10

Geolocation-Centric Monitoring and Characterization of Social Media Chatter for Public Health

Abeed Sarker

Abstract

The adoption of social media is currently at an all-time high. More than half of the world has access to social media. The large-scale adoption and growth of social media have demonstrated the benefits and drawbacks of human activities over such platforms. As the digital footprint of human behavior via social media platforms continues to evolve, it is essential to identify strategies and execute actions that can utilize the data generated for the benefit of humankind. Since most of the human footprint on social media is in the form of free text, the field of natural language processing holds substantial promise in converting such data into valuable and actionable knowledge. Geolocation-related metadata available with or inferred from social media posts enable knowledge to be aggregated at various spatiotemporal granularities. Fine-grained area-level insights about human behavior can, for instance, be obtained through social media-based surveillance in close to real time. Geolocation-specific statistics derived from social media data may also be combined with other area-level data from more traditional sources to obtain comprehensive knowledge on chosen topics. Following a brief introduction to social media and natural language processing, the utility of social media data, particularly when combined with geolocation-based information, is discussed. Two examples—COVID-19 and substance use—are used as case studies.

Introduction

Social media refer to Internet-based platforms over which communications involving text, voice, video, and/or images take place. Growth in the use of social media has been primarily driven by social networking websites, which enable people to connect with others and share information. The adoption of social media is currently at an all-time high, and it is estimated that over 4.5 billion people in the world use social media (Statista 2022c). Despite the staggering

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

number of existing social media users, the adoption of such platforms continues to grow. Globally, the most commonly used social network is Facebook. Other popular social networks include but are not limited to Instagram (primarily used for image sharing), Twitter/X (supports microblogging), YouTube (video sharing), and Reddit (topic-specific forums that allow subscribers to remain anonymous if they desire). While social media are still disproportionately popular among younger people, adoption is currently happening at a faster rate among older people according to the Pew Research Center (2021). As demographics shift, it is only a matter of time before the global social media user base becomes quite accurately reflective of the world population. In fact, there is perhaps no other platform currently available that has a better reach than social media.

The widespread use of social media has resulted in the continuous generation of massive data. Such data encapsulate knowledge on essentially any topic. Connected networks also enable the rapid dissemination of information to many people, typically without any geolocation-based limitation. Both the volume of knowledge and the rapidity with which it can spread have the potential to be leveraged to determine and influence population-level behaviors. Consequently, over the last decade, social media platforms have been utilized for a variety of purposes, including (but not limited to) politics, health, and finance. The role of social media in the presidential elections of the United States, for example, has been extensively studied (Bossetta 2018). In the broad field of finance, the power of social media-based communication and behavioral influence was demonstrated in 2021 when a group of subscribers coalesced on a Reddit forum to invest collectively in stocks of GameStop—a company in the United States that was on the verge of bankruptcy according to many institutional investors (Anand and Pathak 2022). It was reported that the collective trading of small investors on Reddit in January 2021 surpassed the previous trading volume record set in 2008 in the New York Stock Exchange by a factor of six (from approximately four billion shares to 24 billion). This collective behavior, which was specific in the United States from the perspective of geolocation, led to a steep, unprecedented rise in the market valuation of the company, by over 1000% in two weeks, baffling institutional and seasoned investors. These events demonstrated the utility of social media and the influence that social media-based human activities can have within specific spatial and temporal windows. The utility of social media-based data for health-related tasks, particularly the possibility of deriving geolocation-specific insights for public health, has been realized over recent years, and substantial research efforts are currently ongoing to utilize data effectively from this ever-growing resource. The primary focus of this chapter is to outline some of the opportunities associated with social media data in the realm of public health, with particular emphasis on the geospatial aspects, and the research challenges that such data present. Two case studies—COVID-19 and substance use—are used to illustrate the use of social media data in real life.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

