	Global/Country	01/01/ 2020–12/31/ 2021	(Lopez and Gallemore, 2021)
	GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information	Yes	No	No	Sentiment and entity recognition	Global/location	02/01/ 2020–03/31/ 2022	(Qazi et al., 2020)
	COVID-19 Twitter Dataset	No	Yes	No	No	Global	04/01/ 2020–09/31/ 2020	(Gruzd and Mai, 2020)
	Preliminary Extraction from Geotweet Archive v2.0 for COVID-19 Tweets	Yes	Yes	No	No	Global/location	03/01/ 2020–04/30/ 2020	
Sina Weibo	Weibo COVID dataset	No	No	Yes	No	China	12/07/ 2020–04/04/ 2020	(Leng et al., 2020)
	COVID-19 related Weibo Data	No	No	Yes	No	China	12/01/ 2019–02/27/ 2020	(Fu and Zhu, 2020)
Reddit	Reddit Mental Health Dataset	No	No	Yes	Sentiment and emotion	28 mental health and non-mental health subreddits	01/01/ 2018–01/01/ 2020	
	Coronavirus subreddit	No	No	Yes	Sentiment and topic modeling	r/Coronavirus subreddit	01/20/ 2020–01/31/ 2021	
	The Reddit COVID dataset		No	Yes	Sentiment	posts and comments mentioning COVID in their title and body text	N/A- 25/10/ 2021	(Tan, 2021)
Youtube	YouTube's Pseudoscientific Video Recommendations	No	No	Yes	No	Search terms: 'covid-19', 'coronavirus', 'anti-vaccination', 'anti-vaxx', 'anti-mask', or 'flat earth'		(Papadamou et al., 2020)
	Covid-related misinformation videos	No	Yes	No	No	Global	11/ 012019–06/ 30/2020	
Instagram	COVID19 Instagram Post IDs	No	Yes	No	No	Global	01/05/ 2020–3/30/ 2020	(Zarei et al., 2020)

from December 7, 2019, to April 4, 2020, and shared the datasets on Harvard Dataverse. Fu and Zhu (2020) collected 11,362,502 posts between December 1, 2019, and February 27, 2020, which contains at least one outbreak-related keyword (e.g., mask, virus, or coronavirus).

3.2. Media sharing networks (YouTube and Instagram)

Media sharing networks such as YouTube and Instagram have also been important channels where people receive COVID-19 information through venues such as videos and photos. The YouTube Data API could be used to find videos through search queries. Video metadata, including title descriptions, tags, video statistics, comments, as well as the recommended videos, can be collected through the YouTube Data API. For example, Papadamou et al. (2020) collected COVID-related videos and recommendations through the API to analyze the effect of a user's watch history on video recommendations. Notably, YouTube has also served as a source of COVID-19 misinformation (Allington et al., 2021). These videos are often linked by content on other social media sites, including Reddit, Twitter, and Facebook. Knuutila et al. (2021) displayed a dataset of COVID-related video identifiers that were removed by YouTube, though the video's metadata were recovered through archive.org's Wayback Machine. Researchers have also actively studied the content of YouTube videos, as it could be both useful as a source of information (D'Souza et al., 2020) and play a role in spreading

misinformation (Li et al., 2020b). A common approach is to select the most viewed videos by search queries and analyze the video content along with its metadata. Basch et al. (2020) identified the 100 most widely viewed YouTube videos in January 2020, using the search term "Coronavirus". Their analysis revealed that only one-third of the videos covered key prevention behaviors.

Instagram data, including post comments, geotags, and captions, could be retrieved with open-source tools such as the Instaloader. Researchers are able to share data through the Post IDs. For example, Zarei et al. (2020) used the Instagram Hashtag search API to retrieve public posts with a set of COVID-19 hashtags and crawl the reactions (comments or likes) for further analysis. Researchers have previously applied Natural Language Processing and deep learning techniques to Instagram posts as well. For example, Mackey et al. (2020) analyzed illicit COVID-19 product sales from Twitter and Instagram posts using unsupervised topic modeling and a recurrent neural network with long short-term memory (LSTM) unit to identify online sellers.

3.3. Discussion forums

Discussion forums, e.g., Quora, Yahoo answers (shut down on May 4, 2021), Infobot, and Reddit, provide users with online spaces to discuss news and answer questions, where the public comments and statements can be collected by researchers to study the influences of COVID-19. Among the above-mentioned discussion forums, Quora and Reddit are the two commonly used forums that enabled for many COVID-19 studies.

Quora is a popular question-and-answer (Q&A) website where users are allowed to ask questions and connect with people who contribute unique insights and quality answers. The COVID-19 pandemic has greatly stimulated people's interest in asking and answering COVID-19 related questions, and a large amount of content can be harnessed for COVID-19 studies. George et al. (2020), for example, analyzed the content, type, and quality of Q&As in Quora regarding the pandemic and compared the information with that on the WHO website by manually categorizing the tone of the question as either positive, negative or ambivalent and grading questions for accuracy, authority, popularity, readability, and relevancy. Another notable effort is by McCreery et al. (2020), who designed a fine-tuning neural network approach trained by Quora question pairs to identify similar posted questions.

Reddit, one of the most widely used discussion forums, allows registered users to submit content to the site, such as links, text posts, images, and videos, among which lots of content are related to current events. A popular method to collect Reddit data is through the PushShift API, which serves as a copy of Reddit objects. For example, Low et al. (2020) introduced the Reddit Mental Health Dataset that contains posts from 28 subreddits from 2018 to 2020. Reddit data provide a new lens for researchers to study emotion, gender differences, and mental health during the COVID-19 pandemic. Text mining and natural language processing techniques are the major analytical tools employed in research. Naseem et al. (2020) leveraged Non-negative Matrix Factorization (NMF) topic modeling on Reddit posts to study life during the pandemic and the effects of social distancing. Aggarwal et al. (2020) analyzed emotions through the Valence-Arousal-Dominance (VAD) affect representation. Word embeddings of Reddit data were used to train beta regression models in order to predict VAD scores. The results revealed considerable differences between male and female authors across all three emotional dimensions.

4. Challenges and our paths forward

4.1. Challenges

4.1.1. Biased population spectrum

In 2020, social media platforms were used by over 3.6 billion people worldwide, and this number is projected to increase to almost 4.4 billion in 2025 (Tankovska, 2021). Despite this growing trend, however, there has been an argument that the current demographics of social media active users are unrepresentative of the entire population across the world in terms of age, gender, race, education, or socioeconomic status. Jiang et al. (2019) found that Twitter users in the entire U.S. are biased towards certain age groups (18–29 and 30–39), females, and people with Bachelor's and Graduate degrees). They also discovered that U.S. Twitter users' spectrum presents strong spatial non-stationarity, suggesting that the biases of Twitter users vary by geographical location (Jiang et al., 2019). Facebook users are most represented by individuals between the ages of 25 and 35 years (Barnhart, 2022). The demographic representation on one of China's largest social media platforms, Sina Weibo, also has a user demographic that is considerably different from that of the national population statistics, with males composing 56.3% of users, 20–35 years old comprising 82% of users, and with 91% of users with Bachelor's degrees (Weibo-Sina, 2017). Such biases are also observed in other social media platforms, including WeChat and Instagram. Thus, it remains debatable whether place visitations, mobility patterns, sentiment, or emotions captured from social media space are representative of those of the entire population. Applying such findings derived from a small minority towards the general public is cautioned against, unless they are statistically compared with and supported by other means of data collection that are less biased, such as questionnaires and surveys.

