Addictive Behaviors 65 (2017) 289–295

Contents lists available at ScienceDirect

Addictive Behaviors

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

Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning

Janani Kalyanam [a], Takeo Katsuki [b], Gert R.G. Lanckriet [a], Tim K. Mackey [c,d,e,*]

[a] Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA, USA
[b] Kavli Institute for Brain and Mind, University of California, San Diego, San Diego, CA, USA
[c] Department of Anesthesiology, University of California, San Diego, School of Medicine, San Diego, CA, USA
[d] Division of Global Public Health, University of California, San Diego, School of Medicine, Department of Medicine, San Diego, CA, USA
[e] Global Health Policy Institute, San Diego, CA, USA

HIGHLIGHTS

- A methodology using unsupervised machine learning analyzed 11 million tweets filtered for commonly abused prescription opioid drugs
- Analyses identified 2.3 million tweets with content relevant to nonmedical use of prescription medications/drugs (NMUPD)
- Twitter content was associated with a high degree of discussion (approximately 80%) about polydrug abuse involving multiple types of substances
- The methodology can filter large volumes of twitter data with minimal human intervention to identify macro NMUPD themes and trends

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form 29 July 2016
Accepted 13 August 2016
Available online 17 August 2016

Keywords:
Nonmedical use of prescription drugs
Prescription opioid abuse
Digital surveillance
Substance abuse
Twitter
Social media

ABSTRACT

Introduction: Nonmedical use of prescription medications/drugs (NMUPD) is a serious public health threat, particularly in relation to the prescription opioid analgesics abuse epidemic. While attention to this problem has been growing, there remains an urgent need to develop novel strategies in the field of "digital epidemiology" to better identify, analyze and understand trends in NMUPD behavior.
Methods: We conducted surveillance of the popular microblogging site Twitter by collecting 11 million tweets filtered for three commonly abused prescription opioid analgesic drugs Percocet® (acetaminophen/oxycodone), OxyContin® (oxycodone), and Oxycodone. Unsupervised machine learning was applied on the subset of tweets for each analgesic drug to discover underlying latent themes regarding risk behavior. A two-step process of obtaining themes, and filtering out unwanted tweets was carried out in three subsequent rounds of machine learning.
Results: Using this methodology, 2.3M tweets were identified that contained content relevant to analgesic NMUPD. The underlying themes were identified for each drug and the most representative tweets of each theme were annotated for NMUPD behavioral risk factors. The primary themes identified evidence high levels of social media discussion about polydrug abuse on Twitter. This included specific mention of various polydrug combinations including use of other classes of prescription drugs, and illicit drug abuse.
Conclusions: This study presents a methodology to filter Twitter content for NMUPD behavior, while also identifying underlying themes with minimal human intervention. Results from the study track accurately with the inclusion/exclusion criteria used to isolate NMUPD-related risk behaviors of interest and also provides insight on NMUPD behavior that has a high level of social media engagement. Results suggest that this could be a viable methodology for use in big data substance abuse surveillance, data collection, and analysis in comparison to other studies that rely upon content analysis and human coding schemes.

1. Introduction

As recently highlighted by President Obama's widely publicized participation at the National Rx Drug Abuse Summit, nonmedical use of prescription medications/drugs (NMUPD), particularly prescription analgesic opioid abuse, is a grave threat to the nation's health ("President

* Corresponding author at: Global Health Policy Institute, 6256 Greenwich Drive, Mail Code: 0172X, San Diego, CA 92122, USA.
E-mail address: tmackey@ucsd.edu (T.K. Mackey).

http://dx.doi.org/10.1016/j.addbeh.2016.08.019
0306-4603/© 2016 Elsevier Ltd. All rights reserved.