Social Media and Health

A considerable portion of chatter on social media is concerned with health-related topics. People often share their health problems, discuss treatment options and efficacies, ask questions, describe personal experiences, and provide suggestions, including self-management strategies for myriad health conditions. These discussions capture important information about health topics in an unstructured form. Such data are often referred to as patient-generated big data and have been shown to contain information not available through other, more traditional, sources such as electronic health records and published literature. The information is typically enriched with metadata[1] including geolocations, which may enable spatial aggregation. Even when geolocation information is not explicitly present in the metadata, researchers have developed tools that can estimate geolocation based on other profile-level data (Dredze et al. 2013). In theory, patient-generated social media data can be categorized, aggregated, and analyzed to obtain population- and area-level insights in close to real time and at low cost. Importantly, the data are collected in an unobtrusive manner, which may mitigate biases that typically arise in synthetic experimental settings (Fan et al. 2018). The value of patient-generated data from social media for public health has been realized over recent years, and it is being used increasingly for health-related tasks, such as pandemic surveillance (Chen et al. 2020), pharmacovigilance (Sarker et al. 2015), mental health-related topics (Chancellor et al. 2021), and substance use surveillance (Sarker et al. 2019), to name but a few.

Challenges and Limitations of Social Media Data Processing

While the knowledge contained in social media big data holds considerable promise, the extraction and utilization of such knowledge have been limited for years by our capabilities, or lack thereof, in big data and natural language processing (NLP). NLP is the field of computer science that broadly addresses the problem of automatic understanding of human language in text or verbal form. The flow of natural language, by nature, is nondeterministic; thus, traditional, rule-based computational models are not capable of effectively processing such data. Automatic processing of health-related natural language data from social media is particularly difficult due to the presence of colloquial expressions, misspellings, noise, and context-ambiguous statements. The conversion of health-related social media big data into valuable and actionable knowledge has required the development of advanced NLP and machine-learning (ML) methods—research areas in which enormous advances have been made in recent years. We are therefore at an important point in time in our understanding

[1] Data that summarizes or provides additional information about other data.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

of how best to leverage social media chatter for improving public health, including in the context of geolocation-centric surveillance.

Currently, social media text mining systems employ pipelines of NLP and ML modules that gradually filter out the noise and convert unstructured chatter into aggregated knowledge. NLP methods do not have to rely on explicitly coded rules; rules are learned automatically from the chatter itself via ML methods, which have in some fields reached human-level performances (Montejo-Ráez and Jiménez-Zafra 2022). Within NLP, the most exciting advances have perhaps been brought about by innovative strategies in text representations. Early NLP methods simply used rules such as character patterns (often referred to as regular expressions) on the text-based representations. The incorporation of ML into NLP approaches necessitated the use of vector-based representations of texts, resulting in the creation of sparse vector models such as the bag-of-words[2] and n-gram[3] models. The next leap was in the generation of dense vector representations of words or phrases that required large, unlabeled datasets—of which there is an abundance on the Internet and social media—and the representations were capable of capturing the semantics of the texts such that similar words/phrases would appear close together in vector space (e.g., word2vec models; Mikolov et al. 2013). One shortcoming of such word- or phrase-level models was that they were unable to capture contextual differences; for instance, homonyms[4] would have the same vector representations. These challenges were overcome very recently with the creation of contextual vector models that better captured meanings with large sequences of texts, as in the bidirectional encoder representation from transformers or BERT (Devlin et al. 2019). In addition to these advances in text representation, the capabilities of computing large volumes of data and optimizing complex ML models have also made large strides. While many challenges still exist in the automatic processing of health-related natural language data (e.g., in cases when the relevant concepts are sparse or rare), advances have enabled the utilization of social media chatter for many targeted tasks. Parallel advances in geolocation inference strategies when metadata are not available (Harrigian 2018; Mahajan and Mansotra 2021) have improved our capabilities to conduct geolocation-specific studies.

