Journal of the American Medical Informatics Association, 27(2), 2020, 315–329
doi: 10.1093/jamia/ocz162
Advance Access Publication Date: 4 October 2019
Review

Review

Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework

Abeed Sarker ,[1] Annika DeRoos,[2] and Jeanmarie Perrone[3]

[1]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, USA, [2]College of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and [3]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Corresponding Author: Abeed Sarker, PhD, Department of Biomedical Informatics, Emory University School of Medicine, 101 Woodruff Circle, Office 4101, Atlanta, GA 30322, USA; abeed@dbmi.emory.edu

Received 9 July 2019; Revised 14 August 2019; Editorial Decision 15 August 2019; Accepted 0 Month 0000

ABSTRACT

Objective: Prescription medication (PM) misuse and abuse is a major health problem globally, and a number of recent studies have focused on exploring social media as a resource for monitoring nonmedical PM use. Our objectives are to present a methodological review of social media–based PM abuse or misuse monitoring studies, and to propose a potential generalizable, data-centric processing pipeline for the curation of data from this resource.

Materials and Methods: We identified studies involving social media, PMs, and misuse or abuse (inclusion criteria) from Medline, Embase, Scopus, Web of Science, and Google Scholar. We categorized studies based on multiple characteristics including but not limited to data size; social media source(s); medications studied; and primary objectives, methods, and findings.

Results: A total of 39 studies met our inclusion criteria, with 31 (~79.5%) published since 2015. Twitter has been the most popular resource, with Reddit and Instagram gaining popularity recently. Early studies focused mostly on manual, qualitative analyses, with a growing trend toward the use of data-centric methods involving natural language processing and machine learning.

Discussion: There is a paucity of standardized, data-centric frameworks for curating social media data for task-specific analyses and near real-time surveillance of nonmedical PM use. Many existing studies do not quantify human agreements for manual annotation tasks or take into account the presence of noise in data.

Conclusion: The development of reproducible and standardized data-centric frameworks that build on the current state-of-the-art methods in data and text mining may enable effective utilization of social media data for understanding and monitoring nonmedical PM use.

Key words: social media, prescription drug misuse, substance abuse detection, natural language processing, text mining

INTRODUCTION

Prescription medication (PM) abuse (we use the terms *abuse*, *misuse*, and *nonmedical use* interchangeably in this article to represent all forms of use that are not medically prescribed, unless explicitly stated otherwise) is a major public health crisis that has reached epidemic proportions in many countries including the United States.[1] According to a report published in 2011 by the Drug Abuse Warning Network, about half of all emergency department visits for

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contactjournals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

drug misuse were attributed to PMs.[2] A national survey conducted in 2014 showed that over 50 million people in the United States have used PMs nonmedically—a significant portion of which can be classified as abuse.[3] Commonly abused PMs include opioids, depressants and stimulants,[4] and the consequences range from minor side effects such as nausea to serious adverse outcomes including addiction and death. Owing to the rapidly escalating morbidity and mortality, the problem is now receiving international attention, particularly for opioids and their relation to illicit analogs such as heroin and fentanyl.[5] Despite the enormity of the problem, there is a lack of surveillance mechanisms that would enable investigations on the factors contributing to PM abuse, the natural history of the individuals who develop substance use disorders, and the characteristics of the populations affected (eg, age and gender) by distinct classes of abuse-prone PMs. This is emphasized in a recent study delineating 10 steps that the United States government should take to curb the opioid epidemic, where the top suggestion was new and innovative methods of surveillance.[6]

The 2016 National Drug Threat Assessment Summary published by the Drug Enforcement Agency (DEA) revealed that the number of deaths involving PMs has outpaced those from cocaine and heroin combined, for every year since 2002,[7] with approximately 52 people dying each day in the United States from PM overdose. More recently, the Centers for Disease Control and Prevention published a report[8] showing that in the year 2017, there were 70 237 deaths due to drug overdose, of which 17 029 were attributable to prescription opioids, 11 537 to benzodiazepines and 5269 to antidepressants.[9] A portion of these deaths were due to coingestion, and more than half of these deaths involved an opioid, including prescription opioids.[10] Statistics from the WONDER database[11] suggest that overdoses from prescription opioids were a pivotal factor in the 15-year increase in opioid overdose deaths, with the sales of pain-related PMs quadrupling since 1999. This multifold increase in the prescribing and sales of pain medications occurred despite the total volumes of office-based physician visits and emergency department visits due to pain as the primary symptom remaining stable from 2000 to 2010.[12,13] While the long-term impact and costs of prescription opioids are now well understood, less is known about other classes of PMs,[14] although the recently published survey by the Substance Abuse and Mental Health Services Administration presents some alarmingly high numbers.[15] The survey, which estimated abuse based on self-reports, revealed the following statistics: 3.3 million Americans misused opioid pain relievers, 2.0 million misused tranquilizers (eg, benzodiazepines and muscle relaxants), 1.7 million misused stimulants (eg, Adderall), and 0.5 million misused sedatives (eg, zolpidem). Financial costs associated with PM abuse have been on the rise as well. Prescription opioid abuse alone amounted to an estimated total cost of $55.7 billion in 2007[16] and $78.5 billion in 2013,[17] and recent estimates made by the Centers for Disease Control and Prevention suggest that PM misuse costs health insurers up to $72.5 billion annually in direct healthcare costs.[18]

Owing to the enormity of the problem of drug abuse and overdose, the White House announced widespread programs in 2015, which included monitoring and raising awareness about PM abuse, particularly among young people.[19] In an earlier report by the Office of National Drug Control Policy, 4 major areas of focus were detailed, including the improvement of tracking and monitoring techniques to detect and prevent diversion and abuse.[20] Current PM abuse monitoring strategies are aimed primarily at distributors and licensed practitioners. The DEA requires that wholesalers have monitoring programs in place to identify suspicious orders. For licensed, prescribing health practitioners, most states have Prescription Drug Monitoring Programs, and pharmacies are required to report the patients, prescribers, and specific medications dispensed for controlled substances. This data is used by prescribers and law enforcement agencies to identify and limit possible medication abuse. Data at the national level is obtained through large-scale surveys by the DEA and others.[3,7] These surveys are expensive to conduct and there are significant lags between the survey dates and the release of the results (eg, report for the 2016 National Survey on Drug Use and Health was made available in September 2017). Current PM monitoring programs are also plagued with numerous limitations, with efficacies varying widely.[21] Other existing control measures and interventions lack critical information such as the patterns of usage of various PMs and the demographics of the users. Such information can be crucial in designing control measures and outreach programs. For example, warnings to deter PM abuse might be more successful if broadcast during high abuse periods, if known. In response to the necessity of identifying novel strategies for monitoring PM abuse, the National Institute on Drug Abuse launched PA-18-058,[22] encouraging applicants to "develop innovative research applications on prescription drug abuse," "examine the factors," and "characterize this problem in terms of classes of drugs abused and combinations of drug types, etiology of abuse, and populations most affected."

Social media and medication abuse

Recent studies, including our preliminary studies on the topic,[23–26] have validated the use of social media as a platform for monitoring PM abuse. For example, they have shown that although nonmedical users of PMs may not voluntarily report their actions to medical practitioners, their self-reports are often detectable in the social media sphere.[23,24,27] To summarize, these studies have shown that (1) many people publicly self-report PM abuse information in social media, (2) automatic natural language processing (NLP) and machine learning methods are capable of detecting PM abuse-indicating posts, and (3) additional information such as temporal patterns of abuse and common coingestion behaviors can be detected from social media chatter. The Social Media Fact Sheet[28] from Pew Research Center shows that currently 69% of all adult Americans use social media, with particularly high numbers for younger adults (86% for 18- to 29-year-olds; 80% for 30- to 49-year-olds), and the trend of adoption is still upward. Similar trends are also visible globally. Social media may also provide access to communities and information generated through social interactions that may not be available from other sources.[29,30] Thus, social media presents a unique opportunity to study PM abuse at the population level, and discover unique information.

Challenges of social media–based text mining frameworks

Social media provides unfiltered information in near real time, posted by people from diverse demographic groups.[28,31–35] While the volume of data available from this resource is an asset, proper utilization of this data for knowledge discovery is challenging. Knowledge from social media must be automatically curated, as it is not feasible to process such big data manually. Identifying and filtering out relevant data automatically is arduous, requiring customized methods. Knowledge generation typically requires standardization of the data, which in turn requires advanced NLP methods to parse the texts. The language used in social media is unique and

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

complicated—due to the presence of colloquialisms, misspellings, emojis and ambiguities, and often the lack of context.[36,37] Additionally, the language in social media is ever evolving, requiring the development of adaptable, intelligent systems that can evolve with the data. Consequently, while early works attempted to manually create static consumer health vocabularies from social networks and online health comminities,[38,39] some recent research tasks have attempted to develop data-centric methods for automatically discovering common nonstandard consumer health terms[40] and misspellings.[41] PM abuse-related chatter also present mining challenges that illicit drug abuse-related chatter does not present. For example, any expression of consumption of illicit drugs is by definition abuse. However, for PMs, consumption information may represent medical use, misuse or abuse, consequently complicating automated mining further.