Obama Is Taking More Steps to Address the Prescription Drug Abuse and Heroin Epidemic," 2016). In fact, the U.S. Centers for Disease Control and Prevention recorded a record number of drug overdose deaths in 2014, headlined by a nearly fourfold increase in prescription opioid-related drug mortality since 1999, further coupled by increased heroin injection drug use associated with NMUPD behavior (Centers for Disease Control and Prevention (CDC), 2011, 2013; Longo, Compton, Jones, & Baldwin, 2016; Rudd, Aleshire, Zibbell, & Gladden, 2016). As this public health crisis continues to gain public attention, so does the need for better data identifying underlining NMUPD behaviors, trends, and risk factors in order to optimize efforts at improving access to treatment, preventing overdose, and ensuring community interventions are effective. Currently, existing NMUPD data is largely derived from national population-based surveys that measure the prevalence estimates, attitudes, and associated trends of various forms of substance abuse (including NMUPD) and rely upon respondents to self-report their past drug use behaviors via face-to-face interviews or self-administered questionnaires (Katsuki, Mackey, & Cuomo, 2015; Schepis & McCabe, 2016). These survey-based instruments are critical in identifying generalizable trends of prescription drug abuse behavior in a national population, assessing changes in what classes of prescription drugs are becoming popular targets of abuse, and aid in the development of targeted interventions and policy to address risk and protective factors common to NMUPD (Han, Compton, Jones, & Cai, 2015).

However, even powerful nationally representative surveys, such as the National Survey on Drug Use and Health and the Monitoring the Future survey (which focuses on students and young adults), have certain inherent limitations (McCabe, West, & Boyd, 2013; McCabe, West, Teter, & Boyd, 2012; Schepis & McCabe, 2016). Most evidently, they largely rely on respondents to self-report and recall recent and past drug abuse behavior, a methodology that can be subject to recall bias (Harrison & Hughes, 1997). Further, results from these surveys generally take time to compile after data collection is completed, with trends reported from these observations possibly changing by the time survey results are reported (Katsuki et al., 2015).

Hence, alternative methods for conducting surveillance of NMUPD behavior are needed to augment findings from national substance abuse surveys, including leveraging the power of "big data" analysis and social media platforms that are now heavily populated by a wide demographic of the substance using population, a practice now popularized as "digital epidemiology" or "infoveillance (Salathé et al., 2012)."

Despite, growing opportunities in a growing digitized social sphere, the massive size of these datasets and accompanying challenges of filtering, processing and analyzing the data in a meaningful way, has left the field ripe for innovation and improvement, particularly through cross-disciplinary research collaborations. In response, this study advances prior studies that have examined the linkages between Twitter and NMUPD and introduces a new methodology leveraging recent developments in computer science in order to gain a "bigger" picture of national NMUPD trends. Previous studies have used methodologies focusing on content analysis and human coding/annotation using keyword searches, identifying subsets of Twitter NMUPD-related social circles, and analyzing a random sample of filtered tweets (Hanson, Burton et al., 2013; Hanson, Cannon, Burton, & Giraud-Carrier, 2013; Katsuki et al., 2015; Shutler, Nelson, Portelli, Blachford, & Perrone, 2015).

In this study, we similarly conducted surveillance of the popular microblogging site Twitter (which now commands 310 million active users) filtered for content posted by users that specifically mentioned prescription analgesic opioid drugs. However, because of the massive amount of data required to be analyzed and given that content on Twitter is not curated for information specifically relevant to NMUPD, Twitter datasets often contain a high number of tweets that are non-relevant to NMUPD (i.e. noise) compared to content that actually describes NMUPD-related behavior (i.e. signal content). This condition necessitates a methodology that can iteratively filter tweets to eliminate noise and only retain tweets relevant to NMUPD behavior similar to those used in other studies examining other important public health issues (Chen, Hossain, Butler, & Ramakrishnan, 2016; Prier, Smith, Giraud-Carrier & Hanson, 2011.) Concomitantly, by filtering only for NMUPD relevant content, this methodology subsequently analyzes the tweets to discover and identify the different underlying latent themes that exist in the entire dataset. Hence, our study's methodology differs from prior studies because it can be used to highly automate filtering and coding of an extremely large dataset of tweets and simultaneously identify key NMUPD themes that are occurring in the entire Twittersphere in order to better inform researchers and the public about changes in the prescription opioid epidemic.