In addition to the technical challenges associated with mining knowledge from social media, there are limitations inherent to this resource that may not be solvable through technological advances. At the area-level, a major limitation concerns the issue of representativeness. Social media cohorts at specific

2 A text representation commonly used for sentences or documents. Each word is represented as a number, in a list or vector, that specifies its presence/absence or count. Word order is not preserved.

3 A text representation approach that uses contiguous sequences of *n* words. Unlike bag-of-words models, n-gram models preserve information about word sequences.

4 Two or more words with the same meaning or pronunciation but different meanings.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

geolocations are not necessarily representative of the entire population. It is well known that social media data generally underrepresent older age groups while overrepresenting younger ones. Representations may also vary based on the problem being studied. Given a specific health problem, certain segments of the population may be more likely to self-report personal information than others. In the case of substance use, discussed in more detail below, studies have shown that college students are more likely to report nonmedical use of stimulants (Sarker et al. 2016), whereas opioid use may be underreported due to stigma and other factors (Chenworth et al. 2021; Graves et al. 2022). In areas where substance use is criminalized, people may also underreport compared with those living in areas where the issue is treated as a public health issue. The extent to which such under- and overreporting happens among particular cohorts is not fully understood. Absent this knowledge, the best strategy to validate findings from social media is perhaps to compare them with information from traditional sources, such as surveys. Some recent studies have attempted to calibrate problem-specific demographic distribution statistics by developing automatic methods to detect self-reported demographic information (e.g., gender, age-group, and race) and then adjusting the distributions against the distribution detected using the same methods from generic social media data (Yang et al. 2023). Such methods are promising, but the limitations associated with representativeness, and other limitations of social media data, remain important open problems.

Types of Social Networks and Their Contents

While this discussion has mostly projected social media as a sphere of homogeneous data, in reality, that is not the case. Data generated over each social network are unique, as are the utilities associated with the data. Facebook, Twitter/X, Instagram, and Reddit, mentioned above, can be broadly classified as generic social networks. On such networks, subscribers can essentially post on any topic they desire. Consequently, much of the content can simply be considered to be noise, and NLP pipelines processing the chatter must first filter out such noise. The structures of the posts can also be significantly different. Facebook and Reddit, for example, allow long posts. In contrast, Twitter/X posts are length-limited, and so posts are short and often lack context. In addition to these generic social networks, others are dedicated specifically to health-related topics (e.g., MedHelp, PatientsLikeMe), and are generally rich in information but lack metadata, such as geolocation, and attract lower numbers of daily active users. The distinct structures and contents of these social networks have naturally led to distinct digital footprints of their subscriber cohorts. Since subscriber behaviors evolve over time based on the characteristics of the social networks, these differing behaviors provide exciting data for digital ethology.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

Geolocation-Centric Data Analysis and Application Programming Interfaces

As mentioned above, the knowledge encapsulated in social media data goes beyond just natural language chatter or images. Most social networks allow subscribers to make geolocation information visible in their posts. Posts by subscribers on Twitter/X, for example, often contain geolocation information in the form of exact coordinates or information obtained at the city or state level. Such information that complements the contents of social media posts are called metadata. Metadata, such as geolocation and timestamps, are automatically encoded in the posts. Therefore, geolocation-based metadata, when available, can be used to study geolocation-centric digital behavioral patterns among the subscribers. Data posted at specific geolocations by many subscribers at defined time periods can be aggregated and analyzed to study subpopulation-level behavior digitally. Studying aggregated data from many subscribers, as opposed to data from a single subscriber, is invariably more valuable from social media sources. Individual subscribers may not post all information relevant for behavioral or other analyses, but when posts from large numbers of subscribers are aggregated, the most important topics relevant to that group of subscribers tend to become visible as they surface above the rest. Aggregating by geolocations may reveal important distinctions in topics relevant to people from different locations.