MATERIALS AND METHODS

Data search and selection

We searched the databases Medline and Embase, the citation database Scopus and Web of Science, and Google Scholar to find relevant articles published within the last 15 years. We searched for keywords indicating *social media AND prescription medication AND abuse*. Besides searching the databases, we also reviewed the reference lists of studies that met our inclusion criteria, to find additional related studies that may not be identifiable by our keyword-based approaches (eg, studies naming specific medications and utilizing social media data along with data from other sources). Table 1 presents the variants of the keywords used for each of the 3 categories.

We sorted the search engine results by relevance, filtered a selective set for review, and obtained their full texts. We included articles if the titles or abstracts suggested that they used data from social media for detecting, characterizing or studying PM abuse or misuse. Studies that met our inclusion criteria were those that presented original data, utilized any internet-based resource of consumer-generated data (eg, online health communities, forums, message boards, social networks), and presented qualitative or quantitative analyses or well-defined outcomes or results that were directly relevant to at least 1 PM. We included articles that employed manual analysis as well as those that employed NLP or machine learning approaches. We excluded studies that solely focused on illicit drug abuse or trade, or utilized sources such as electronic health records or published literature. Studies were also excluded if they only described clinical trials or extracted information from medication labels suggesting possibility of abuse, if they were news articles or other non–peer-reviewed sources, or if they were not published in English. Additionally, we excluded short commentaries, letters, and responses, unless they provided methodological insights. Articles focused on computational methodologies, which are not relevant or unique to the PM abuse problem, were also excluded unless they included at least a case study involving a named social network (eg, the study by Yakushev and Mityagin[42] was excluded based on this criterion).

Table 1. Sample search queries used to retrieve articles for this review

Social media	Prescription medication	Abuse
social media	prescription medication	abuse
social network	medication	misuse
forum	drug	use
online health community	substance	usage
discussion board		nonmedical use
twitter		
reddit		
instagram		

Data abstraction

For all the included studies, we abstracted the pertinent information presented in them, such as study sizes, sources of data, medications studied, and the primary objectives, methods and findings of the studies. For study size, we focused on the sample size of the data (eg, number of tweets) and the number of medications. We broadly categorized studies into "big" and "small," with big studies including at least 10 000 posts in the articles' primary analyses. We also identified the medication classes studied, when available (eg, opioids, benzodiazepines). For studies presenting multiple objectives or findings, we focused on the primary ones only or those that are related to misuse or abuse. In our analyses of primary methods and results, we attempted to critique the data processing method(s) employed, the primary contributions of the methods, and the relevance and strengths of the evaluation methods employed.

RESULTS

Data collection

Our searches resulted in an initial set of over 1000 articles. Many of these articles focused more generally on substance abuse (eg, illicit drugs and alcohol) and social media, or PM abuse from non–social media data sources. It was particularly challenging to identify studies that included both prescription and illicit drugs. Based on an inspection of the titles and abstracts of these articles, we selected a sample of 63 articles for further review. From this set, 39 studies—journal articles and conference proceedings—were deemed to meet our inclusion criteria.

The earliest study we identified, which suggested the possibility of utilizing web-based, consumer-generated sources for studying drug abuse, was from 2006.[43] Research on this topic, however, began gaining attention from 2012, with 3 articles published in that year. Since then, generally speaking, there has been an increasing trend in the number of articles published on the topic every year (Figure 1).

Study characterizations

Detailed characterizations of the included studies across several dimensions are summarized in Tables 2 and 3. The articles in the 2 tables are listed in the same chronological order. Table 2 shows the years of publication of the articles, the data sources utilized by the studies, the number of medications, medication categories studied, the sizes of the datasets and whether the datasets could be categorized as big data or not. Twitter has been the most commonly used data source, with 20 (51.3%) studies relying on it. This is particularly due to the early availability and popularity of Twitter's public streaming application programming interface (https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html). The application programming interface makes available a sample of public Twitter posts in real-time, which can be collected using keywords for research purposes. Among generic social networks, other than Twitter, Instagram and Reddit are increasing in popularity due to the growing user bases and the typically public nature of the posts.

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

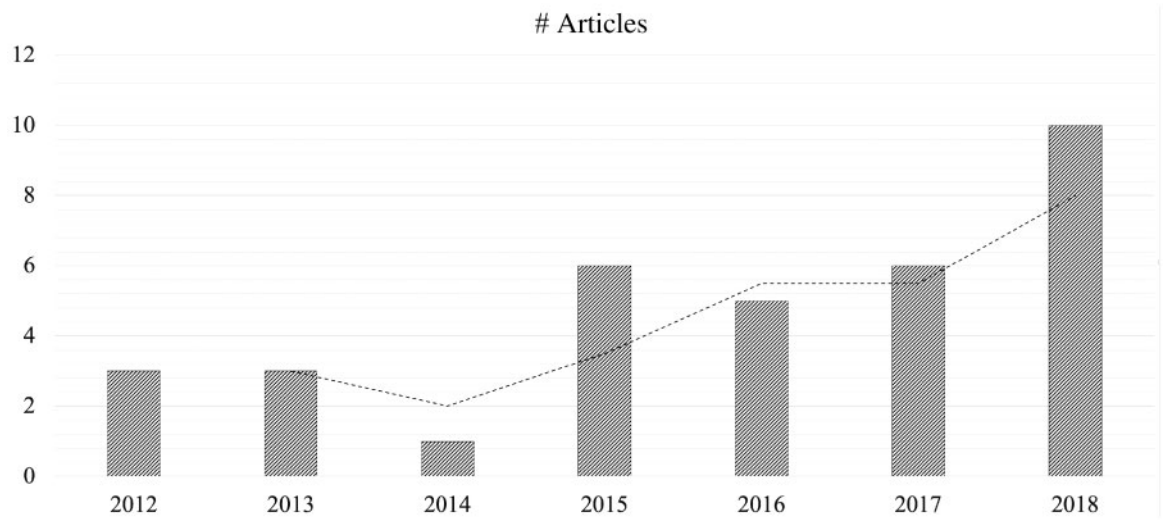

Figure 1. Number of articles meeting the inclusion criteria of our review from 2012 to 2018. The figure does not include the year 2019 because full data will not be available until the end of the year.

Many studies attempted to utilize specialized topic-oriented web forums for research.

As depicted in Table 2, only 6 studies focused on a single medication, and at least 10 studies included both prescription and illicit. Opioids have been the most common medication category studied, with 16 (41%) papers focusing solely on this category. This is unsurprising, considering the growing interest in opioids following the opioid crisis in the United States. Based on our categorization threshold for study size, 20 (51.3%) studies included big data, with 3 studies from this set also performing elaborate manual analyses on smaller samples.

Table 3 details the (1) objectives of the included studies, (2) the primary methods employed by them, and (3) their primary findings. A number of studies included multiple objectives, approaches, or findings, and in the table, we focus on the main contributions of the articles according to our review guidelines. The objectives of the articles varied considerably and included studies to assess if social media chatter contained evidence of abuse, characterize chatter about specific medications manually or automatically, assess user sentiments, develop new methods for automating the surveillance of drug misuse or abuse via social media, discover nonstandard names or terms associated with abuse-prone drugs, and analyze the geographic distributions of abuse-related chatter. Methods for data analysis or characterization included manual analyses, and unsupervised and supervised automatic approaches. We now provide a brief summary of the key findings.

Summary of methodologies and findings

Early studies mostly relied on manual analyses and characterizations to ascertain that user posts contained information about misuse or abuse and the types of the information posted. Typical studies manually annotated small samples for further analyses.[47,79] For Twitter, keyword-based approaches were utilized to analyze the volumes of chatter mentioning specific medications over time, followed by analyses of the chatter to better understand the patterns in volume.[25] Following the publication by Cameron et al,[47] many studies employed NLP to parse conversations and better categorize the meanings of the posts, moving beyond keyword-based approaches. More recently, due to the availability of big data and the absence of manually annotated data, some studies have employed unsupervised topic modeling methods such as latent Dirichlet allocation (LDA) to identify themes associated with the chatter mentioning specific substances,[52,56,66] and identify the abuse-associated topics. The evaluation approaches for such studies, however, have been ad hoc in nature, and no standard method has been proposed to determine the performances of the topic generation methods. Only 9 of the reviewed studies employed some form of supervised machine learning using manually annotated data. The performances of the employed methods suggest that such methods are still very much in their exploratory phases and the annotated datasets used are rather small. Due to the sensitive nature of the topic, there is also a lack of publicly available manually annotated data, which has perhaps acted as an obstacle to community-driven method development.