4.1.2. Multilingual investigations

Social media data presents several advantageous characteristics, such as facilitating the process of intra- and inter-continental investigations due to its breadth of foreign languages and allowing comparisons between data derived from different regions where cellphone records from certain providers can differ geographically. Despite their advantages, multilingual posts in the social media space pose challenges towards contextual interpretation. For example, every month, there are over 330 million active Twitter users across the world, using tens of languages, with English (31.8%), Japanese (18.8%), and Spanish (8.46%) as the three most popular languages (VICINITAS, 2018). Current studies that extract situational awareness and perform sentiment/emotion analysis on COVID-19 related posts tend to focus on monolingual posts (Griffith et al., 2021; Mansoor et al., 2020; Shofiya and Abidi, 2021) or multilingual posts with naïve translating approaches (Lin et al., 2021; Zhang et al., 2021). When applied to study areas with two or more dominant languages though, such investigative procedures ultimately ignore specific groups of people and introduce uncertainties when summarizing emotions and sentimental preferences across different languages. Despite the development in multilingual translation, which is supported by the advances in natural language processing techniques, the potential biases in extracting and quantifying sentiment and emotions across different languages are still deserving further exploration.

4.1.3. Posting incentives

For geotagged social media posts, the active sharing characteristics of social media data inevitably lead to a "warped reality", when compared to actual human-to-human interactions and place visitations. That is to say, human mobility patterns extracted from the social media space are a biased representation of actual human mobility. For example, geotagged social media posts derived from check-in records generally have to satisfy two requirements: 1) users are geographically close to the check-in locations (or at least they claim themselves to be); 2) the check-in locations are worth posting (i.e., "interesting" enough for them to create a post). In comparison, geolocations obtained via passive collecting means (e.g., WIFI, Call Detail Records, and GPS signals) only need to satisfy the former requirement. Such a biased and inevitably generalized representation may lead to uncertainties or even mistakes when they are applied to the decision-making process for COVID-19 mitigation. For sentiment and emotion mining, studies have shown

that bursts of posting tend to occur following major events (Pohl et al., 2012; Zhou and Chen, 2014). In other words, a considerable amount of social media posts are event/news-driven. Therefore, the question as to whether emotions and sentiments from event-triggered posts largely reflect options towards the event itself or the general topic remains to be explored. Unfortunately, it remains a challenge to grasp the contextual meaning behind sentiments and emotions using the current natural language processing techniques.

4.1.4. Positioning accuracy

The levels to which social media data are geotagged vary greatly (depending on the social media platforms' terms of use and users' specific settings), posing challenges to studies that prefer certain geolocational accuracy for social media posts. In general, the geotagging levels include country, first-level subdivision, second-level subdivision, city, neighborhood/point of interest (POI), and exact coordinates. A study conducted by Li et al. summarized the positioning levels of 1.4 billion geotagged tweets worldwide: 1.1 billion (79%) at the city level, 138.1 million (9.8%) at the first-level subdivision (state or province), 90.4 million (6.4%) with exact coordinates, 46.2 million (3.3%) at country level, and 21.4 million (1.5%) at neighborhood/point of interest (Li et al., 2021a). Certainly, different social media platforms have varying preferences towards certain positioning levels. For example, Sina Weibo check-in data returned from Sina Weibo API are mostly positioned at the POI level (Hu et al., 2019), whereas Facebook Data for Good only provides re-aggregated data at certain administrative levels due to privacy concerns (Edsberg Møllgaard et al., 2022). The varying positioning levels of social media posts impose a great influence on the statistical findings of studies that summarize statistics within certain geographic units due to the modifiable areal unit problem (MAUP). For applications that demand accurate human moving patterns, integrating social media posts with mixed positioning levels produces significant uncertainties that should not be overlooked.

4.1.5. Uncertainties in sentiment and emotion

Uncertainties in sentiment analysis and the emotions that it extracts from social media posts have been widely acknowledged. Despite the fact that advanced natural language processing techniques, when applied to multilingual posts, enable reliable translation for certain languages, they still have relatively less consistency and lower performances for those that are less spoken (Balahur and Jacquet, 2015). This leads to increased uncertainties in the results of sentiment and emotion analysis when they are applied to multilingual regions, especially in those with less spoken languages. Certain social media platforms, such as Twitter and Weibo, have character limits, creating oddities (e.g., the usage of abbreviations and acronyms) found in posts that would otherwise not be present in normal language. Furthermore, the unique character-limit restrictions imposed on posts made on certain social media platforms demand the application of word vectors trained specifically from short-text documents instead of the ones that are from popular word representation models, such as Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) and Embeddings from Language Models (ELMo) (Peng et al., 2019). In addition, within the context of COVID-19, we should note that certain words have sentimental tendencies that are opposite to their original meaning. For example, the sentence "*I have been tested positive*" has a negative sentiment polarity, despite the fact that the word "*positive*" presents a strong positive polarity in many sentiment analysis models. Another challenge is the treatment of neutral reporting of valenced information, e.g., "the daily death toll dropped to 1,000" and "100 more have been tested positive today". It is unclear whether these statements should be considered as neutral unemotional reporting of developments or assumed that users are in negative/positive emotional states.

4.1.6. Bots, retweets, and skewed posting behaviors

From 50 million tweets, Al-Rawi & Shukla (2020) identified the top 1,000 most active accounts that mention COVID. Within these accounts, 127 (12.7%) were identified as highly likely to be bots. In an early study involving Weibo, a random sample of roughly 30,000 users was found to contain 57% of either inactive users or "zombie accounts" due to these accounts' lack of consistent postings over time (Fu and Chau, 2013). The method by which social media bots are handled in social media analytics is important for studies that address COVID-19 challenges by mining information from authentic human users. Despite the bots composing a smaller population than that of authentic human users, relatively high posting volumes can greatly contaminate researchers' data. However, we have yet to find an automatic and correct approach to identifying bots. Hence, this issue remains a challenge. In addition, we must acknowledge the skewed posting behaviors of social media users, given that a majority of social media posts come from a minority of users. For example, 80% of tweets come from the top 10% of the most active users (Wojcik and Hughes, 2019), which means that our analysis of a collection of social media posts is likely to be skewed towards a small subset of users. Optimized weighting mechanisms based on posting frequencies and user ID indexed analytical workflows can be adopted to address this issue. However, our literature review suggests that few efforts have considered such a skewed user representation when performing social media mining in the context of COVID-19. The question of how re-posting behaviors should be managed is yet another challenge because there is a multitude of methods to account for re-posting, which may alter the analytical results. Despite the fact that many studies have designated re-posts as an agreement to the original post, Metaxas et al. (2014) found that, on many occasions, this assumption may actually not be the case.