Due to the growing utility of social media data, many platforms have made them available through application programming interfaces (APIs), which allow computer programs to connect to the data streams on networks and collect data based on the relevant protocols. Twitter/X, for example, recently released an academic API to support noncommercial research. This API allows researchers to collect the contents of the posts as well as the metadata associated with such posts. Two key metadata elements that have been utilized heavily in research are timestamps and geolocation. Specifically, these meta contents are used to aggregate posts on the platform, given a specific time and topic, and to analyze them over specific geolocations. As mentioned earlier, recent studies have also proposed methods for inferring geolocation from social media posts when explicit geocoding is not available (Dredze et al. 2013; Mahajan and Mansotra 2021). These inference methods have substantially increased the proportion of posts that can be aggregated by geolocation to derive insights. Researchers use geolocation-specific data for tasks such as infectious disease outbreak surveillance, and a number of recent studies have utilized such data to study the COVID-19 pandemic. Below, we look at two case studies that have utilized metadata from Twitter/X for geolocation-centric analyses.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

COVID-19

The pandemic caused by the novel coronavirus provides a current example of a recent global health crisis and has received considerable research attention since the outbreak of the virus in late 2019. This research led to the development of effective mRNA and other vaccines in record time. Ongoing research includes, but is not limited to, studies that focus on the long-term impacts of COVID-19 infection (typically referred to as long COVID), identification and analysis of new mutated variants of the virus, and methods for detecting potential future outbreaks in a timely manner. There is now general understanding and acceptance that future infectious disease outbreaks like COVID-19 may happen. It is also generally accepted that no one mechanism of infectious disease surveillance is by itself sufficient to provide timely alerts; a combination of approaches is required. Localized infectious disease outbreaks, including future variants of COVID-19, can exert tremendous strain on health systems, causing large numbers of deaths (Carinci 2020), as was observed in some countries (e.g., Italy and Spain) as well as in big metropolitan cities (e.g., New York City) during the early waves of the pandemic. Traditional surveillance methods struggled to keep up with the pace of the outbreaks due to the time and effort required to compile data (González-Padilla and Tortolero-Blanco 2020; Gupta and Katarya 2020; Lakamana et al. 2022; Sabouret et al. 2020), which typically comes from sources such as hospitals. The need to develop novel surveillance strategies with the potential to forecast upcoming outbreaks was realized during the COVID-19 outbreak. Infodemiology-oriented data-centric methods for surveillance (Eysenbach 2009), such as those that rely on social media posts, have the potential to detect patterns in chatter associated with geolocation-specific outbreaks and provide timely alerts to relevant health agencies.

Social media proved to be of high utility during the early COVID-19 outbreaks, as it became the primary mode of communication for many, particularly after "lockdowns" and/or "social distancing" measures went into effect. Research during the early months of COVID-19 revealed that social media chatter was rich in first-person reports of COVID-19 positive test results (Guo et al. 2021; Myrick and Willoughby 2022). Many people shared the symptoms they were experiencing, often with day-to-day updates. Research also showed that these self-reports of positive test results and expressions of symptoms can be detected and extracted automatically using NLP methods. In fact, early research showed that about one-third of the people discussed symptoms up to two weeks before they tested positive for COVID-19, and some relevant symptoms were reported before their associations with COVID-19 were common knowledge. For example, the first report of anosmia was observed on Twitter/X in the first week of March, while Google Trends showed that search queries for the symptom peaked after March 20, 2020 (Sarker et al. 2020). This suggests that information specific to COVID-19, including self-reported symptoms, may be available and detectable from social media. Self-reported

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

symptoms represented only a subset of all COVID-19 topics covered in social media chatter, and the knowledge generated often preceded those from other, more traditional, sources. Findings from numerous studies conducted during the COVID-19 pandemic suggest that strategic mining of knowledge from social media chatter can provide valuable and timely insights about infectious diseases, including insights on upcoming outbreaks and potential communities at risk, enabling relevant experts to plan appropriate responses (Callard and Perego 2021; Chen and Wang 2021; Golinelli et al. 2020; Li and Zhang 2021; Matharaarachchi et al. 2022; Tsao et al. 2021; Turiel et al. 2021).