In terms of findings, all studies have reported the presence of important information regarding nonmedical use of PMs—early studies typically verified the presence of such information, while a number of recent studies have attempted to develop methods for automatically detecting and extracting the knowledge contained within the posts. In addition to the presence of abuse-related information, studies reported finding chatter involving illicit trading of drugs, discovering population subgroups engaged in abuse of specific PMs (eg, high prevalence of Adderall usage among college students), quantifying relapse rates during recovery, measuring geographic distributions of misuse, and their associations with other topics (eg, overdose-related deaths). Although some studies reported the presence of noise in generic social networks, none of the proposed unsupervised methods addressed the issue. Supervised methods that apply a classification filter prior to data analysis have the potential of filtering out varying levels of noise. Some studies computed agreement/correlations between social media signals and other sources, such as metrics from National Survey on Drug Use and Health surveys[58] and geolocation-specific overdose deaths.[66] Broadly speaking, there is still a paucity of studies that have proposed full data-centric processing pipelines for automating the use of social media data for monitoring or characterization of PM abuse, or to find novel insights about abuse-prone medications.

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 2. Articles published on social media mining for prescription medication abuse or misuse monitoring; their years of publication, data sources, and medications/drugs of focus, and sizes of datasets studied

Study	Year	Data source	Number of medications	Medications	Medication categories	Data size/number of instances	Big /small data
Schifano et al[43]	2006	Multiple websites	Unspecified	Multiple (prescription and illicit)	Multiple	290 websites (for prescription medications)	Big
McNaughton et al[44]	2012	Unspecified	6	Oxycodone, hydrocodone, hydromorphone, oxymorphone, morphine, tramadol	Opioids	12 838	Big
Davey et al[45]	2012	Unspecified	Multiple	Unspecified	Unspecified	Data from 8 forums	Big
Daniulaityte et al[46]	2012	Unspecified	1	Loperamide	Diarrhea medication	1290	Small
Cameron et al[47]	2013	Unspecified	Multiple	Unspecified	Multiple: cannabinoids, buprenorphine, opioids, sedatives, and stimulants are mentioned.	1 066 502	Big
Hanson et al[25]	2013	Twitter	1	Adderall	Stimulant	213 633	Big
Hanson et al[48]	2013	Twitter	Multiple (all prescription medications)	Multiple	Multiple	3 389 771 initial tweets	Big
McNaughton et al[49]	2014	7 Forums; unspecified	3	Oxycontin (oxycodone), Vicodin (hydrocodone), Dilaudid (hydromorphone)	Opioids	88 484	Big
McNaughton et al[50]	2015	7 Forums; unspecified	1	Tapentadol	Opioid	1 940 121	Big
MacLean et al[51]	2015	Forum77	Unspecified	Unspecified	Opioids	2848	Small
Shutler et al[23]	2015	Twitter	Unspecified	Unspecified	Opioids	2100	Small
Buntain and Goldbeck[52]	2015	Twitter	21	Multiple	Opioids + mostly illicit drugs	821 000 000	Big
Katsuki et al[53]	2015	Twitter	Unspecified	Multiple	Opioids, benzodiazepines + others	1000	Small
Chan et al[54]	2015	Twitter	11 (keywords)	duragesic, fentanyl, hydrocodone, hydros, oxy, oxycodone, oxycotin, oxycotton, vicodin, vikes, oxycontin	Opioids	540	Small
Seaman and Giraud-Carrier[55]	2016	Twitter	73	Multiple	Opioids, benzodiazepines, stimulants, and others (including illicit opioids)	98 691	Big
Ding et al[56]	2016	Twitter	Unspecified	Multiple (prescription and illicit)	Multiple	116 885; 255 annotated	Big/small
Jenhani et al[57]	2016	Twitter	Unspecified	Multiple (prescription and illicit)	Multiple	80 000	Big
Zhou et al[58]	2016	Instagram	Unspecified	Vicodin + other prescription drugs; illicit drugs	Opioids and others	1000 posts initially, followed by 16+ million posts and all posts from 2362 users	Big
Sarker et al[24]	2016	Twitter	4	Oxycodone, Adderall, Quetiapine, and metformin (control medication)	Multiple	6400 annotated; followed by 100 000+ unlabeled posts	Big/small
Anderson et al[59]	2017	Bluelight, Opiophile	1	Bupropion + 2 comparators (amitriptyline and venlafaxine)	Antidepressant	7756	Small

(continued)

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 2. continued

Study	Year	Data source	Number of medications	Medications	Medication categories	Data size/number of instances	Big /small data
Kalyanam et al[60]	2017	Twitter	3	Percocet, OxyContin and Oxycodone	Opioids	11 million	Big
Phan et al[61]	2017	Twitter	Unspecified	OxyContin, Ritalin and opiates + illicit drugs	Opiates (illicit and prescription)	300	Small
Yang et al[62]	2017	Instagram	Unspecified	Multiple (prescription + illicit)	Multiple	4819 from Instagram; 4329 from Google	Small
Chary et al[63]	2017	Twitter	Unspecified	Prescription opioids	Opioids	3 611 528	Big
D'Agostino et al[64]	2017	Reddit	Unspecified	Unspecified	Opioids	100 posts	Small
Cherian et al[65]	2018	Instagram	1	Codeine	Opioid	1156	Small
Graves et al[66]	2018	Twitter	Unspecified	Multiple (prescription + illicit)	Opioids	84 023	Big
Hu et al[67]	2018	Twitter	Unspecified	Multiple (prescription + illicit)	Multiple	More than 3 million raw tweets; 1794 annotated	Big/small
Chary et al[68]	2018	Lycaeum	Unspecified	Multiple (prescription + illicit)	Sedative-hypnotic, hallucinogen, stimulant, nootropic, psychiatric, anticholinergic, analgesic, antipyretic, antiemetic, antihypertensive, cannabinoid, and contaminant	9289	Small
Fan et al[69]	2018	Twitter	Unspecified	Multiple (prescription + illicit)	Opioids	4 447 507 tweets from 4 051 423 users; 19 722 tweets from 2312 users annotated	Big
Bigeard et al[70]	2018	Doctissimo	Unspecified	Multiple (prescription + illicit)	Antidepressants, antixiolytics, and mood disorder drugs	1850 annotated posts	Small
Chen et al[71]	2018	French forums: Atoute, Doctissimo, e-sante, onmeda, sante-medicine	1	Methylphenidate (trade names: Ritalin, Quasym, Concerta, Medikinet)	Stimulant	3443	Small
Pandrekar et al[72]	2018	Reddit	Unspecified	Multiple (prescription + illicit)	Opioids	51 537	Big
Lossio-Ventura and Bian[73]	2018	Twitter	13 (prescription keywords)	Multiple (prescription + illicit)	Opioids	310 323	Big
Hu et al[74]	2018	Twitter	Unspecified	Multiple	Multiple	3 million tweets with 6794 annotated tweets	Big/small
Adams et al[75]	2019	Reddit and Twitter	Unspecified	Opioids, fentanyl, cocaine, methamphetamine, marijuana, and stimulants	Multiple	Not Available or Applicable	Big
Lu et al[76]	2019	Reddit	Unspecified	Unspecified	Opioids	309 528	Big
Tibebu et al[77]	2019	Twitter	14 (prescription keywords)	Multiple (prescription + illicit)	Opioids	2602	Small
Chancellor et al[78]	2019	Reddit	Unspecified	Multiple	Opioids and opioid use disorder recovery drugs	1 446 948 posts from 63 unique subreddits	Big

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 3. Summary of the primary objectives, approaches, and findings from the studies included in this review