2. Methods

2.1. Overall aims

This study was conducted in two distinct phases: data collection and data analysis. The goal of this study included two distinct aims related to data processing and analysis to identify risk behaviors associated with NMUPD as reported in content generated by Twitter users. The first aim was to increase the signal to noise ratio in the dataset of all tweets analyzed and weed out tweets that are not relevant to NMUPD behavior. The second involved identifying themes and patterns from the large corpus of Tweets in order to gain a broader understanding of NMUPD behaviors for a larger Twitter user population that included more than 11 million tweets collected during a six-month period. In order to handle a dataset of this magnitude it is necessary to develop methodologies to be as automated as possible and limit the amount of human coding, in order to scale big data collection, surveillance and analysis. Section 2.2 describes the data collection process used to generate large amounts of tweets on NMUPD from the Twitter Application Programming Interface (API) stream. Section 2.3 describes the iterative methodology that was employed to increase the signal to noise ratio in the dataset, and to identify patterns and themes present in the data specific to analgesic opioid NMUPD.

2.2. Data collection

Twitter provides a public API that enables the collection of messages publicly posted by its users via its online platform. We used a data collection methodology involving cloud-based computing services offered by Amazon Web Services (AWS) and virtual computers via Amazon EC2 t2.micro instances set to filter and collect tweet objects containing specific NMUPD keywords.

Keywords included brand and international non-proprietary (e.g. generic) names (INN names) of commonly abused prescription analgesic opioid drugs. The data collection methodology used for this study has been previously described in detail in a prior published study (Katsuki et al., 2015). INN names of prescription opioid drugs included: Percocet® (acetaminophen/oxycodone) OxyContin® (oxycodone) and Oxycodone and were used in conjunction with the streaming API in order to track tweet objects that potentially contain at least one of these keywords. This generated a total of approximately 11M English-language tweets that were collected between the period of June and November 2015. A summary of the number of tweets collected for each drug is provided in Table 1. Additionally, during the preprocessing of the data, any identifiable information (e.g. Twitter account names, etc.) was removed from prior to data analysis.

2.3. Analysis plan

In order to appropriately code, identify, and characterize large volumes of data collected from social media sites in the context of public health issues, the application of machine learning has become a critical strategy in digital surveillance. The need for the application of machine

Table 1
Summary of keywords used in conjunction with the Twitter Streaming API for data collection and the corresponding number of tweets collected.

Keywords	#-of-tweets	% of English tweets
Percocet®	8,004,229	94.24
OxyContin®	3,280,910	92.16
Oxycodone	2,315,321	88.90

Table 2
Summary of rules applied to determine if a theme was relevant for inclusion in subsequent iterations.

Condition 1: Contains INN or slang names identified by NIDA of a prescription drug subject to abuse.
Condition 2: Mentions of other illicit drugs (e.g., heroin, cocaine, marijuana) and/or alcohol.
Condition 3: Mentions of identified substance abuse risk behavior (e.g. overdose, injection, withdrawal).
Condition 4: Contains adjectives related to prescription drug abuse behavior (e.g., popping).

learning has been necessitated by the sheer enormity of the data used for analysis when generating data from public sources like the Twitter API. Machine learning, especially unsupervised machine learning models are efficient at automatically finding patterns in the data and summarizing the content of the text corpus. The algorithms are computationally efficient and results have the advantage of not being affected by potential human judgment bias.

In the machine learning literature, there exist models which, when given a text corpus, are able to identify the underlying latent patterns or themes or (more formally in the machine learning literature) topics present in the corpus. Such models are referred to as topic models owing to their ability to summarize the content of the corpus concisely in terms of a few distinct topics (Blei, Ng, & Jordan, 2003). With the advent of Twitter and other microblogging sites as a potent source of textual data, a model called the Biterm Topic Model (BTM) was proposed specifically to detect themes and patterns in corpora of short-texts (Yan, Guo, Lan, & Cheng, 2013).

The methodology for our work is built by using BTM as the core for recognizing patterns in our corpus of tweets. BTM, when given a corpus of tweets as a parameter, k, for the number of themes to be identified as inputs identifies k underlying themes in the corpus. This is the *learning phase* of BTM. As the output of this phase, it produces a discrete probability distribution for all words for each theme. Ideally, this distribution would place large weights on the words that are most representative of that theme. Hence, a ranking of the top-10 words produced by BTM for each theme can be used as a summary representing all the themes present in the input corpus.