Geolocation-Centric Surveillance

In response to the COVID-19 pandemic, Twitter/X released a customized API for the collection of data in real time. Combining metadata, particularly about geolocation, with NLP methods showed excellent potential for conducting localized surveillance of outbreaks. Since this was the first global pandemic in the era of social media, the data collected since its beginning served as testbeds for innovative approaches. In summary, research showed that social media chatter provided both challenges and opportunities for geolocation-centric monitoring. An outline of these is provided below.

Prior to COVID-19, studies using social media to detect infectious disease outbreak relied primarily on volumes of data that emerged from targeted geolocations. Keyword-based methods were used to detect relevant social media posts (e.g., about *flu* or *flu-like* symptoms) to estimate the incidence of new infections (Broniatowski et al. 2013). Such methods proved too simplistic in the case of COVID-19, as the behaviors of people over digital platforms evolved substantially during the COVID-19 pandemic compared with past infectious outbreaks. First, the volume of chatter associated with COVID-19 was dissimilar to any previous infectious disease outbreak. Once awareness about it increased throughout the population, posts on the topic surged. The vast majority of posts were not first-person experiences; rather, they were mostly focused on sharing information. Second, the large-scale sharing of information also led to the spread of misinformation, which included but not limited to, conspiracies about the pandemic and vaccines, promotion of fraudulent products for the treatment and diagnosis of COVID-19, and the sharing of unverified news. Consequently, an effective strategy for conducting geolocation-centric surveillance proved to be a multistep process.

The first step in utilizing social media chatter for geolocation-centric monitoring of infectious disease outbreaks is to *collect* the right data. For COVID-19, the specialized API provided by Twitter/X served this purpose. The metadata that accompanies the posts often includes geolocation information. Even if a small fraction of the post-level metadata contains geolocation information, having a large volume of data generated on a daily basis enables the collection of very representative geolocation-specific behavioral digital

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

footprints on the chosen topic. For a topic such as COVID-19, however, most of the digital footprint may be noise or misinformation; thus, a crucial step in processing is to *characterize* the data so that irrelevant or unwanted content (e.g., misinformation) can be separated from relevant or useful content (e.g., firsthand reports of positive tests). This characterization problem is also perhaps the hardest to automate. Here, the latest developments in NLP research can help. To solve this characterization step, recent studies have proposed modeling it as a supervised classification[5] problem. Supervised classification is an ML approach where models are trained based on manually annotated data. In this case, efforts were made to annotate data manually to identify misinformation, firsthand reports of symptoms, and informative contents (Gerts et al. 2021). Next, state-of-the-art supervised classification models, such as transformer-based ones (Li and Zhang 2021; Nguyen et al. 2020), were trained on the annotated data and deployed to characterize streaming data automatically. Posts deemed to be relevant are mapped onto their origin location. In the United States, significant correlation (Spearman $r = 0.88$, $p = 0.000$) was found between the distribution of automatically detected posts at the state level and real COVID-19 case counts. Figure 10.1 shows the population-adjusted distribution of automatically characterized Twitter/X posts from early 2021 at the state level.

Social media-based surveillance is not limited to the United States or high-income countries. Due to its widespread global adoption, social media-based surveillance can be implemented almost anywhere in the world. This may be particularly useful for geographical areas where testing services are limited or slow and traditional surveillance of outbreaks is ineffective. To test the utility of social media-based geolocation-centric monitoring outside of the United States, one study focused on India—a large country with a population of over one billion where surveillance at the national level is extremely challenging (Lakamana et al. 2022). In the study, between February and June 2021, over 500,000 tweets about COVID-19 were geolocated to be from India. The chatter about COVID-19 increased almost at the same time as the number of confirmed cases in India, with a high correlation (Spearman $r = 0.944$; $p = 0.001$). The top tweeting states were Maharashtra, Karnataka, Tamil Nadu, and Uttar Pradesh—states that also recorded some of the highest numbers of COVID-19 cases. There was also a significant correlation between the state-level case numbers and the number of tweets emerging from those states (Spearman $r = 0.84$, $p = 0.0003$). Fatigue, dyspnea, and cough were the top reported symptoms emerging from India, and emotion analysis showed a surge in negative emotions in 2021 compared with the previous year. Anxiety levels and concerns about black fungus (mucormycosis) also surged—the latter was known as a problem during the outbreak there. The strong correlations between actual