Study	Primary objective(s) and/or significance	Primary approach(es)	Primary finding(s)
Schifano et al[43]	First study to explore web forums for drug abuse research. Objective was to analyze data from "web pages" related to information on consumption, manufacture and sales of psychoactive substances.	Manual exploration of search engines using drug names as keywords. User posts from 1633 websites were analyzed primarily for contents (personal intake and/or trading) and stance (pro- vs anti-drug).	18% websites included pro-drug chatter, 10% included harm reduction, and 10% included drug trading. Previously unknown coingestion patterns were discovered.
McNaughton et al[44]	To explore the sentiment expressed by opioid abusers and their endorsement behavior on internet forums. First study to employ automated methods for analyzing social media chatter related to abuse or misuse.	Mixed-effects multinomial logistic regression was applied to model the probability of endorsing, discouraging, mixed, or unclear messages per compound. Endorsement to discouragement ratios were estimated for each compound.	The following list (ordered), in terms of endorsement ratio, was obtained for the included drugs: oxymorphone, hydromorphone, hydrocodone, oxycodone, morphine, and tramadol.
Daniulaityte et al[46]	To analyze nonmedical use of loperamide, as reported on a specific patient forum.	Retrieved posts mentioning from 2005 to 2011. A random sample of 258 posts were manually annotated to identify intent, dosage, and side effects.	The discussion suggested that high doses of loperamide are used to address opioid withdrawal symptoms or as a methadone substitute.
Davey et al[45]	To analyze the key features of drug-related Internet forums and the communities.	Categories, themes, and attributions were manually analyzed from 8 forums (qualitative).	The study identified unique communities of recreational drug users that can provide information about new drugs and drug compounds.
Cameron et al[47]	The development of a semantic web platform called PREDOSE (PREscription Drug abuse Online Surveillance and Epidemiology), designed to facilitate the epidemiologic study of prescription (and related) drug abuse practices using social media.	A drug abuse ontology is used to recognize 3 types of data, namely (1) entities, (2) relationships, and (3) triples. Basic natural language processing approaches are used to extract entities and relationships, and to identify sentiment.	The reported approach obtains 85% precision and 72% recall in entity identification, on a manually created gold standard dataset. In manual evaluation, the system obtains 36% precision in relationship identification, and 33% precision in triple extraction.
Hanson et al[25]	To identify variations in the volume of Adderall chatter by time and geographic location in the United States, as well as commonly mentioned side effects and coingested substances.	Tweets containing the term *Adderall* were collected from November 2011 to May 2012, and a keyword-based approach was used to detect coingested substances and side effects using manual analysis of geolocation clusters and temporal pattern.	Twitter posts confirm Adderall as a study aid among college students. Twitter may contribute to normative behavior regarding its abuse.
Hanson et al[48]	To analyze the networks of users who report abusing/misusing prescription medications.	Tweets mentioning prescription medications were collected from Twitter as well as users mentioning prescription medications multiple times. Social circles of 100 users were analyzed, particularly their discussions associated with prescription drug abuse.	Twitter users who discuss prescription drug abuse online are surrounded by others who also discuss it—potentially reinforcing a negative behavior and social norm.
McNaughton et al[49]	To evaluate the reactions to the introduction of reformulated OxyContin. To identify methods aimed to defeat the abuse-deterrent properties of the product.	Posts spanning over 5 years collected from 7 forums were evaluated before and after the introduction of reformulated OxyContin on August 9, 2010. Qualitative and quantitative analyses of the posts were performed to assess proportions and sentiments.	Sentiment profile of OxyContin changed following reformulation. OxyContin was discouraged significantly more following reformulation. Frequency of posts reporting abuse decreased over time.
McNaughton et al[50]	To assess the amount of discussion and endorsement for abuse of tapentadol and comparator drugs.	Internet messages posted between January 1, 2011, and September 30, 2012, on 7 web forums were evaluated. Proportions of posts and unique authors discussing tapentadol were compared with 8 comparator compounds.	Recreational abusers appeared to be less interested in discussing tapentadol abuse.
MacLean et al[51]	To assess the effectiveness of a specialized forum in helping misusers/abusers of prescription opioids.	A taxonomy describing the phases of addiction was developed, and the activities and linguistic features across phases of use/abuse, withdrawal, and recovery were examined. Statistical classifiers were developed to identify addiction, relapse, and recovery phases.	According to the forum data, almost 50% of recovering abusers relapsed, but their prognosis for recovery is favorable.
Shutler et al[23]	Qualitatively assess tweets mentioning prescription opioids to determine if	Manual categorization of posts into predefined categories—abuse, nonabuse, and	Twitter can be a potential resource for monitoring prescription opioid use, as

(continued)

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 3. continued

Study	Primary objective(s) and/or significance	Primary approach(es)	Primary finding(s)
	they represent abuse or nonabuse, or were not characterizable. To assess the connotation (positive, negative, noncharacterizable).	not characterizable; and, in terms of connotation, positive, negative, and not characterizable.	abuse is commonly described by users (mostly with a positive connotation).
Buntain and Goldbeck[52]	To assess how tweets can augment a public health program that studies emerging patterns of illicit drug use.	The article proposed an architecture for collecting vast numbers of tweets over time. Automatic topic modeling was employed to identify topics, and temporal and geolocation-based analyses were discussed.	An architecture for mining Twitter data for drug abuse monitoring (illicit and prescription).
Katsuki et al[53]	To conduct surveillance and analysis of tweets to characterize the frequency of prescription medication abuse-related chatter, and identify illegal online pharmacies involved in drug trading.	Tweets collected using medication keywords and street names were manually coded to indicate misuse or abuse behavior and attitude (positive/negative). Supervised machine learning automatically identified over 100 000 tweets mentioning abuse or promotion. Word frequency–based experiments identified associations. Geolocations were analyzed for geographic distributions.	The study found a large number of tweets (over 45 000) that directly marketed prescriptions medications illegally. Supervised machine learning showed adequate performance in automatic detection.
Chan et al[54]	To manually analyze opioid chatter from Twitter.	Data was collected from Twitter over 2 weeks and manually coded (eg, personal vs general experiences including nonmedical use, and user sentiments toward opioids) for analysis.	Personal opioid misuse was the most common theme among the tweets analyzed.
Seaman and Giraud-Carrier[55]	To present statistics about volume as well as attitudes toward distribution (selling/buying) and need.	Only a small number (500) of tweets were manually analyzed. New York–based tweets showed that buying/selling and "need" were the most common topics associated with the drug names.	Twitter users often express the need for Adderall and Xanax; chatter related to specific drugs is directly impacted by media events involving such substances.
Ding et al[56]	To detect abuse-related posts and discover new, unknown street names for drugs.	A sample of Instagram posts was annotated for medical use, illicit use, not related, or not sure. Topic modeling (LDA) was used to track changes in hashtags. Hand-annotated tweets were used to identify proportions for abuse-related tweets. Manual analysis of hashtags performed to assess the performance of the word embeddings.	The topic modeling approach retrieves drug-related posts with 78.1% accuracy. Word embeddings learned from social media data are useful for finding new hashtags and street terms associated with abuse.
Jenhani et al[57]	To propose methods for automatically detecting drug-abuse-related events from Twitter.	A hybrid approach consisting of a rule-based component and supervised machine learning is described. Automatically annotated tweets are used for evaluation, showing 0.51 F-score.	Machine learning based approach can detect events not detected by rules. Findings are limited by the fact that only automatically annotated data is used for evaluation, which is prone to errors.
Zhou et al[58]	To explore the possibility of using multimedia data (images and text) to discover drug usage patterns at a fine-grained level with respect to demographics.	Posts were retrieved from Instagram using drug-related hashtags. An initial set of hashtags was used to create a dictionary of hashtags. User demographics, such as age and gender, were predicted using face-image analysis algorithms. Patterns of drug-usage associated with demographics, time and location were then analyzed.	Findings from social media mining are consistent with findings of the NSDUH (qualitatively), even at a fine-grained level.
Sarker et al[24]	To verify that abuse information for abuse-prone medications in social media is higher than non–abuse-prone medications. To assess the possibility of automatically detecting abuse via NLP and machine learning. To compare automatically classified temporal data with past manual analysis.	Manually annotated 6400 tweets to indicate abuse vs nonabuse. Evaluation of automatic classification was performed via 10-fold cross-validation; tests for proportions of abuse-related posts between case and control medications. Compared classified Adderall tweets with past manual analysis.	There is significantly more abuse-related information for abuse prone medications compared with non–abuse-prone medications. Supervised machine learning is an effective approach for automated monitoring.
Anderson et al[59]	To determine if misuse or abuse could be detected via social media listening. To	Posts were collected using generic, brand, and vernacular brand names and were reviewed manually by coders.	Agreement among raters in manual categorization was low (0.448). Analysis of posts revealed that 8.61% refer-

(continued)