4.1.7. Data sharing

Data sharing in social media, especially during the COVID-19 pandemic, has become a crucial driving force in motivating social media studies to address COVID-19 challenges. Properly shared social media data archives support validity by advancing reproducibility, replicability, and comparability, addressing the 'digital divides' in data accessibility and saving efforts in the data collection processes (Weller and Kinder-Kurlanda, 2016). However, existing efforts often fail to be grounded in the general principles that underlie institutionalized data archiving and sharing, as the lack of standardized metadata, consistent documentation, and sustainable claim in current COVID-19 social media sharing efforts can be clearly observed. The lack of guidance for social media data sharing, especially during the COVID-19 pandemic, leads to sharing procedures that vary by different social media research communities. Thus, the current practices of social media data sharing need to be coherent and universally agreed upon in order to benefit not only future COVID-19 studies but also investigations on other public health emergencies.

4.2. Future directions

Based on the aforementioned challenges, we propose a number of research directions along which future efforts can be made to broaden and deepen the current research paradigm. These future directions are discussed in the context of the quantity and quality of social media data, the techniques used to process social media data, its application across multi-disciplines, and data archiving and sharing.

First, future efforts should be made to have a better understanding of the nature of social media data and to improve the quantity and quality of social media data. In particular, efforts towards enriching social media data with the demographic attributes of social media users under the protection of data privacy are much needed. Social media users' demographic attributes affect their participation in the social network and further influence their behaviors (e.g., mental health status) (Sinnenberg et al., 2017). Obtaining users' demographic attributes can be difficult because they cannot be directly collected from social media platforms. However, such demographic information, including age,

gender, socioeconomic status, religion, and personality type, can be extrapolated from a user's tweets via machine learning, with an accuracy ranging from 60% to 90% (Bi et al., 2013; Burger et al., 2011; Ikeda et al., 2013; Pennacchiotti and Popescu, 2011; Rao et al., 2010). This is an underutilized resource for studies using social media data and can be applied towards understanding the subjects of investigation and reducing sampling biases. More specifically, studies based on individual tweets rather than individual users face the issue of skewed data problems, given that one user may post multiple tweets in a certain period of time. It can be addressed by the user indexed analytics, which is based on users' ID (e.g., as individuals or organizations). Such demographic information would also enable us to calibrate and justify the representativeness and reliability of social media data by cross-data validation based on other data sources (e.g., survey or census data).

Second, future work could explore new approaches and techniques in data retrieval, processing, and analytics to provide potential solutions to conquer the constraints inherent in social media data-based studies. For data retrieval, using Twitter APIs has been the most common approach to retrieve tweets that target certain topics. In early 2021, a new academic-oriented API was released by Twitter, which grants free access to full-archive search for researchers to obtain more precise, complete, and unbiased data, greatly benefitting future Twitter-based analytics thanks to its increased data representativeness (Twitte, 2021). Facebook posts can be retrieved via CrowdTangle API, a public tool owned and operated by Facebook (CrowdTangle, 2016). However, the representativeness of retrieved Facebook posts deserves further investigation (Yang et al., 2021). Posts in the discussion forums, such as Reddit and Quora, are valuable sources to gauge public attention. Additional cautions are needed when retrieving topic-relevant questions and answers. In addition, further efforts are encouraged to conduct cross-comparison on analytical results from different social media platforms, given that social media platforms can have user bases that vary in population spectrum. For data processing and analytics, technical solutions lie in the rapid development of computational skills and platforms (e.g., artificial intelligence, digital twins, and crowdsourcing) as they are more effective and efficient ways to quantify human behaviors (e.g., fuzzy logic lexical metrics and multilingual sentiment analysis). Comparison studies across different methods for data pre-processing are also needed. Taking Twitter as an example, analytical results might be different when using tweets V.S. retweets, tweets with or without URLs, tweets including emoticons or not, and tweets generated by robots or not. We also need to establish standards for social media data reporting and a generic metadata architecture to better compare the scalability, replication, and reliability of social media data-based studies.

Third, we call for investigations on social media bi-directional communications (e.g., organization-to-individual, individual-to-individual, and individual-to-organization) before, during, and after the COVID-19 pandemic as well as during other disruptive events. We also call for broader potential and opportunities for using social media data in multidisciplinary studies across social, geographic, environmental, and computational sciences to better understand the impact of COVID-19 on human-environment interaction. In addition to the popular domains mentioned in Section 2, social media data can be applied to a wider network of fields under the context of the COVID-19 pandemic, such as commercial industry, transportation, security and information management, and social phycology. For example, social media data has great potential for understanding and addressing COVID-19 related cyber-bullying (Das et al., 2020), detecting suicide (Morese et al., 2022) or mental disorders within a certain population (Sher, 2020), and evaluating the recovery of restaurants (Laguna et al., 2020) and hospitality industry (Park et al., 2020). Social media data provides a unique opportunity to support governments and public/private sectors by monitoring the recovery of human society in the later stages of the pandemic and preparing for future public health emergencies and crises.

Finally, there is a need for universal guidelines that address the ethics of social media research, with a focus on maintaining the privacy and anonymity of social media users. Sharing social media data, though they are largely claimed as 'anonymized', via public repositories and platforms should be supported by discussions of obtaining consent and/or ethical approval for research purposes (Fadda et al., 2022). This is particularly important for datasets containing information of users' profiles due to the fact that such datasets have the risk of being identifiable via cross-referencing data attributes (Sinnenberg et al., 2017). Under the protection of data sharing regulations and the spirit of reproductivity, we should endeavor to facilitate the sharing of the processed social media data via public repositories and platforms and establish reproducible workflow that can be employed by end-users without a coding background.

5. Conclusion

Social media have been widely used as platforms and virtual communities where users worldwide can create and share information. The vast sensing network constituted by millions of active users and billions of posts allow information on certain topics to be mined and analyzed in a rapid manner. During the COVID-19 pandemic, we have witnessed extensive social media data mining efforts with diversified data mining techniques. These efforts address COVID-19 challenges from various perspectives, including early warning and detection, human mobility monitoring, communication and information conveying, gauging public attitudes and emotions, monitoring infodemic and misinformation, and mitigating hatred and violence. We also notice that an increasing number of COVID-19 related social media datasets have been made publicly available to benefit research communities by promoting replicability and reproducibility. Despite the remaining challenges (e.g., biased population spectrum and difficulty in multilingual investigations), we believe the future is bright for social media analytics to address future public health emergencies.

CRediT authorship contribution statement

Xiao Huang: Conceptualization, Resources, Writing – original draft, Writing – review & editing, Supervision. **Siqin Wang:** Supervision, Conceptualization, Writing – original draft, Writing – review & editing. **Mengxi Zhang:** Supervision, Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Tao Hu:** Supervision, Conceptualization, Writing – original draft, Writing – review & editing. **Alexander Hohl:** Writing – original draft. **Bing She:** Writing – original draft. **Xi Gong:** Writing – original draft. **Jianxin Li:** Writing – original draft. **Xiao Liu:** Writing – original draft. **Oliver Gruebner:** Writing – review & editing. **Regina Liu:** Writing – original draft, Writing – review & editing. **Xiao Li:** Conceptualization, Writing – review & editing. **Zhewei Liu:** Writing – review & editing. **Xinyue Ye:** Conceptualization. **Zhenlong Li:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.J., 2011. Sentiment analysis of twitter data, Proceedings of the workshop on language in social media (LSM 2011), pp. 30-38.

Aggarwal, J., Rabinovich, E., Stevenson, S., 2020. Exploration of gender differences in COVID-19 discourse on reddit. arXiv preprint arXiv:2008.05713.

Alenezi, M.N., Alqenaei, Z.M., 2021. Machine learning in detecting COVID-19 misinformation on twitter. Future Internet 13, 244.