[5] A machine-learning approach in which labeled examples are used to train a model, which is then used to classify unlabeled samples.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

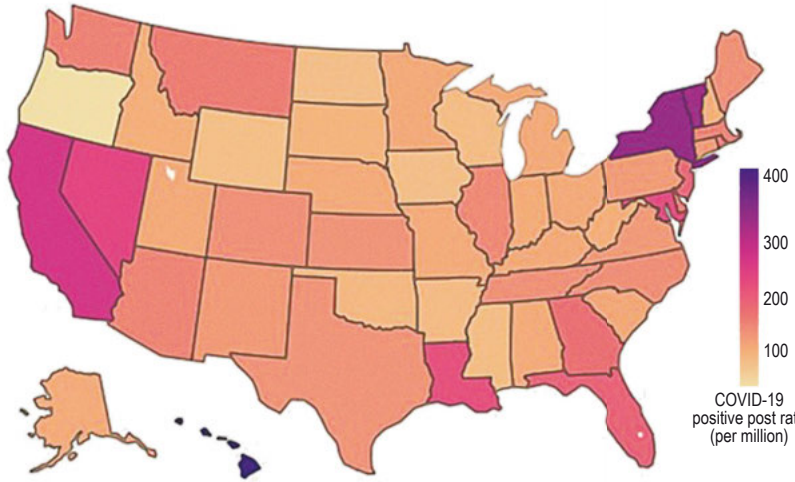

Figure 10.1 Population-adjusted state-level distribution of firsthand reports of positive COVID-19 tests on Twitter/X.

COVID-19 cases and the numbers reported on social media, as discussed above, illustrate the potential of social media-based geolocation-centric pandemic monitoring. With the growing adoption of social media globally, it has the potential to serve as a future platform for geolocation-centric surveillance on a global level. In fact, social media-based surveillance has the potential to reach populations that are hard to reach via traditional surveillance mechanisms—in close to real time and at low cost.

Substance Use

Social media platforms have emerged as potential sources of knowledge for studying topics about which information is either not available or scarce to obtain from traditional sources. One such topic is substance use and substance use disorder. Substance use and its impact constitute a major public health problem globally, and in some countries, like the United States, it is currently considered to be a national crisis. In 2020, over 90,000 Americans died from drug overdoses according to the National Center for Health Statistics (2021), and more than 100,000 overdose deaths occurred in the 12 months leading up to December 2021, an average of over 270 deaths per day. Whereas nonmedical use of prescription medications has historically contributed significantly to the drug overdose epidemic, recent years have seen notable increases in the use of synthetic opioids and psychostimulants. The current epidemic of substance

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

use-related deaths and substance use disorder, including opioid use disorder, is the continuation of decades of constantly evolving trends (Jalal et al. 2018). Within the United States, inequitable access to treatment and enforcement of drug use laws have led to racial disparities in substance use, addiction, treatments, and outcomes (Sanmartin et al. 2020). Over recent years, many studies have highlighted disparities in the treatment of people who use substances that can be traced to socioeconomic status, race/ethnicity, gender identity, community, criminal history, and health-care coverage (Burlew et al. 2021; Lagisetty et al. 2019). The COVID-19 pandemic exacerbated the substance use epidemic, disproportionally affecting communities of color and minority populations (Volkow and Blanco 2021). It has been realized that the implementation of public health approaches to fight the substance use crisis across the globe needs to be multifaceted, focusing on evidence-based programs (Becker et al. 2021), actively addressing barriers to treatment, such as treatment access and stigma (Volkow 2020), and improved surveillance of emerging substance use trends (Kolodny and Frieden 2017; Strickland et al. 2019). Surveillance must be timely to detect emerging "waves" of the epidemic, which is currently believed to be in the early phases of a "fourth wave" in the United States, involving polysubstance use, illicit fentanyl analogs and stimulants (Ciccarone 2021), and responses to these evolving trends need to be tailored to the underlying needs of the affected populations.