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 3. continued

Study	Primary objective(s) and/or significance	Primary approach(es)	Primary finding(s)
	describe and characterize social media posts.		enced misuse or abuse, including routes of intake. Web forums present a valuable new source for monitoring nonmedical use of medications.
Kalyanam et al[60]	To demonstrate that the geographic variation of social media posts mentioning prescription opioid misuse strongly correlates with government estimates of prescription opioid misuse in the previous month.	Tweets were collected from 2012 to 2014, using *opioid keywords*. Tweets were automatically quantified using semantic distance with word centroids. Unsupervised classification/clustering used to group tweets mentioning opioid misuse. Volume of abuse-related chatter was correlated with NSDUH surveys, with separate correlations for different age groups.	Mentions of misuse or abuse of prescription opioids on Twitter correlate strongly with state-by-state NSDUH estimates.
Phan et al[61]	To verify that tweets contain patterns of drug abuse. To study the correlations among different levels of drug usage including abuse, addiction and death, and assess the applicability of large-scale systems for online social network-based drug abuse monitoring.	Manual annotation of opiate-mentioning tweets and basic feature selection methods were developed. Several machine learning classifiers were then trained and evaluated. Word co-occurrence patterns for abuse-indicating tweets were identified and used as features in machine learning experiments. Correlations between words and drug terms were computed.	The best performance was obtained by a decision tree-based classifier, but performance was low compared with human judgment.
Yang et al[62]	To propose a multitask learning method to leverage images from Instagram for recognition of drug abuse. To identify user accounts involved in illicit drug trading.	A multitask learning method was employed for image classification (stage 1) and accounts of interest were identified. Drug-related patterns, temporal patterns, and relational information patterns were detected from the user timelines and potential dealer accounts were detected (stage 2).	A reproducible machine learning model for tracking and combating illicit drug trade on Instagram. The framework can be reused and improved for practical tracking and combating of illicit drug trade on Instagram.
Chary et al[63]	Demonstrate that the geographic variation of tweets mentioning prescription opioid misuse strongly correlates with government estimates in the previous month.	Basic preprocessing was performed on tweets from 2012 to 2014 (signal tweets and basal tweets) collected by keywords linked to prescription opioid use (misspellings as well). Tweets were manually annotated and geodata was collected. Compared tweets with NSDUH.	State-by-state correlation between Twitter and NSDUH data was high. Correlation was strongest in NSDUH data for 18- to 25-year-olds.
D'Agostino et al[64]	To examine the online Reddit community's ability to target and support individuals recovering from opiate addiction.	Collected 100 Reddit posts and their comments from August 19, 2016. Manually annotated the posts/comments according to DSM-5 criteria to determine the addiction phases of individual users.	Demonstrated the supportive environment of the online recovery community and the willingness to share self-reported struggles to help others.
Cherian et al[65]	To characterize information about codeine misuse through analysis of public posts on Instagram to understand text phrases related to misuse.	1156 posts were collected over 2 weeks from Instagram via hashtags and text associated with codeine misuse. Themes and culture around misuse were identified through manual analysis.	50% of reported abuse involved combining codeine with soda (lean). Common misuse mechanisms included coingestion with alcohol, cannabis, and benzodiazepines.
Graves et al[66]	To determine whether Twitter data could be used to identify geographic differences in opioid-related discussion. To study whether opioid topics were significantly correlated with opioid overdose death rate.	Tweets collected using keywords from 2009 to 2015. Topic modeling (LDA) used to summarize contents into 50 topics. The correlations between topic distribution and census region, census division and opioid overdose death rates were quantified.	Selected topics were significantly correlated with county- and state-level opioid overdose death rates.
Hu et al[67]	To build a system for effective drug abuse related data collection from social media and develop an annotation strategy for categorization of data (abuse vs nonabuse) and a deep learning model that can automatically categorize tweets.	More than 800 keywords were used to collect data, followed by crowd-sourced annotation of 4985 tweets. Deep learning model built on small annotated data and evaluated via 10-fold cross-validation. Geographic distribution over 100 000 tweets (positively classified) were analyzed.	The crowd-sourced annotation method enabled annotation at a much faster rate and lower cost. Deep learning model achieved state-of-the-art classification performance. Semantic analysis of tweets revealed drug abuse behaviors. Geolocation-based analysis

(continued)

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 3. continued

Study	Primary objective(s) and/or significance	Primary approach(es)	Primary finding(s)
			enabled the identification of geographic hotspots.
Chary et al[68]	To demonstrate that data concerning polysubstance use can be extracted from online user posts, and that these data can be used to infer novel as well as known coingestion patterns.	Posts were retrieved via web scraping and basic natural language processing methods were applied to identify possible mentions of drugs. Correlation was computed between mentions of pairs of drugs to identify common ingestion patterns based on mentions of drugs.	183 coingestion combinations were discovered, including 44 that had not been studied before.
Fan et al[69]	To propose a novel framework named AutoDOA to automatically detect opioid addicts from Twitter.	Five groups of annotators (18 persons) with domain expertise labeled 19 722 tweets from 2312 users to identify potential addicts. Using only annotations with full agreement, an approach relying on metapath–based similarity was used to perform transductive classification of the users based on the tweets, their likes, and their networks.	Evaluation on annotated data shows that this method outperforms other approaches; A case study on 1132 identified heroin addicts qualitatively show similarities with CDC estimates of overdoses.
Bigeard et al[70]	To create a typology for drug abuse or misuse and methods for automatic detection and propose methods for classification of drug misuses by analyzing user-generated data in French social media.	1850 posts were annotated into 4 categories—misuse, normal use, no use, and unable to decide. Categories were used to create a typology of misuses and to evaluate an automatic system. Several machine learning algorithms were then trained on artificially balanced data to categorize among misuse, no use, and normal use.	Multinomial naïve Bayes is shown to achieve the best performance on the artificially balanced data. The manual categorization of the data reveals an elaborate typology of intentional and unintentional misuse. The annotator agreements are relatively low, showing the difficulty of the misuse annotation task.
Chen et al[71]	To qualitatively analyze posts about methylphenidate from French patient forums including an analysis of information about misuse or abuse.	Data were collected from French social networks that mentioned methylphenidate keywords. Text mining methods such as named entity recognition and topic modeling where used to analyze the chatter, including the identification of adverse reactions.	Analysis of the data revealed cases of misuse of the medication and abuse.
Pandrekar et al[72]	To demonstrate the potential of analyzing social media (specifically Reddit) data to reveal patterns about opioid abuse at a national level	Collected 51 537 Reddit posts between January 2014 to October 2017; evaluated psychological categories of the posts and characterized the extent of social support; performed topic modeling to determine major topics of interests and tracked differences between anonymous and nonanonymous posts.	The information shared on Reddit can provide a candid and meaningful resource to better understand the opioid epidemic.
Lossio-Ventura and Bian[73]	To study and understand (1) the contents of opioid-related discussions on Twitter, (2) the coingestion of opioids with other substances, (3) the trajectory of individual-level opioid use behavior, and (4) the vocabulary used to discuss opioids.	310 323 tweets were collected over 4 months, and 124 143 tweets were included in the study following rule-based filtering. Keyword frequency and co-occurrence based methods were applied to meet the objectives of the study.	Although most of the chatter talked about use of opioids as legitimate pain relievers, there was considerable discussion about misuse or abuse and coingestion of opioids with other substances; 18 new terms for opioids, which were previously not encoded, were discovered.
Hu et al[74]	To establish a framework for automatic, large-scale collection of tweets based on supervised machine learning and crowd sourcing, with a self-taught learning approach for automatic detection.	Data were collected from Twitter using keywords and following an initial annotation by the authors, crowdsourcing was utilized for obtaining reliable annotations. An iterative automatic classification approach is applied where the training data is augmented with machine-classified tweets to improve performance. Both traditional and neural network–based classifiers were experimented with.	The neural network–based (convolutional and recurrent) deep, self-taught learning algorithms outperformed traditional models in the binary classification task with ~86% accuracy.
Adams et al[75]	To demonstrate the benefit of mining platforms other than Twitter, and the	The synonym discovery method was compared for finding terms relevant to mari-	The synonym discovery method yielded more synonyms from Reddit than

(continued)

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

Table 3. continued

Study	Primary objective(s) and/or significance	Primary approach(es)	Primary finding(s)
	use of word embeddings for keyword synonym discovery resulting in increased collected data.	juana and opioids from 2 sources—Twitter and Reddit.	Twitter. Twitter, however, provided more slang terms.
Lu et al[76]	To demonstrate the insights that can be obtained from employing data mining techniques on social media to better understand drug addiction.	Collected 309 528 posts from 125 194 unique Reddit users between January 2012 and May 2018. Used a trained classifier to predict transition from casual drug discussion to drug recovery. Used a Cox regression model to calculate the likelihood of the transition.	Found that certain utterances and linguistic features of one's post can help predict the transition to drug recovery and determined specific drugs that are associated more with transition to recovery, which offers insight into drug culture.
Tibebu et al[77]	To assess if Twitter maybe used as a data source for studying population-level opioid use and perceptions in Canada.	Collected 2602 tweets over 1 month and manually categorized 826 tweets to study usage and perceptions.	The analyzed tweets presented information about medical usage of opioids, impacts of opioid use on family and friends, and drug use in public places. Tweets representing user perceptions were mostly associated with the keywords *heroin*, *fentanyl*, and *opioids*.
Chancellor et al[78]	To assess if Reddit contains information on clinically unverified alternative treatments to opioid use disorder, develop a machine learning approach for discovering posts representing alternative treatments, and identifying commonly reported agents for successful recovery.	A transfer learning approach was developed to automatically detect posts discussing recovery from opioid use disorder and was applied to all the posts collected from 63 subreddits. An approach involving regular expressions and word embeddings is used identify alternative treatments from the positively classified posts.	The transfer learning–based classification approach obtained accuracy of 91.7%, leading to 93 104 recovery posts. Common drugs discovered for alternative treatments included both prescription (eg, Loperamide, Xanax, Valium, Klonopin, gabapentin) and nonprescription (eg, kratom) drugs.