Alkhalifa, R., Yoong, T., Kochkina, E., Zubiaga, A., Liakata, M., 2020. QMUL-SDS at CheckThat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. arXiv preprint arXiv:2008.13160.

Allington, D., Duffy, B., Wessely, S., Dhavan, N., Rubin, J., 2021. Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychol. Med. 51, 1763–1769.
Al-Rakhami, M.S., Al-Amri, A.M., 2020. Lies kill, facts save: detecting COVID-19 misinformation in twitter. IEEE Access 8, 155961–155970.
Al-Rawi, A., Shukla, V., 2020. Bots as active news promoters: A digital analysis of COVID-19 tweets. Information 11, 461.
Alsaeedi, A., Khan, M.Z., 2019. A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications 10, 361–374.
Ameur, M.S.H., Aliane, H., 2021. AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. Procedia Comput. Sci. 189, 232–241.
Balahur, A., Jacquet, G., 2015. Sentiment analysis meets social media–Challenges and solutions of the field in view of the current information sharing context. Elsevier 428–432.
Balcombe, L., De Leo, D., 2020. An integrated blueprint for digital mental health services amidst COVID-19. JMIR mental health 7, e21718.
Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G., 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. Epidemiologia 2 (3), 315–324.
Barbieri, F., Anke, L.E., Camacho-Collados, J., 2021. Xlm-t: A multilingual language model toolkit for twitter. arXiv preprint arXiv:2104.12250.
Barkur, G., Vibha, G.B.K., 2020. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry 51, 102089.
Barnhart, B. (2022, March 2). Social media demographics to inform your Brand's strategy in 2022. Sprout Social. Retrieved March 13, 2022, from https://sproutsocial.com/insights/new-social-media-demographics/.
Basch, C.H., Hillyer, G.C., Meleo-Erwin, Z.C., Jaime, C., Mohlman, J., Basch, C.E., 2020. Preventive behaviors conveyed on YouTube to mitigate transmission of COVID-19: cross-sectional study. JMIR public health and surveillance 6, e18807.
Bashar, M.A., Nayak, R., Luong, K., Balasubramaniam, T., 2021. Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts. Social Network Analysis and Mining 11, 1–18.
Beria, P., Lunkar, V., 2020. Presence and mobility of the population during Covid-19 outbreak and lockdown in Italy.
Bi, B., Shokouhi, M., Kosinski, M., Graepel, T., 2013. Inferring the demographics of search users: Social data meets search queries. In: Proceedings of the 22nd international conference on World Wide Web, pp. 131–140.
Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M., Wallach, D.S., 2015. Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. ACM Transactions on Internet Technology (TOIT) 15, 1–24.
Bisanzio, D., Kraemer, M.U., Bogoch, I.I., Brewer, T., Brownstein, J.S., Reithinger, R., 2020. Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale. Geospatial health 15.
Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A.L., Valensise, C.M., Scala, A., Quattrociocchi, W., 2020. Economic and social consequences of human mobility restrictions under COVID-19. Proc. Natl. Acad. Sci. 117, 15530–15535.
Boon-Itt, S., Skunkan, Y., 2020. Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. JMIR Public Health and Surveillance 6, e21978.
Boukouvalas, Z., Mallinson, C., Crothers, E., Japkowicz, N., Piplai, A., Mittal, S., Joshi, A., Adalı, T., 2020. Independent component analysis for trustworthy cyberspace during high impact events: an application to Covid-19. arXiv preprint arXiv:2006.01284.
Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation (Doctoral dissertation, University of Oxford).
Burger, J.D., Henderson, J., Kim, G., Zarrella, G., 2011. Discriminating gender on twitter. Association for Computational Linguistics.
Ceron, W., de-Lima-Santos, M.-F., Quiles, M.G., 2021. Fake news agenda in the era of COVID-19: Identifying trends through fact-checking content. Online Social Networks and Media 21, 100116.
Chang, M.-C., Kahn, R., Li, Y.-A., Lee, C.-S., Buckee, C.O., Chang, H.-H., 2021. Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. BMC Public Health 21, 1–10.
Chen, E., Lerman, K., Ferrara, E., 2020a. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR public health and surveillance 6, e19273.
Chen, Q., Min, C., Zhang, W., Wang, G., Ma, X., Evans, R., 2020b. Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis. Comput. Hum. Behav. 110, 106380.
Cheng, I., Heyl, J., Lad, N., Facini, G., Grout, Z., 2021. Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. Sci. Rep. 11, 1–13.
Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A., 2020. The COVID-19 social media infodemic. Sci. Rep. 10, 1–10.
Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
Coppersmith, G., Dredze, M., Harman, C., 2014. Quantifying mental health signals in Twitter, Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, pp. 51-60.
Croucher, S.M., Nguyen, T., Rahmani, D., 2020. Prejudice toward Asian Americans in the COVID-19 pandemic: The effects of social media use in the United States. Frontiers. Communication 39.
CrowdTangle. 2016. Content discovery and Social Monitoring Made Easy. CrowdTangle. Retrieved June 14, 2022, from https://www.crowdtangle.com/.
Cui, L., Lee, D., 2020. Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.
D'Souza, R.S., D'Souza, S., Strand, N., Anderson, A., Vogt, M.N., Olatoye, O., 2020. YouTube as a source of medical information on the novel coronavirus 2019 disease (COVID-19) pandemic. Global public health 15, 935–942.
Das, S., Kim, A., & Karmakar, S. (2020). Change-point analysis of cyberbullying-related twitter discussions during covid-19. arXiv preprint arXiv:2008.13613.
Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, pp. 512–515.
Dharawat, A., Lourentzou, I., Morales, A., Zhai, C., 2020. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation. arXiv preprint arXiv:2010.08743.
Edsberg Møllgaard, P., Lehmann, S., Alessandretti, L., 2022. Understanding components of mobility during the COVID-19 pandemic. Philosophical Transactions of the Royal Society A 380, 20210118.
Elhadad, M.K., Li, K.F., Gebali, F., 2020. Detecting misleading information on COVID-19. IEEE Access 8, 165201–165215.
Ewing, L.-A., Vu, H.Q., 2021. Navigating 'home schooling'during COVID-19: Australian public response on twitter. Media International Australia 178, 77–86.
Fadda, M., Sykora, M., Elayan, S., Puhan, M. A., Naslund, J. A., Mooney, S. J., .et al. 2022. Ethical issues of collecting, storing, and analyzing geo-referenced tweets for mental health research. Digital Health, 8, 20552076221092539.
Fan, L., Yu, H., Yin, Z., 2020. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. Proceedings of the Association for Information Science and Technology 57, e313.
Fersini, E., Rosso, P., Anzovino, M., 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. Ibereval@ sepln 2150, 214–228.
Fraser, T., Aldrich, D.P., 2020. Social ties, mobility, and covid-19 spread in Japan.
Fritz, C., Kauermann, G., 2020. On the interplay of regional mobility, social connectedness, and the spread of COVID-19 in Germany. arXiv preprint arXiv:2008.03013.
Fu, K.-W., Chau, M., 2013. Reality check for the Chinese microblog space: a random sampling approach. PLoS ONE 8, e58356.
Fu, K.-W., Zhu, Y., 2020. Did the world overlook the media's early warning of COVID-19? J. Risk Res. 23, 1047–1051.
Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., Wang, Y., Fu, H., Dai, J., 2020a. Mental health problems and social media exposure during COVID-19 outbreak. PLoS ONE 15, e0231924.
Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., Galesic, M., 2020. Countering hate on social media: Large scale classification of hate and counter speech. arXiv preprint arXiv:2006.01974.
Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., Galesic, M., 2022. Impact and dynamics of hate and counter speech online. EPJ Data Sci. 11, 3.
George, J., Gautam, D., Kesarwani, V., Sugumar, P.A., Malhotra, R., 2020. What Does the Public Want to Know About The COVID-19 Pandemic? A Systematic Analysis of Questions Asked in The Internet. medRxiv. https://doi.org/10.1101/2020.09.15.20192039.
Gesser-Edelsburg, A., 2021. Using narrative evidence to convey health information on social media: the case of COVID-19. Journal of Medical Internet Research 23, e24948.
Giustini, D., Ali, S.M., Fraser, M., Boulos, M.N.K., 2018. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. Online journal of public health informatics 10.
Glazkova, A., Glazkov, M., Trifonov, T., 2021. g2tmn at constraint@ aaai2021: exploiting CT-BERT and ensembling learning for COVID-19 fake news detection, International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation. Springer 116–127.
Gong, X., Lane, K.M.D., 2020. Institutional Twitter usage among US geography departments. The professional geographer 72, 219–237.
Gong, X., Yang, X., 2020. Social media platforms. The geographic information science & technology body of knowledge 1–9.
Gong, X., Ye, X., 2021. Governors Fighting Crisis: Responses to the COVID-19 Pandemic across US States on Twitter. The Professional Geographer 73, 683–701.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.
Grajales III, F.J., Sheps, S., Ho, K., Novak-Lauscher, H., Eysenbach, G., 2014. Social media: a review and tutorial of applications in medicine and health care. Journal of medical Internet research 16, e2912.
Griffith, J., Marani, H., Monkman, H., 2021. COVID-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. Journal of medical Internet research 23, e26874.
Gruzd, A., Mai, P., 2020. Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. Big Data & Society 7, 2053951720938405.
Gundapu, S., Mamidi, R., 2021. Transformer based automatic COVID-19 fake news detection system. arXiv preprint arXiv:2101.00180.
Gupta, R.K., Vishwanath, A., Yang, Y., 2020. COVID-19 Twitter dataset with latent topics, sentiments and emotions attributes. arXiv preprint arXiv:2007.06954.
He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., Kumar, S., 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 90–94.
Head, B.W., Alford, J., 2015. Wicked problems: Implications for public policy and management. Administration & society 47, 711–739.