BTM also operates in what is called the *inference phase*. In this phase, BTM has the ability to decompose each tweet in terms of the themes identified in the *learning phase*. As the output of this phase, BTM produces a histogram for each tweet where each bin represents a theme discovered in the *learning phase*. The value in that bin represents how correlated the given tweet is to that particular theme. A tweet highly correlated to a particular theme will have, in its histogram representation, a large weight placed in the corresponding bin. The *inference phase* of BTM essentially enables us to retrieve the tweets most correlated to a particular theme.

First, the dataset was separated according to the keywords filtered in Table 1. Once a subset of tweets corresponding to each drug was obtained, a two-step iterative process was employed. The first step involved detecting k themes from the set of tweets corresponding to each drug using the *learning phase* of BTM. The second step involved manually identifying themes that produced false positive content; i.e., content irrelevant to NMUPD behavior and promotion. For this step, the top ranked words in each theme (produced in the first step) were analyzed in conjunction with the inclusion/exclusion rules laid out in Table 2 to determine whether or not a theme was relevant to NMUPD behavior.

Prior studies that have utilized content analysis and human annotation in order to filter out tweets that are not explicitly related to individual NMUPD behavior (e.g. tweets containing news reports, automated feeds, tweets by commercial entities, illicit online sales, etc.) (Hanson, Burton et al., 2013; Hanson, Cannon et al., 2013; Katsuki et al., 2015; Shutler et al., 2015). We used a similar filtering scheme instead relying upon our machine learning BTM protocol to identify tweets that contained NMUPD-related behavior content of interest. Specifically, relevance to NMUPD behavior focused on four specific conditions: (1) contained INN or "street"/slang term for a prescription opioid drug subject to abuse; (2) mentioned other prescription and/or illicit drug abuse (i.e. polydrug abuse); (3) mentioned identified substance abuse risk behavior (e.g. overdose, injection, adverse event); and (4) contained a commonly used adjective or verb related to NMUPD. Once the themes satisfying conditions laid out in Table 2 were identified, using the *inference phase* of BTM, the most correlated tweets to these themes were obtained and those not related were discarded.

These two steps were repeated in three iterative BTM rounds to improve content saturation and ensure that the filtered content was highly relevant to NMUPD behavior of interest. After each iteration, what remains is a smaller corpus of tweets with a higher signal to noise ratio than before. The number of themes to be discovered, i.e. the parameter k, needs to be set by the user. We set this number to 20, 10, and 7 in the

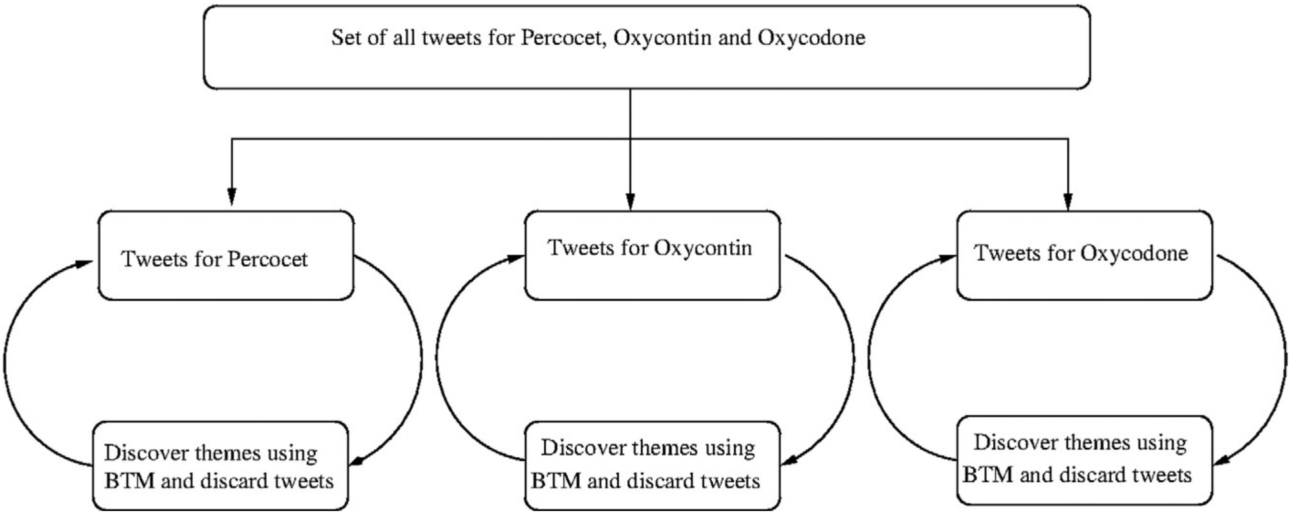
Set of all tweets for Percocet, Oxycontin and Oxycodone
Tweets for Percocet
Tweets for Oxycontin
Tweets for Oxycodone
Discover themes using BTM and discard tweets
Discover themes using BTM and discard tweets
Discover themes using BTM and discard tweets