CDC: Centers for Disease Control and Prevention; LDA: latent Dirichlet allocation; NSDUH: National Survey on Drug Use and Health.

DISCUSSION

Our review covers research efforts that have attempted to mine user-posted web and social media data for studying, curating, monitoring, or characterizing PM abuse-related information. The 39 studies that met our inclusion criteria unanimously concluded that social media is a potentially useful resource for studying PM abuse due to the presence of considerable amounts of unfiltered information available. The studies reviewed fall into 3 broad categories from the perspective of methodology employed: (1) manual analysis, (2) automatic unsupervised analysis, and (3) supervised analysis. Most studies employed some form of manual analyses, and these analyses were primarily targeted toward hypotheses generation (eg, "does social media provide information about PM abuse?" and "can we study information about mechanisms of PM abuse from social media?"), and hypotheses testing via manual annotation of samples of data. Such analyses of social media data generated the crucial early hypotheses and helped establish it as a valuable resource for toxicovigilance research. But such analyses are limited to small data samples, are difficult to reproduce, and cannot be used for continuous analysis. Therefore, despite their effectiveness in some cases, manual approaches are not suitable for long-term, data-centric efforts that take advantage of the primary attraction of social media—the continuous generation of big data. We have also reached a point in which further manual validation of hypotheses regarding the presence of abuse-related information at the post level are not required.

Unsupervised approaches have primarily focused on big data to identify trends, for example, through analyses of volume of data to estimate abuse rates at specific time periods or, more recently, topic modeling to identify abuse-related topics associated with selected medications. Volume-oriented unsupervised approaches (eg, keyword based) are capable of tracking interests and discovering trending hidden topics in real-time (eg, via LDA), but studies have shown that only small proportions of the data may present abuse information, and so, such methods are likely to be significantly affected by unrelated chatter, and the conclusions derived may be particularly unreliable when the proportions of abuse indicating posts for specific medications are low. Some of the studies mentioned in Tables 2 and 3 have shown that for certain medications a very minute portion of the social media chatter may be associated with abuse.[24] For example, a significant portion of Twitter chatter mentioning opioids is generated by users sharing general information, such as news articles, rather than personal experiences. This characteristic of the data is not unique to the problem of PM abuse, but is generalizable across social media–based datasets, and has been observed in other studies including influenza and vaccine monitoring,[80] cancer communications,[81] and pharmacogivilace.[36] Thus, especially when working with generic social media data, applying a supervised classification filter before the analysis of topics or trends is perhaps methodologically more robust.

Few studies have employed supervised classification approaches to identify salient information, as supervised learning algorithms require large volumes of data to be manually annotated for training, which is time consuming and expensive. However, supervised approaches, due to their ability to filter out irrelevant information, are likely to have greater longevity in the constantly evolving sphere of social media. The time spent in annotating data for supervised classifications may be valuable for long-term studies and stable systems, provided the annotations follow explicit guidelines and are portable across studies.

Despite the promise of supervised classification approaches, the performances reported by the reviewed systems are typically

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

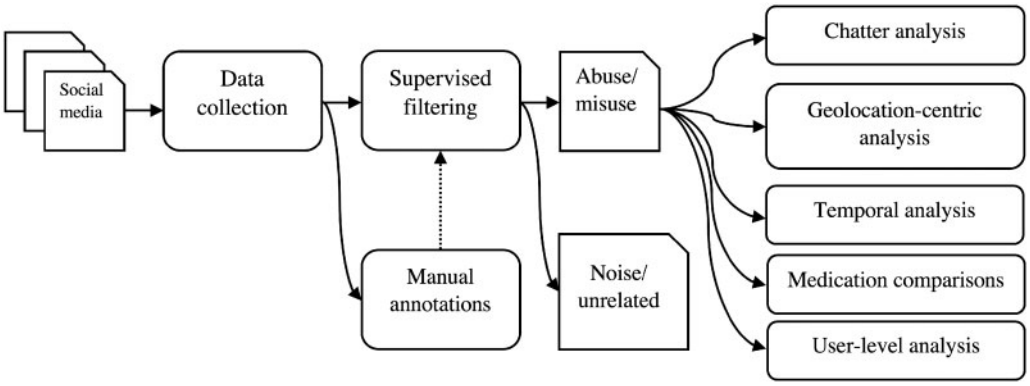

Figure 2. High-level framework for deriving knowledge about prescription medication abuse from social media big data.

low.[24,44,57] This is a known issue for social media data—the text is very difficult to automatically classify due to the factors discussed previously. Social media data can be hard to decipher even for humans, as contents can be ambiguous. Studies that double-annotated sample data, typically reported low agreement rates.[24,59,70] To improve performances of future classification methods, it is essential to increase human agreement rates during annotation tasks. Only 10 (25.6%) reviewed articles[24,44,46,49–51,54,59,65] in our sample reported the creation, presence, or use of detailed annotation guide or guidelines or coding rules which the annotators followed to improve agreement rates. In our view, future research should put more focus on developing thorough annotation guidelines that can be used as reference for annotating data. For researchers from distinct institutions attempting to perform identical tasks, use of publicly available elaborate guidelines will enable the direct comparison of research methodologies (eg, classification performances), even if the data are not shared. There is also a shortage of publicly available annotated data for tasks such as automatic abuse detection. The recent adoption of social media for similar tasks have been accelerated by the creation of publicly available annotated data (eg, for pharmacovigilance).[82] However, there have been no such efforts for studying PM abuse from social media, and such efforts should accelerate the research in this space as well. Such data preparation and release efforts need also consider the potential ethical implications.

CONCLUSION

We conclude our review by proposing a possible data-centric NLP and machine learning framework informed by the extensive review presented in this paper. The proposed framework may be used for monitoring PM abuse from social media and for related research problems within the broader health domain, which have characteristics similar to PM abuse.

Framework for mining social media for prescription medication abuse

Our proposed framework consists of a data processing pipeline that starts from data collection, which is often not trivial for social media–based studies. The data collection strategy has to take into account common misspellings,[41] and street names for medications, as many abuse-prone medications have commonly used street names (eg, "oxy," "percs," "addy," "xanny"; a list of such street names provided by the DEA can be found at: https://ndews.umd.edu/sites/ndews.umd.edu/files/dea-drug-slang-terms-and-code-words-july2018.pdf). Collection is particularly difficult for generic social networks, such as Twitter, due to the presence of large numbers of misspellings and nonstandard terms, compared with targeted online health communities. Following data collection, it is essential to filter out noise or irrelevant posts, which most of the retrieved data are likely to comprise. This is best achieved by classification methods, which not only filter out noise, but may also classify the posts into relevant categories (eg, medical consumption vs abuse). Considering the reported performances of past systems, there need to be future efforts for improving the state of the art in PM abuse classification. These strategies and steps of data collection followed by supervised classification are also applicable to research problems that resemble that of PM abuse monitoring. Such studies, for example, include research on alcohol misuse or abuse,[83,84] and medical and nonmedical consumption of marijuana[85,86] from social media—for both these research topics, like PM abuse, consumption alone, without additional evidence, may not indicate misuse or abuse.

Following the effective removal of unrelated data or noise, the relevant chatter can be passed on for further NLP and machine learning based processing for the discovery of knowledge. In Figure 2, we have specified a few possible studies. For example, once the noise has been removed, it is appropriate to employ unsupervised chatter analysis methods such as topic modeling to discover salient topics closely related to PM misuse or abuse. While topic modeling methods, such as LDA, without any prior filters may retrieve mostly irrelevant latent topics, the application of a classification filter ensures the relevance of the topics to PM abuse. Geotagged social media data, if available, can be utilized to compare abuse or misuse related information across different locations. Similarly, timestamps can be used to analyze temporal patterns of abuse for different medications. Combinations of unlabeled methods, coupled with geolocation and temporal information can be used to compare information about distinct medications (eg, Vicodin and Percocet) and categories of medications (eg, opioids and benzodiazepines). Finally, studying longitudinal data related to abuse from groups of users may enable us to detect cohort-level behavioral patterns and trends.