Heidari, M., Zad, S., Hajibabaee, P., Malekzadeh, M., HekmatiAthar, S., Uzuner, O., Jones, J.H., 2021. Bert model for fake news detection based on social bot activities in the covid-19 pandemic, 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE 0103–0109.

Hohl, A., Choi, M., Yellow Horse, A. J., Medina, R. M., Wan, N., & Wen, M. (2022). Spatial Distribution of Hateful Tweets Against Asians and Asian Americans During the COVID-19 Pandemic, November 2019 to May 2020. In American Journal of Public Health (Vol. 112, Issue 4, pp. 646–649). American Public Health Association. https://doi.org/10.2105/ajph.2021.306653.

Holtz, D., Zhao, M., Benzell, S.G., Cao, C.Y., Rahimian, M.A., Yang, J., Allen, J., Collis, A., Moehring, A., Sowrirajan, T., 2020. Interdependence and the cost of uncoordinated responses to COVID-19. Proc. Natl. Acad. Sci. 117, 19837–19843.

Horse, A.J.Y., Jeung, R., Lim, R., Tang, B., Im, M., Higashiyama, L., Schweng, L., Chen, M., 2021. Stop AAPI hate national report. Stop AAPI Hate: San Francisco, CA, USA.

Hossain, T., Logan IV, R.L., Ugarte, A., Matsubara, Y., Young, S., Singh, S., 2020. COVIDLies: Detecting COVID-19 misinformation on social media.

Hu, Q., Bai, G., Wang, S., Ai, M., 2019. Extraction and monitoring approach of dynamic urban commercial area using check-in data from Weibo. Sustainable cities and society 45, 508–521.

Hu, T., Wang, S., Luo, W., Zhang, M., Huang, X., Yan, Y., Li, Z., 2021. Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: spatiotemporal perspective. Journal of Medical Internet Research 23 (9), e30854.

Hu, Y., Wang, R.-Q., 2020. Understanding the removal of precise geotagging in tweets. Nat. Hum. Behav. 4, 1219–1221.

Huang, X., Li, Z., Jiang, Y., Li, X., Porter, D., 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS ONE 15, e0241957.

Huang, X., Li, Z., Jiang, Y., Ye, X., Deng, C., Zhang, J., Li, X., 2021. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the US during the COVID-19 pandemic. Int. J. Digital Earth 14, 424–442.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., Higashino, T., 2013. Twitter user profiling based on text and community mining for market analysis. Knowl.-Based Syst. 51, 35–47.

Ilin, C., Annan-Phan, S., Tai, X.H., Mehra, S., Hsiang, S., Blumenstock, J.E., 2021. Public mobility data enables covid-19 forecasting and management at local and global scales. Sci. Rep. 11, 1–11.

Imran, M., Qazi, U., Ofli, F., 2022. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. Data 7, 8.

Jang, H., Rempel, E., Roth, D., Carenini, G., Janjua, N.Z., 2021. Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. Journal of medical Internet research 23, e25431.

Jiang, Y., Huang, X., Li, Z., 2021. Spatiotemporal Patterns of Human Mobility and Its Association with Land Use Types during COVID-19 in New York City. ISPRS Int. J. Geo-Inf. 10, 344.

Jiang, Y., Li, Z., Ye, X., 2019. Understanding Demographic and Socioeconomic Bias of Geotagged Twitter Users at the County Level. Cartography and Geographic Information Science 46 (3).

Kabir, M., Madria, S., 2020. CoronaVis: a real-time COVID-19 tweets data analyzer and data repository. arXiv preprint arXiv:2004.13932.

Kar, D., Bhardwaj, M., Samanta, S., Azad, A.P., 2020. No rumours please! a multi-indic-lingual approach for COVID fake-tweet detection, 2021 Grace Hopper Celebration India (GHCI). IEEE 1–5.

Kelly, K.J., 2013. The effectiveness of Twitter as a communication tool in college recruitment. Texas A&M. University-Kingsville.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. Bus. Horiz. 54, 241–251.

Kim, J., Aum, J., Lee, S., Jang, Y., Park, E., Choi, D., 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. Telematics Inform. 64, 101688.

Knuutila, A., Herasimenko, A., Au, H., Bright, J., Howard, P.N., 2021. A Dataset of COVID-Related Misinformation Videos and their Spread on Social Media. Journal of Open Humanities Data 7.

Kogan, N.E., Clemente, L., Liautaud, P., Kaashoek, J., Link, N.B., Nguyen, A.T., Lu, F.S., Huybers, P., Resch, B., Havas, C., 2021. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. Science. Advances 7, eabd6989.

Koirala, A., 2020. COVID-19 fake news classification using deep learning.

Kostkova, P., Szomszor, M., St. Louis, C., 2014. # swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Transactions on Management Information Systems (TMIS) 5, 1-25.