Fig. 1. An overview of the steps undertaken in the interative BTM rounds applied to set of all tweets.

Table 3
This table illustrates some of examples of themes discovered for each drug, along with the filtering decision made during the first round of iteration.

	Theme example 1	Theme example 2	Theme example 3
Theme summary for Percocet® (through top words)	Percocet, super, high, best, buy, online, place, offer compare, quality	Percocet, xanax, pop, strippers	Percocet, liquor, pour, dose, money, weed
Filtering decision	Exclude (example of illicit online sale of controlled substances)	Include	Include
Theme summary for OxyContin® (through top words)	Oxycontin, bottle, cocaine, drug, love, wrong	Oxycontin, addiction, dangerous, abuse	Oxycontin, richest, Forbes, list, family, newcomer
Filtering decision	Include	Include	Exclude (example of news/media content)
Theme summary for Oxycodone (through top words)	Oxycodone, drug, approval, fda, media, reports	Heroin, oxycodone, cocaine, appearance, terrifying, change	Canada, monopoly, rules, oxycodone, drugs
Filtering decision	Exclude (example of news/media content)	Include	Exclude (example of news/media content)

first, second and third round of iterations respectively in order to ensure appropriate thematic saturation balancing missing important emerging themes that could be detected. Fig. 1 summarizes our iterative data filtering methodology.

Before the BTM model was applied on the full corpus of tweets as described above, the tweets were subjected to several standard data preprocessing steps. The first step was to produce a subset of tweets corresponding to each drug INN. Subsequently, the lang field and the user_lang field provided by the Twitter API were used to remove any non-English tweets. From the remaining tweets, all the stopwords were removed. For each drug category, a list of vocabulary and the corresponding counts were built. Word tokens that occurred less than 10 times in the corpus of that particular drug were removed from all tweets to prevent the model from fitting to what could be noisy outliers. Only alphanumeric strings were retained, any string with special characters was discarded. In addition, any tweet with two or less words was also discarded given limited interpretability.

3. Results

Table 3 illustrates some examples of the themes produced by our BTM machine learning protocol for the three prescription opioid analgesic drugs during the first round of the iteration, along with the filtering decision that was made as to whether or not to retain the tweets pertaining to this theme in the subsequent rounds.

The themes listed in Table 3 were annotated manually according to the inclusion/exclusion rules laid out in Table 2. For Percocet®, the first example contains keywords like “buy”, “online”, “offer”, “quality”, “compare” etc. This suggests that this topic could be highly correlated with tweets promoting illicit prescription drug sales through illegal online pharmacies, also a recognized public health threat (Forman, 2003; Forman & Block, 2006; Forman, Woody, McLellan, & Lynch, 2006; Mackey, Liang, & Strathdee, 2013; Raine, Webb, & Maxwell, 2009). Hence, this theme does not satisfy the rules of inclusion from Table 2, though warrants further examination, which is currently being undertaken in a separate study. In order to validate the application of these inclusion and exclusion rules, a list of 1000 tweets most correlated to this theme was retrieved and analyzed by a human coder (first author with training from last author). It was observed that 82% of all the tweets were about sales of prescription drugs through online pharmacies. This confirms that the majority of the tweets most correlated with this theme indeed do not satisfy the requirements for inclusion.