FUNDING

Research reported in this publication was supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number R01DA046619. The content is solely the responsibility of the authors and

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

AS contributed significantly to the article review and selection process, wrote the majority of the content in the manuscript, and performed critical analysis and comparison of the included studies. AD contributed significantly to the article review and selection process, helped the primary author to summarize included studies, and contributed to the preparation of the manuscript. JP provided toxicology domain expertise for the review, helped identify key articles, and contributed to the manuscript writing and revisions.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Cicero TJ, Ellis MS. Abuse-deterrent formulations and the prescription opioid abuse epidemic in the United States. *JAMA Psychiatry* 2015; 72 (5): 424–30.
2. Substance Abuse and Mental Health Services Administration. Highlights of the 2011 Drug Abuse Warning Network (DAWN) findings on drug-related emergency department visits. https://www.samhsa.gov/data/sites/default/files/DAWN127/DAWN127/sr127-DAWN-highlights.htm. Accessed September 9, 2019.
3. Center for Behavioral Health Statistics. *2014 National Survey on Drug Use and Health: Detailed Tables. Substance Abuse and Mental Health Services Administration*. Rockville, MD: Center for Behavioral Health Statistics; 2015.
4. National Institute on Drug Abuse. *Misuse of Prescription Drugs*. Bethesda, MD: National Institute on Drug Abuse; 2016.
5. Compton WM, Jones CM, Baldwin GT. Relationship between nonmedical prescription-opioid use and heroin use. *N Engl J Med* 2016; 374 (2): 154–63.
6. Kolodny A, Frieden TR. Ten steps the federal government should take now to reverse the opioid addiction epidemic. *JAMA* 2017; 318 (16): 1537–8.
7. Drug Enforcement Administration. *2016 National Drug Threat Assessment Summary*. Springfield, VA: Drug Enforcement Administration; 2016.
8. Centers for Disease Control and Prevention, Opioid Overdose. https://www.cdc.gov/drugoverdose/data/statedeaths.html. Accessed September 9, 2019.
9. National Institute on Drug Abuse. Overdose Death Rates. https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates. Accessed September 09, 2019.
10. Rudd RAR, Seth P, David F, Scholl L. Increases in drug and opioid-involved overdose deaths—United States, 2010–2015. *MMWR Morb Mortal Wkly Rep* 2016; 65 (5051): 1445–52.
11. Centers for Disease Control and Prevention. *Wide-Ranging Online Data for Epidemiologic Research (WONDER)*. Atlanta, GA: Centers for Disease Control and Prevention; 2018. https://wonder.cdc.gov/wonder/help/about-cdc-wonder-508.pdf. Accessed September 09, 2019.
12. Daubresse M, Chang H-Y, Yu Y. Ambulatory diagnosis and treatment of nonmalignant pain in the United States, 2000–2010. *Med Care* 2013; 51 (10): 870–8.
13. Chang H-Y, Daubresse M, Kruszewski SP, Alexander GC. Prevalence and treatment of pain in EDs in the United States, 2000 to 2010. *Am J Emerg Med* 2014; 32 (5): 421–31. doi: 10.1016/j.ajem.2014.01.015
14. Jena AB, Goldman DP. Growing Internet use may help explain the rise in prescription drug abuse in the United States. *Health Aff (Millwood)* 2011; 30 (6): 1192–9.
15. Ahrnsbrak R, Bose J, Hedden SL, Lipari RN, Park-Lee E, Tice P. *Key Substance Use and Mental Health Indicators in the United States: Results from the 2016 National Survey on Drug Use and Health*. Rockville, MD: Center for Behavioral Statistics and Quality, Substance Abuse and Mental Health Services Administration; 2017.
16. Birnbaum HG, White AG, Schiller M, Waldman T, Cleveland JM, Roland CL. Societal costs of prescription opioid abuse, dependence, and misuse in the United States. *Pain Med* 2011; 12: 657–67.
17. Florence CS, Zhou C, Luo F, Xu L. The economic burden of prescription opioid overdose, abuse, and dependence in the United States, 2013. *Med Care* 2016; 54 (10): 901–6.
18. Centers for Disease Control and Prevention. Prescription painkiller overdoses in the US. https://www.cdc.gov/vitalsigns/painkilleroverdoses/ Accessed September 09, 2019.
19. White House Office of the Press Secretary. FACT SHEET: Obama Administration announces public and private sector efforts to address prescription drug abuse and heroin use; 2015. https://obamawhitehouse.archives.gov/the-press-office/2015/10/21/fact-sheet-obama-administration-announces-public-and-private-sector Accessed September 09, 2019.
20. U.S. Executive Office of the President. Epidemic: responding to America's prescription drug abuse crisis. https://www.ncjrs.gov/App/Publications/abstract.aspx? ID=256103 Accessed September 09, 2019.
21. Manasco AT, Griggs C, Leeds R, *et al*. Characteristics of state prescription drug monitoring programs: a state-by-state survey. *Pharmacoepidemiol Drug Saf* 2016; 25 (7): 847–51.
22. National Institute on Drug Abuse, National Institutes of Health, Department of Health and Human Services. *PA-18-058: Prescription Drug Abuse (R01 Clinical Trial Optional). Prescription Drug Abuse*. North Bethesda, MD: National Institute on Drug Abuse.
23. Shutler L, Nelson LS, Portelli I, Blachford C, Perrone J. Drug use in the twittersphere: a qualitative contextual analysis of tweets about prescription drugs. *J Addict Dis* 2015; 34 (4): 303–10.
24. Sarker A, O'Connor K, Ginn R, *et al*. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug Saf* 2016; 39 (3): 231–40.
25. Hanson CL, Burton SH, Giraud-Carrier C, West JH, Barnes MD, Hansen B. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *J Med Internet Res* 2013; 15 (4): e62.
26. Jouanjus E, Mallaret M, Micallef J, Ponté C, Roussin A, Lapeyre-Mestre M. Comment on social media mining for toxicovigilance: monitoring prescription medication abuse from Twitter. *Drug Saf* 2017; 40 (2): 183.
27. Chary M, Genes N, McKenzie A, Manini AF. Leveraging social networks for toxicovigilance. *J Med Toxicol* 2013; 9 (2): 184–91.
28. PEW Research Center. *Demographics of Social Media Users and Adoption in the United States|Pew Research Center. Social Media Fact Sheet*. Washington, DC: PEW Research Center; 2017.
29. Felt M. Social media and the social sciences: how researchers employ big data analytics. *Big Data Soc* 2016; 3 (1): 205395171664582.
30. Cao B, Gupta S, Wang J, *et al*. Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. *J Med Internet Res* 2017; 19 (11): e394.
31. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. *Am J Public Health* 2017; 107 (1): e1–8.
32. Culotta A, Kumar Ravi N, Cutler J. Predicting the demographics of Twitter users from website traffic data. In: AAAI'15 *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*; 2015: 72–8.
33. Woods HC, Scott H. #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *J Adolesc* 2016; 51: 41–9.
34. Wong CA, Merchant RM, Moreno MA. Using social media to engage adolescents and young adults with their health. *Healthcare (Amsterdam, Netherlands)* 2014; 2 (4): 220–4.
35. Nguyen D, Gravel R, Trieschnigg D, Meder T. "How old do you think I am?": a study of language and age in Twitter. *In: Proceedings of the Sev-*