Kouloumpis, E., Wilson, T., Moore, J., 2011. Twitter sentiment analysis: The good the bad and the omg!, Proceedings of the international AAAI conference on web and social media, pp. 538-541.

Koulouris, A., Vraimaki, E., Koloniari, M., 2020. COVID-19 and library social media use. Reference Services Review.

Kruspe, A., Häberle, M., Kuhn, I., Zhu, X.X., 2020. Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. arXiv preprint arXiv: 2008.12172.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics-Theory and methods 26, 1481–1496.

Kumar, S., Pranesh, R.R., Carley, K.M., 2021. A fine-grained analysis of misinformation in covid-19 tweets.

Kwok, S.W.H., Vadde, S.K., Wang, G., 2021. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. Journal of medical Internet research 23, e26953.

Laguna, L., Fiszman, S., Puerta, P., Chaya, C., Tárrega, A., 2020. The impact of COVID-19 lockdown on food priorities. Results from a preliminary study using social media and an online survey with Spanish consumers. Food Qual. Prefer. 86, 104028.

Lai, S., Floyd, J., Tatem, A., 2021. Preliminary risk analysis of the spread of new COVID-19 variants from the UK. South Africa and Brazil.

Lamsal, R., 2021. Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence 51 (5), 2790–2804.

Leng, Y., Zhai, Y., Sun, S., Wu, Y., Selzer, J., Strover, S., Fensel, J., Pentland, A., Ding, Y., 2020. Analysis of misinformation during the COVID-19 outbreak in China: cultural, social and political entanglements. arXiv preprint arXiv:2005.10414.

Li, C., Chen, L.J., Chen, X., Zhang, M., Pang, C.P., Chen, H., 2020a. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance 25, 2000199.

Li, H.-O.-Y., Bailey, A., Huynh, D., Chan, J., 2020b. YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ global health 5, e002604.

Li, X., Xu, H., Huang, X., Guo, C.A., Kang, Y., Ye, X., 2021a. Emerging geo-data sources to reveal human mobility dynamics during COVID-19 pandemic: opportunities and challenges. Computational Urban Science 1, 1–9.

Li, Y., Jiang, B., Shu, K., Liu, H., 2020c. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. arXiv preprint arXiv: 2011.04088.

Li, Y., Shin, J., Sun, J., Kim, H.M., Qu, Y., Yang, A., 2021b. Organizational sensemaking in tough times: The ecology of NGOs' COVID-19 issue discourse communities on social media. Comput. Hum. Behav. 122, 106838.

Li, Y., Zeng, Y., Liu, G., Lu, D., Yang, H., Ying, Z., Hu, Y., Qiu, J., Zhang, C., Fall, K., 2020d. Public awareness, emotional reactions and human mobility in response to the COVID-19 outbreak in China–a population-based ecological study. Psychol. Med. 1–8.

Li, Z., Huang, X., Hu, T., Ning, H., Ye, X., Huang, B., Li, X., 2021c. ODT FLOW: A Scalable Platform for Extracting, Analyzing, and Sharing Multi-source Multi-scale Human Mobility. PLoS ONE 16 (8), e0255259.

Li, Z., Huang, X., Ye, X., Jiang, Y., Martin, Y., Ning, H., Hodgson, M.E., Li, X., 2021d. Measuring global multi-scale place connectivity using geotagged social media data. Sci. Rep. 11, 1–19.

Liao, Q., Yuan, J., Dong, M., Yang, L., Fielding, R., Lam, W.W.T., 2020. Public engagement and government responsiveness in the communications about COVID-19 during the early epidemic stage in China: infodemiology study on social media data. Journal of medical Internet research 22, e18796.

Lin, B., Zou, L., Duffiel, N., Mostafavi, A., Cai, H., Zhou, B., Tao, J., Yang, M., Mandal, D., Abedin, J., 2021. Revealing the Global Linguistic and Geographical Disparities of Public Awareness to Covid-19 Outbreak through Social Media. arXiv preprint arXiv: 2111.03446.

Liu, W., Xu, W., John, B., 2021. Organizational disaster communication ecology: Examining interagency coordination on social media during the onset of the COVID-19 pandemic. American Behavioral Scientist 65, 914–933.

Lopez, C.E., Gallemore, C., 2021. An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. Social Network Analysis and Mining 11, 1–14.

Lopreite, M., Panzarasa, P., Puliga, M., Riccaboni, M., 2021. Early warnings of COVID-19 outbreaks across Europe from social media. Sci. Rep. 11, 1–7.

Low, D.M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., Ghosh, S.S., 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. Journal of medical Internet research 22, e22635.

Lu, Y., Zhang, L., 2020. Social media WeChat infers the development trend of COVID-19. J. Infect. 81, e82–e83.

Lyu, J.C., Le Han, E., Luli, G.K., 2021. COVID-19 vaccine–related discussion on Twitter: topic modeling and sentiment analysis. Journal of medical Internet research 23, e24435.

Mackey, T., Purushothaman, V., Li, J., Shah, N., Nali, M., Bardier, C., Liang, B., Cai, M., Cuomo, R., 2020a. 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, e19509.

Mackey, T.K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., Liang, B., 2020b. Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram. JMIR public health and surveillance 6, e20794.

Manguri, K.H., Ramadhan, R.N., Amin, P.R.M., 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan Journal of Applied Research 54–65.

Mansoor, M., Gurumurthy, K., Prasad, V., 2020. Global sentiment analysis of COVID-19 tweets over time. arXiv preprint arXiv:2010.14234.

Matošević, G., Bevanda, V., 2020. Sentiment analysis of tweets about COVID-19 disease during pandemic, 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE 1290–1295.

McCreery, C.H., Katariya, N., Kannan, A., Chablani, M., Amatriain, X., 2020. August). Effective transfer learning for identifying similar questions: matching user questions to COVID-19 FAQs. In: In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3458–3465.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 5, 1093–1113.

Meisner, B.A., 2021. Are you OK, Boomer? Intensification of ageism and intergenerational tensions on social media amid COVID-19. Leisure Sciences 43, 56–61.

Mendelsohn, J., Tsvetkov, Y., Jurafsky, D., 2020. A framework for the computational linguistic analysis of dehumanization. Frontiers in artificial intelligence 3, 55.