A necessary aspect of curbing the epidemic of substance use is to obtain insights about its trends in a timely manner so that responses can be executed accordingly. Traditional surveillance systems consist of surveys (e.g., those conducted by the National Survey on Drug Use and Health, NSDUH), poison control centers, hospital data about treatment admissions and discharge, overdose-related emergency department visits, overdose death records, and others. Such traditional surveillance systems have considerable lags associated with them (Flores and Young 2021). For mortality data, for example, there is a lag time of 12 to 18 months (Anwar et al. 2020). Due to these major delays, emerging trends can only be detected and understood retrospectively. Here, social media can potentially provide an excellent source of real-time information. Indeed, the utility of social media for conducting substance use surveillance (toxicovigilance) has been realized in recent years, resulting in a fast increase in the number of studies exploring social media for substance use-related topics. Social media sources hold substantial promise for toxicovigilance research and signals comparable to NSDUH surveys and NEDS[6] can be discovered from social media via automatic characterizing and mapping of data (Chary et al. 2017; Sarker et al. 2019). Social media are also well-suited for studying aggregated behaviors from targeted cohorts since the social media user base is fairly diverse. Social media data can potentially be used to understand

[6] NEDS (Nationwide Emergency Department Sample) is part of a family of databases and software tools developed for the Healthcare Cost and Utilization Project.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

substance use and substance use disorder among subpopulations such as those who have rising rates of overdose deaths and worse treatment outcomes (e.g., Black and Hispanic populations), lower chances of seeking treatment (e.g., women), or are often excluded (e.g., uninsured).

Social Media-Based Monitoring Strategies

Strategies for conducting monitoring of social media chatter effectively for substance use are similar, in principle, to those for COVID-19 described above. Unlike COVID-19, however, substance use-related chatter is sparse, so the data collection process requires additional innovations. In the past, studies have used automatic, data-centric methods for generating street names and misspellings of substances for data collection (Sarker and Gonzalez-Hernandez 2018). Following data collection, supervised ML needs to be applied to filter out most of the posts that mention substances but are not reports of personal use. Once self-reported substance use posts are identified, they can be mapped to geolocations to obtain an understanding of how substance use is distributed spatially at specific time periods. Figure 10.2 shows the distribution of county-level substance use-related chatter in the United States in 2019, estimated purely via automatic characterization of Twitter/X data.

Research that attempted to establish social media as a potential source for geolocation-centric monitoring had to first validate whether signals detected from these resources were meaningful. Since it is not possible to ascertain if individual social media subscribers at specific geolocations are self-reporting accurate information, this validation focused on comparing aggregated social media data on substance use with established traditional sources. In the case of substance use, these established sources include, for example, overdose deaths from the CDC WONDER database (Spencer et al. 2022) and national surveys such as the NSDUH (SAMSHA 2017, 2020). A study conducted using this strategy of geolocation-centric analysis showed that for the state of Pennsylvania, estimates derived from Twitter/X about opioid use were correlated with opioid overdose-related deaths (Spearman $r=0.331$, $P=.004$) at the county level (Sarker et al. 2019). At the substate level, tweet-level estimates were also found to be correlated with prescription opioid use (Spearman $r = 0.346$), illicit drug use (Spearman $r=0.341$), illicit drug dependence (Spearman $r=0.495$), and illicit drug dependence or abuse (Spearman $r=0.401$). This study demonstrated the utility of analyzing geolocation-specific patterns of Twitter/X chatter on substance use, as it can be applied to understand behavior at a large scale accurately and in close to real time. Social media-based monitoring thus offers the possibility of detecting patterns faster than any other traditional form of surveillance.