For each of the themes marked as “Exclude” in the OxyContin® and Oxycodone category, it appears that some of the identified themes are related to news media reports but not individual NMUPD user behavior. To evaluate this, the top 1000 most correlated tweets from each of the themes were analyzed. 25%–40% of these tweets were retweets of news headlines. Another 40% of the tweets were repetitions of the same news headlines, even though they were not retweets. In fact, in this set of 1000 tweets, there were only a handful of unique tweets (ranging between 4–25), most of which were news items. Importantly, by eliminating these tweets from subsequent BTM rounds, content that is not useful in inferring specific NMUPD behavior from individual users can be filtered out in the iterative machine learning process applied to a large set of data. These analyses also suggest that the top words discovered by BTM are indicative of whether or not the theme, and the tweets correlated to it, need to be included in the subsequent iterations of filtering.

In Table 4, the percentage of tweets retained between the first and the second rounds based on our inclusion and exclusion criteria was 24%–36% (for all three drugs). The percentage of tweets retained between the second and the third rounds is 72%–84%. This increase in the percentage of tweets that satisfy the inclusion criteria indicates that better NMUPD content saturation is achieved with each round of iteration.

Table 5 illustrates the top words from some of the topics obtained after the final round of data pruning. All the themes satisfy rules for inclusion that were prespecified in Table 2 suggesting that as per the rules, we might have attained saturation. Also, provided in Table 6 are some specific examples of tweets randomly sampled from the data after the final round of pruning. It is clear that all the tweet examples are related to at least one of the themes from Table 5. In addition, the content of the tweets itself is indicative of NMUPD behavior and abuse.

In Table 5, almost all of the themes mention more than one prescription drug, and in some cases mention the use of other illicit drugs (e.g. heroin, ecstasy). This suggests that Twitter prescription opioid analgesic abuse content and user behavior is highly associated with self-reporting of other forms of substance abuse, specific to certain classes of drugs. In particular, for the first theme under Percocet®, the top words suggest that abusing Percocet® and Xanax® is relaxing and addictive. While analyzing the top 100 tweets most correlated to this theme, 89% of the tweets were found to be pertinent to the proposed summary. In addition, in all three themes for Percocet, different polydrug combinations are mentioned with different accompanying adjectives, suggesting that each polydrug combination might exhibit its own unique form of user described behavior or effect. Examining specific examples of tweets identified as highly correlated to themes and reproduced in Table 6 further substantiates this pattern. For example, tweets for Percocet® primarily report use of other prescription drugs (examples #1 and #4 mention benzodiazepines though example #2 includes use of alcohol), the OxyContin® tweets describe various use both self-report

Table 4
This table summarizes the number of tweets used in the first round of iteration, and the % of tweets used in the subsequent rounds.

INN	#-tweets, 1st round	% of tweets retained for 2nd round (from the 1st round)	% of tweets retained for 3rd round (from the second round)
Percocet®	5,983,497	24%	84%
OxyContin®	2,812,364	36%	72%
Oxycodone	1,806,900	29%	74%

Table 5
This table illustrates some of examples of themes discovered for each drug after the final iteration of data pruning.

	Theme example 1	Theme example 2	Theme example 3
Theme summary for Percocet® (through top words)	Percocet, addict, taking, relax, xanax	Percocet, ecstasy, adderall, sleep	Percocet, vicodin, gum, ball, machine, withdrawal
Theme summary for OxyContin® (through top words)	Oxycontin, pain, addicted, pills	Oxycontin, dangerous, abuse, bottle, selling, history	Oxycontin, niggas, roxies, droppin, pistols
Theme summary for Oxycodone (through top words)	Oxycodone, ecstasy, pain, hugs, kisses, xanax	Oxycodone, heroin, morphine, addiction, make	Oxycodone, ecstasy, pain, xanax

and observational, and Oxycodone tweets also describe poly-use with other illicit drugs.