Merkley, E., Bridgman, A., Loewen, P.J., Owen, T., Ruths, D., Zhilin, O., 2020. A rare moment of cross-partisan consensus: Elite and public response to the COVID-19

pandemic in Canada. Canadian Journal of Political Science/Revue canadienne de science politique 53, 311–318.
Metaxas, P.T., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., Finn, S., 2014. Do retweets indicate interest, trust, agreement? arXiv preprint arXiv:1411.3555.
Michela, E., Rosenberg, J.M., Kimmons, R., Sultana, O., Burchfield, M.A., Thomas, T., 2022. "We Are Trying to Communicate the Best We Can": Understanding Districts' Communication on Twitter During the COVID-19 Pandemic. AERA Open 8, 23328584221078542.
Moorhead, S.A., Hazlett, D.E., Harrison, L., Carroll, J.K., Irwin, A., Hoving, C., 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. Journal of medical Internet research 15, e1933.
Morese, R., Gruebner, O., Sykora, M., Elayan, S., Fadda, M., Albanese, E., 2022. Detecting suicide ideation in the era of social media: the population neuroscience perspective. Front. Psychiatry. https://doi.org/10.3389/fpsyt.2022.652167.
Müller, M., Salathé, M., Kummervold, P.E., 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv: 2005.07503.
Murugesan, S., 2007. Understanding Web 2.0. IT Prof. 9, 34–41.
Naseem, S.S., Kumar, D., Parsa, M.S., Golab, L., 2020. Text mining of COVID-19 discussions on Reddit, 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE, pp. 687-691.
National Research Council, 1989. Improving risk communication.
Nemes, L., Kiss, A., 2021. Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication 5, 1–15.
Ngai, C.S.B., Singh, R.G., Lu, W., Koon, A.C., 2020. Grappling with the COVID-19 health crisis: content analysis of communication strategies and their effects on public engagement on social media. Journal of medical Internet research 22, e21360.
Nguyen, T., Gupta, S., Raman, J., Bellomo, R., Venkatesh, S., 2020a. Geolocated Twitter-based population mobility in Victoria, Australia, during the staged COVID-19 restrictions. Critical care and resuscitation: journal of the Australasian Academy of Critical Care Medicine.
Nguyen, T.T., Criss, S., Dwivedi, P., Huang, D., Keralis, J., Hsu, E., Phan, L., Nguyen, L. H., Yardi, I., Glymour, M.M., 2020b. Exploring US shifts in anti-Asian sentiment with the emergence of COVID-19. Int. J. Environ. Res. Public Health 17, 7032.
Niu, J., Rees, E., Ng, V., Penn, G., 2021. Statistically Evaluating Social Media Sentiment Trends towards COVID-19 Non-Pharmaceutical Interventions with Event Studies. In: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, pp. 1–6.
Olteanu, A., Vieweg, S., Castillo, C., 2015. What to expect when the unexpected happens: Social media communications across crises. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pp. 994–1009.
Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., Sirivianos, M., 2020. " It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. arXiv preprint arXiv:2010.11638.
Park, E., Kim, W.H., Kim, S.B., 2020. Tracking tourism and hospitality employees' real-time perceptions and emotions in an online community during the COVID-19 pandemic. Current Issues in Tourism 1–5.
Peng, Y., Yan, S., Lu, Z., 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.
Peng, Z., Wang, R., Liu, L., Wu, H., 2020. Exploring urban spatial features of COVID-19 transmission in Wuhan based on social media data. ISPRS Int. J. Geo-Inf. 9, 402.
Pennacchiotti, M., Popescu, A.-M., 2011. Democrats, republicans and starbucks afficionados: user classification in twitter, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 430-438.
Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.
Pérez-Arnal, R., Conesa, D., Alvarez-Napagao, S., Suzumura, T., Català, M., Alvarez-Lacalle, E., Garcia-Gasulla, D., 2021. Comparative analysis of geolocation information through mobile-devices under different Covid-19 mobility restriction patterns in Spain. ISPRS Int. J. Geo-Inf. 10, 73.
Perrio, C., Madabushi, H.T., 2020. CXP949 at WNUT-2020 Task 2: Extracting Informative COVID-19 Tweets–RoBERTa Ensembles and The Continued Relevance of Handcrafted Features. arXiv preprint arXiv:2010.07988.
Petutschnig, A., Albrecht, J., Resch, B., Ramasubramanian, L., Wright, A., 2021. Commuter Mobility Patterns in Social Media: Correlating Twitter and LODES Data. ISPRS Int. J. Geo-Inf. 11, 15.
Pohl, D., Bouchachia, A., Hellwagner, H., 2012. Automatic sub-event detection in emergency management using social media. In: Proceedings of the 21st international conference on world wide web, pp. 683–686.
Qazi, U., Imran, M., Ofli, F., 2020. GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. SIGSPATIAL Special 12, 6–15.
Qin, L., Sun, Q., Wang, Y., Wu, K.-F., Chen, M., Shia, B.-C., Wu, S.-Y., 2020. Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. Int. J. Environ. Res. Public Health 17, 2365.
Quintero Johnson, J.M., Saleem, M., Tang, L., Ramasubramanian, S., Riewestahl, E., 2021. Media Use During COVID-19: An investigation of negative effects on the mental health of Asian versus White Americans. Frontiers in Communication 6, 79.
Rahman, M., Islam, M.N., 2022. Exploring the performance of ensemble machine learning classifiers for sentiment analysis of covid-19 tweets. Sentimental Analysis and Deep Learning. Springer 383–396.
Rao, D., Yarowsky, D., Shreevats, A., Gupta, M., 2010. Classifying latent user attributes in twitter, Proceedings of the 2nd international workshop on Search and mining user-generated contents, pp. 37-44.
Reuter, C., Ludwig, T., Kaufhold, M.-A., Spielhofer, T., 2016. Emergency services' attitudes towards social media: A quantitative and qualitative survey across Europe. Int. J. Hum Comput Stud. 95, 96–111.
Roberts, N., 2000. Wicked problems and network approaches to resolution. International public management review 1, 1–19.
Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., Sievers, N., 2021. Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. Data & Policy 3.
Rufai, S.R., Bunce, C., 2020. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. Journal of public health 42, 510–516.
Saakyan, A., Chakrabarty, T., Muresan, S., 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. arXiv preprint arXiv: 2106.03794.
Samaras, L., García-Barriocanal, E., Sicilia, M.-A., 2020. Syndromic surveillance using web data: a systematic review. Innovation in Health Informatics 39–77.
Schillinger, D., Chittamuru, D., Ramírez, A.S., 2020. From "infodemics" to health promotion: a novel framework for the role of social media in public health. Am. J. Public Health 110, 1393–1396.
Shahi, G.K., Nandini, D., 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. arXiv preprint arXiv:2006.11343.
Shepherd, H.E., Atherden, F.S., Chan, H.M.T., Loveridge, A., Tatem, A.J., 2021. Domestic and international mobility trends in the United Kingdom during the COVID-19 pandemic: an analysis of facebook data. Int. J. Health Geographics 20, 1–13.
Sher, L., 2020. The impact of the COVID-19 pandemic on suicide rates. QJM: An International Journal of Medicine 113 (10), 707–712.
Shofiya, C., Abidi, S., 2021. Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. Int. J. Environ. Res. Public Health 18, 5993.
Sobaih, A.E.E., Hasanein, A.M., Abu Elnasr, A.E., 2020. Responses to COVID-19 in higher education: Social media usage for sustaining formal academic communication in developing countries. Sustainability 12, 6520.
Stechemesser, A., Wenz, L., Levermann, A., 2020. Corona crisis fuels racially profiled hate in social media networks. EClinicalMedicine 23.
Stephens, M., Poorthuis, A., 2015. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. Comput. Environ. Urban Syst. 53, 87–95.
Sutton, J., Renshaw, S.L., Butts, C.T., 2020. The first 60 days: American public health Agencies' social media strategies in the emerging COVID-19 pandemic. Health security 18, 454–460.
Tan, M.J.Z., 2021. Topic extraction and sentiment analysis of subreddit (r/Coronavirus). Final Year Project (FYP), Nanyang Technological University.
Tankovska, H., 2021. Number of social media users 2025| Statista. Statista. https://www. statista. com/statistics/278414/number-of-worldwide ….
Tsao, S.-F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., Butt, Z.A., 2021. What social media told us in the time of COVID-19: a scoping review. The Lancet Digital Health 3, e175–e194.
Tsoy, D., Tirasawasdichai, T., Kurpayanidi, K.I., 2021. Role of social media in shaping public risk perception during Covid-19 pandemic: a theoretical review. International Journal of Management Science and Business Administration 7, 35.
Twitter. 2021. Twitter API for Academic Research | Products | Twitter Developer Platform. CrowdTangle. Retrieved June 14, 2022, from https://developer.twitter.com/en/products/twitter-api/academic-research.
Tziafas, G., Kogkalidis, K., Caselli, T., 2021. Fighting the COVID-19 infodemic with a holistic BERT ensemble. arXiv preprint arXiv:2104.05745.
Velasquez, N., Leahy, R., Restrepo, N.J., Lupu, Y., Sear, R., Gabriel, N., Jha, O., Goldberg, B., Johnson, N., 2021. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Sci. Rep. 11, 1–8.
VICINITAS, 2018. Research on 100 Million Tweets: What it Means for Your Social Media Strategy for Twitter.
Vishwamitra, N., Hu, R.R., Luo, F., Cheng, L., Costello, M., Yang, Y., 2020. On analyzing covid-19-related hate speech using bert attention. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 669–676.
Wang, J., Zhou, Y., Zhang, W., Evans, R., Zhu, C., 2020a. Concerns expressed by Chinese social media users during the COVID-19 pandemic: content analysis of Sina Weibo microblogging data. Journal of medical Internet research 22, e22152.
Wang, S., Huang, X., Hu, T., Zhang, M., Li, Z., Ning, H., Corcoran, J., Khan, A., Liu, Y., Zhang, J., 2022. The times, they are a-changin': tracking shifts in mental health signals from early phase to later phase of the COVID-19 pandemic in Australia. BMJ Global Health 7, e007081.
Wang, X., Zou, C., Xie, Z., Li, D., 2020b. Public opinions towards covid-19 in california and new york on twitter.
Wang, Y., Hao, H., Platt, L.S., 2021. Examining risk and crisis communications of government agencies and stakeholders during early-stages of COVID-19 on Twitter. Comput. Hum. Behav. 114, 106568.
Waseem, Z., Hovy, D., 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter, Proceedings of the NAACL student research workshop, pp. 88-93.
Weber, E.P., Khademian, A.M., 2008. Wicked problems, knowledge challenges, and collaborative capacity builders in network settings. Public administration review 68, 334–349.
Wei, Y., Wang, J., Song, W., Xiu, C., Ma, L., Pei, T., 2021. Spread of COVID-19 in China: analysis from a city-based epidemic and mobility model. Cities 110, 103010.
Weibo-Sina, 2017. Weibo-Sina Weibo user report on 2017.
Weller, K., Kinder-Kurlanda, K.E., 2016. A manifesto for data sharing in social media research. In: Proceedings of the 8th ACM Conference on Web Science, pp. 166–172.