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

enth International AAAI Conference on Weblogs and Social Media; 2013: 439–48.
36. Sarker A, Ginn R, Nikfarjam A, *et al*. Utilizing social media data for pharmacovigilance: A review. *J Biomed Inform* 2015; 54: 202–12.
37. Demner-Fushman D, Elhadad N. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearb Med Inform* 2016; 1: 224–33.
38. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Informatics Assoc* 2006; 13 (1): 24–9.
39. Zielstorff RD. Controlled vocabularies for consumer health. *J Biomed Inform* 2003; 36 (4–5): 326–33.
40. He Z, Chen Z, Oh S, Hou J, Bian J. Enriching consumer health vocabulary through mining a social Q&A site: a similarity-based approach. *J Biomed Inform* 2017; 69: 75–85.
41. Sarker A, Gonzalez-Hernandez G. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *J Biomed Inform* 2018; 88: 98–107.
42. Yakushev A, Mityagin S. Social networks mining for analysis and modeling drugs usage. *Proc Comput Sci* 2014; 29: 2462–71.
43. Schiano F, Deluca P, Baldacchino A, *et al*. Drugs on the web; the Psychonaut 2002 EU project. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2006; 30 (4): 640–6
44. Mcnaughton EC, Black RA, Zulueta MG, Budman SH, Butler SF. Measuring online endorsement of prescription opioids abuse: An integrative methodology. *Pharmacoepidemiol Drug Saf* 2012; 21 (10): 1081–92.
45. Davey Z, Schiano F, Corazza O, Deluca P. e-Psychonauts: conducting research in online drug forum communities. *J Ment Heal* 2012; 21 (4): 386–94.
46. Daniulaityte R, Carlson R, Falck R, *et al*. I just wanted to tell you that loperamide WILL WORK": a web-based study of extra-medical use of loperamide. *Drug Alcohol Depend* 2013; 130 (1–3): 241–4.
47. Cameron D, Smith GA, Daniulaityte R, *et al*. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *J Biomed Inform* 2013; 46 (6): 985–97.
48. Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. *J Med Internet Res* 2013; 15 (9): e189.
49. McNaughton EC, Coplan PM, Black RA, Weber SE, Chilcoat HD, Butler SF. Monitoring of internet forums to evaluate reactions to the introduction of reformulated oxycontin to deter abuse. *J Med Internet Res* 2014; 16 (5): e119.
50. Mcnaughton EC, Black RA, Weber SE, Butler SF. Assessing abuse potential of new analgesic medications following market release: an evaluation of internet discussion of tapentadol abuse. *Pain Med* 2015; 16 (1): 131–40.
51. Maclean D, Gupta S, Lembke A, Manning C, Heer J. Forum77: an analysis of an online health forum dedicated to addiction recovery. In: *CSCW '15 Proc ACM Conference on Computer Supported Cooperative Work & Social Computing*; 2015.
52. Buntain C, Golbeck J. This is your Twitter on drugs. Any questions? In: *Proceedings of the 24th International Conference on World Wide Web-WWW'15 Companion*; 2015.
53. Katsuki T, Mackey TK, Cuomo R. Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of Twitter data. *J Med Internet Res* 2015; 17 (12): e280.
54. Chan B, Lopez A, Sarkar U. The Canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. *PLoS One* 2015; 10 (8): e0135072.
55. Seaman I, Giraud-Carrier C. Prevalence and attitudes about illicit and prescription drugs on Twitter. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI)*; 2016: 14–17.
56. Ding T, Roy A, Chen Z, Zhu Q, Pan S. Analyzing and retrieving illicit drug-related posts from social media. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2016: 1555–60. doi: 10.1109/BIBM.2016.7822752.
57. Jenhani F, Gouider MS, Said LB. A hybrid approach for drug abuse events extraction from Twitter. *Proc Comput Sci* 2016; 96: 1032–40.
58. Zhou Y, Sani N, Luo J. Fine-grained mining of illicit drug use patterns using social multimedia data from Instagram. In: *Proceedings–2016 IEEE International Conference on Big Data (Big Data 2016)*; 2016.
59. Anderson L, Bell HG, Gilbert M, *et al*. Using social listening data to monitor misuse and nonmedical use of bupropion: a content analysis. *JMIR Public Health Surveill* 2017; 3 (1): e6.
60. Kalyanam J, Katsuki T, R.G. Lanckriet G, Mackey TK. Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning. *Addict Behav* 2017; 65: 289–95.
61. Phan N, Bhole M, Ae Chun S, Geller J. Enabling real-time drug abuse detection in tweets. In: *Proceedings International Conference on Data Engineering*; 2017.
62. Yang X, Luo J. Tracking illicit drug dealing and abuse on Instagram using multimodal analysis. *ACM Trans Intell Syst Technol* 2017; 8 (4): 1–15.
63. Chary M, Genes N, Giraud-Carrier C, Hanson C, Nelson LS, Manini AF. Epidemiology from tweets: estimating misuse of prescription opioids in the USA from social media. *J Med Toxicol* 2017; 13 (4): 278–286.
64. D'Agostino AR, Optican AR, Sowles SJ, Krauss MJ, Escobar Lee K, Cavazos-Rehg PA. Social networking online to recover from opioid use disorder: a study of community interactions. *Drug Alcohol Depend* 2017; 181: 5–10.
65. Cherian R, Westbrook M, Ramo D, Sarkar U. Representations of codeine misuse on Instagram: content analysis. *J Med Internet Res* 2018; 4 (1): e22.
66. Graves RL, Tufts C, Meisel ZF, Polsky D, Ungar L, Merchant RM. Opioid discussion in the twittersphere. *Subst Use Misuse* 2018; 53 (13): 2132–9.
67. Hu H, Moturu P, Dharan K, *et al*. Deep learning model for classifying drug abuse risk behavior in tweets. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*; 2018: 386–7.
68. Chary M, Yi D, Manini AF. Candyflipping and Other combinations: identifying drug–drug combinations from an online forum. *Front Psychiatry* 2018; 9: 135.
69. Fan Y, Zhang Y, Ye Y, Li X, Zheng W. Social media for opioid addiction epidemiology: automatic detection of opioid addicts from Twitter and case studies. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management-CIKM '17*; New York, NY: ACM Press; 2017: 1259–67.
70. Bigeard E, Grabar N, Thiessard F. Detection and analysis of drug misuses. *Front Pharmacol* 2018; 9: 791.
71. Chen X, Faviez C, Schuck S, *et al*. Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate. *Front Pharmacol* 2018; 9: 541.
72. Pandrekar S, Chen X, Gopalkrishna G, *et al*. Social media based analysis of opioid epidemic using reddit. *AMIA Annu Symp Proc* 2018; 2018: 867–76.
73. Lossio-Ventura JA, Bian J. An inside look at the opioid crisis over Twitter. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); IEEE; 2018: 1496–9.
74. Hu H, Phan N, Geller J, *et al*. Deep self-taught learning for detecting drug abuse risk behavior in tweets. In: *CSoNet 2018: Computational Data and Social Networks*. Cham, Switzerland: Springer; 2018: 330–42.
75. Adams N, Artigiani EE, Wish ED. Choosing your platform for social media drug research and improving your keyword filter list. *J Drug Issues* 2019; 49 (3): 477–92.
76. Lu J, Sridhar S, Pandey R, Hasan MA, Mohler G. Redditors in recovery: text mining reddit to investigate transitions into drug addiction. In: *2018 IEEE International Conference on Big Data*. Seattle, WA: IEEE; 2018: 2521–30.
77. Tibebu S, Chang VC, Drouin C-A, Thompson W, Do MT. At-a-glance-what can social media tell us about the opioid crisis in Canada? *Health Promot Chronic Dis Prev Can* 2018; 38 (6): 263–7.

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024

78. Chancellor S, Nitzburg G, Hu A, Zampieri F, De Choudhury M. Discovering alternative treatments for opioid use recovery using social media. In: *Proceedings* of the 2019 CHI Conference on Human Factors in Computing Systems-CHI '19. New York, NY: ACM Press; 2019: 1–15.
79. Shutler L, Nelson LS, Portelli I, Blachford C, Perrone J. Drug use in the twittersphere: a qualitative contextual analysis of tweets about prescription drugs. *J Addict Dis* 2015; 34 (4): 303–10.
80. Huang X, Smith MC, Jamison AM, *et al*. Can online self-reports assist in real-time identification of influenza vaccination uptake? A cross-sectional study of influenza vaccine-related tweets in the USA. *BMJ Open* 2019; 9: e024018.
81. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *J Biomed Inform* 2017; 69: 1–19.
82. Sarker A, Belousov M, Friedrichs J, *et al*. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the social media mining for health (SMM4H)-2017 shared task. *J Am Med Inform Assoc* 2018; 25 (10): 1274–83.
83. Tamersoy A, De Choudhury M, Chau DH. Characterizing smoking and drinking abstinence from social media. *HT ACM Conf Hypertext Soc Media* 2015; 2015: 139–48.
84. Salimian PK, Chunara R, Weitzman ER. Averting the perfect storm: addressing youth substance use risk from social media use. *Pediatr Ann* 2014; 43 (10): 411.
85. Cavazos-Rehg PA, Krauss MJ, Sowles SJ, Bierut LJ. Marijuana-related posts on Instagram. *Prev Sci* 2016; 17 (6): 710–20.
86. Dai H, Hao J. Mining social media data on marijuana use for post traumatic stress disorder. *Comput Hum Behav* 2017; 70: 282–90.

Downloaded from https://academic.oup.com/jamia/article/27/2/315/5581276 by guest on 19 April 2024