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

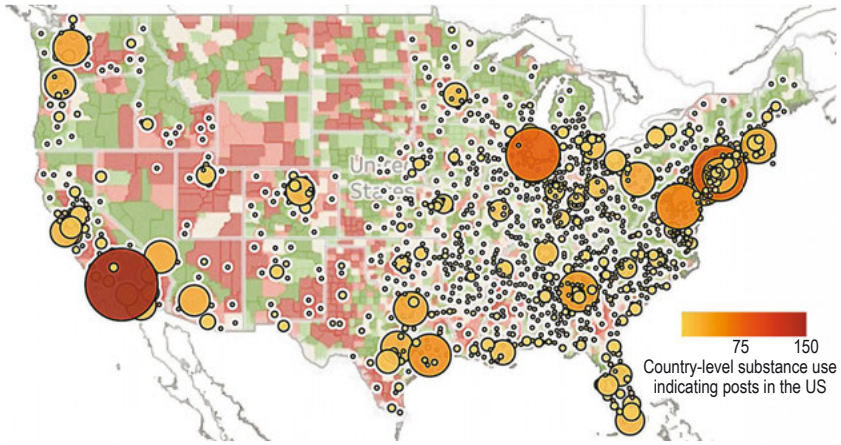

Figure 10.2 Self-reported substance use rates in the United States at the county level on Twitter/X.

Ethical Considerations of Utilizing Social Media Data

The rapid growth of social media and its utility in digital ethology raises important questions regarding the ethical considerations that need to be made when using such data (see also Medeiros et al., this volume). Because of the evolving nature of this research area, there are currently no standardized and universally accepted guidelines for the usage of social media data in health-related or other tasks. The protocols for data use are primarily driven by the organizations behind the social networks and their end-user agreements. For research within the broader medical domain, the protocols for the inclusion of social media data in research are largely guided and approved by institutional review boards. By and large, these boards attempt to ensure that the inclusion of data from social media does not pose any additional risks to the people whose data are being used. Generally speaking, the use of data is considered to be acceptable as long as the data are publicly available. Over recent years, researchers in this space have made efforts to reach consensus regarding the acceptable use of data. Many research groups have also outlined efforts to promote safe use of the data that go beyond what is required by the data use agreements specified by the social networking companies. These efforts include, for example, the removal of user data from studies if subscribers themselves delete their data or make their data private.

Due to the evolving nature of the data and research in this sector, ensuring standards for the ethical use of such data for digital ethology is a moving target. This will perhaps continue to be the case in the near future, much like the field of artificial intelligence itself. This fact is being increasingly recognized by

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813

researchers in this space, and efforts are in place to reach consensus collaboratively and/or raise awareness about concerns.

Conclusions

The evolution of social media, its large-scale adoption, and the rapid advances in data science have opened up unprecedented opportunities for digital ethology. Here, I have focused specifically on the utility of geolocation-centric social media chatter analysis for public health tasks and have outlined two case studies. As the digital footprint of human civilization on social media continues to grow, it is reasonable to expect new opportunities and challenges will arise in the future. The currently known limitations of this data source will also likely evolve over time. The evolving nature of research in this domain means that ethical guidelines will evolve. Consequently, it is imperative that experts from different domains and stakeholders with diverse intentions collaborate to establish protocols that will ensure the responsible use of such data, leveraging it for the common good and minimizing potential harm.

Acknowledgment

The research on social media-based substance use surveillance was supported in part by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH) in the United States (DA057599, DA046619).

From "Digital Ethology: Human Behavior in Geospatial Context,"
edited by Tomáš Paus and Hye-Chung Kum. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978026254813