Wich, M., Räther, S., Groh, G., 2021. German Abusive Language Dataset with Focus on COVID-19, Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), pp. 247-252.
Williams, M.L., Burnap, P., Javed, A., Liu, H., Ozalp, S., 2020. Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. The British Journal of Criminology 60, 93–117.
Wojcik, S., Hughes, A., 2019. How Twitter users compare to the general public. Internet, Science & Tech, Pew Research Center.
Xu, P., Dredze, M., Broniatowski, D.A., 2020. The twitter social mobility index: Measuring social distancing practices with geolocated tweets. Journal of medical Internet research 22, e21499.
Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., Zhu, T., 2020. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. Journal of medical Internet research 22, e20550.
Yamamoto, Y., Kumamoto, T., Nadamoto, A., 2014. Role of emoticons for multidimensional sentiment analysis of Twitter. In: Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, pp. 107–115.
Yang, A., 2020. The issue niche theory of nongovernmental and nonprofit organizations' interorganizational network ecology. Communication Theory 30, 41–63.
Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F., 2017. Big Data and cloud computing: innovation opportunities and challenges. Int. J. Digital Earth 10, 13–53.
Yang, K. C., Pierri, F., Hui, P. M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The COVID-19 infodemic: twitter versus facebook. Big Data & Society, 8(1), 20539517211013861.
Ye, X., Andris, C., 2021. Spatial social networks in geographic information science. International Journal of Geographical Information Science. https://doi.org/10.1080/13658816.2021.2001722.
Ye, X., Jourdan, D., Lee, C., Newman, G., Van Zandt, S., 2021. Citizens as sensors for small communities. Journal of Planning Education and Research. https://doi.org/10.1177/0739456X211050932.
Yin, H., Yang, S., Li, J., 2020. Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media, International Conference on Advanced Data Mining and Applications. Springer, pp. 610-623.
Yoo, W., 2019. How risk communication via Facebook and Twitter shapes behavioral intentions: The case of fine dust pollution in South Korea. Journal of Health Communication 24, 663–673.
Yu, M., Li, Z., Yu, Z., He, J., Zhou, J., 2021. Communication related health crisis on social media: a case of COVID-19 outbreak. Current issues in tourism 24, 2699–2705.
Zachreson, C., Mitchell, L., Lydeamore, M.J., Rebuli, N., Tomko, M., Geard, N., 2021. Risk mapping for COVID-19 outbreaks in Australia using mobility data. J. R. Soc. Interface 18, 20200657.
Zarei, K., Farahbakhsh, R., Crespi, N., Tyson, G., 2020. A first instagram dataset on covid-19. arXiv preprint arXiv:2004.12226.
Zeng, C., Zhang, J., Li, Z., Sun, X., Olatosi, B., Weissman, S., Li, X., 2021. Spatial-temporal relationship between population mobility and COVID-19 outbreaks in South Carolina: time series forecasting analysis. Journal of medical Internet research 23, e27045.
Zhang, X., Yang, Q., Albaradei, S., Lyu, X., Alamro, H., Salhi, A., Ma, C., Alshehri, M., Jaber, I.I., Tifratene, F., 2021. Rise and fall of the global conversation and shifting sentiments during the COVID-19 pandemic. Humanities and social sciences communications 8, 1–10.
Zhou, X., Chen, L., 2014. Event detection over twitter social media streams. VLDB J. 23, 381–400.
Zhou, X., Mulay, A., Ferrara, E., Zafarani, R., 2020. Recovery: A multimodal repository for covid-19 news credibility research. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp. 3205–3212.
Zhu, Y., Fu, K.-W., Grépin, K.A., Liang, H., Fung, I.-C.-H., 2020. Limited early warnings and public attention to coronavirus disease 2019 in China, January–February, 2020: a longitudinal cohort of randomly sampled Weibo users. Disaster medicine and public health preparedness 14, e24–e27.

Further reading

Docquier, F., Golenvaux, N., Nijssen, S., Schaus, P., Stips, F., 2021. Cross-border mobility responses to covid-19 in Europe: new evidence from facebook data. Manuscript (Université catholique de Louvain).
Gao, Z., Wang, S., Gu, J., 2020b. Public participation in smart-city governance: A qualitative content analysis of public comments in urban China. Sustainability 12, 8605.
Spelta, A., Pagnottoni, P., 2021. Mobility-based real-time economic monitoring amid the COVID-19 pandemic. Sci. Rep. 11, 1–15.