Another likely sign of user initiated and self-reported NMUPD behavior is the detection of street or slang terms associated with polydrug abuse combinations or drug abuse related behavior. This includes the term “hugs” and “kisses” in the first theme of Oxycodone, both words which used in combination are slang for the drug combination of ecstasy and oxycodone, which are also keywords included in the theme (Table 6, example #2 under Oxycodone). Similarly, the term “roxies” are included in the OxyContin® second theme, which is a slang term for Roxicodone®, another opioid analgesic (oxycodone hydrochloride). Table 6 also contains other controlled substances like Ritalin, Prednisone, Valium and Marijuana.

In order to assess the quality of themes that emerged after the final round of data pruning two types of evaluations were performed. The first was a *supervised* evaluation that involved manually annotating the tweets from each theme as being relevant or irrelevant to the theme detected. From each theme, a maximum of 2000 most correlated tweets were retrieved and annotated. The average false positive rate was calculated across all the themes for each drug. While manually annotating tweets for their relevance to NMUPD behavior, it was observed that the dataset contained several retweets. So, even if one tweet was found to be irrelevant, it was often the case that the tweet was retweeted several times; thereby increasing the false positive rate. In addition, there were also scenarios where even though a tweet contained keywords indicating polydrug abuse and/or potential adverse effects, the intent of the tweet remained vague. Such tweets were also marked as irrelevant during manual annotation. The results of the total number of tweets after the final round of machine learning, and the false positive rate for each drug is summarized in Table 7.

Table 6
Randomly sampled examples of tweets obtained from the data after the final round of pruning.

Example tweets for Percocet®:

1. popping percocet and xannies like they some tylenol
2. its only 3 pm and ive had a beer and 4 percocets your move bad decisions
3. just when i thought that id rock the mic again my brain was fucked up on percocet and vicodine
4. i got xanex percocet promethazine with codeine

Example tweets for OxyContin®:

1. i fell in love with a trap mami she be snortin cocaine and molly sometimes she be poppin oxycontin blue pill she be smokin them roxis
2. daydreams laced with oxycontin mind elsewhere
3. my mom is ritalin my dad is oxycontin
4. i need the zans and oxycontin christ every 2 hours

Example tweets for Oxycodone:

1. i sure wish i had a few beers and maybe an oxycodone to make this afterglow even more pleasurable
2. w00t oxycodone and morphine i feel like lindsay lohan
3. that moment when you realize the weeknds trademark xo stands for ecstacy and oxycodone hence xo til we overdose
4. high on coke and oxycodone marijuana is too weak4meh

The second was an *unsupervised* evaluation that involved the calculation of a metric called *cluster purity* (Bishop, 2006). This metric quantifies how coherent a theme is. If the tweets belonging to a theme are very diverse in terms of their content, that theme is considered incoherent, and the resulting purity score will be low; and vice versa. In order to obtain the cluster purity score for each theme, we considered the same 2000 most correlated tweets as before and calculated the average similarity between all pairs of tweets from this set.[1] This average similarity is the *cluster purity* of the theme. As baseline, we randomly sampled 2000 tweets from our original dataset of 11 M tweets, and calculated the average similarity between all pairs from this *random* set. The results are summarized in the last two columns of Table 7. The cluster purity for each drug is up to 3 times better than that of a random set of tweets. Both the *supervised* and *unsupervised* evaluations suggest that the themes obtained after the final round of data pruning are of good quality.

4. Conclusion

In summary, the results of this study suggest that the use of an automated methodology that employs iterative rounds of unassisted machine learning has the potential to filter and analyze large and complex social media conversational datasets with minimal human intervention such as through the use of manual content analysis or human annotation. Importantly, the study also demonstrates that a machine learning algorithm, such as the one employed in this study, can be used to identify themes in NMUPD use and behavior that are important in identifying macro and emerging trends in the context of broader prescription opioid analgesic abuse behavior.

Primarily, in this study, the central theme that emerged was that polydrug abuse is predominantly associated with Twitter prescription drug abuse discussions and could be indicative of larger behavioral trends of users abusing multiple prescription drugs and also combining use with other illicit substances. These results could form the basis for future in-depth studies examining the unique health and substance abuse consequences associated highly prevalent polydrug uses (including potentially linkages to mental health issues and behavior among young adults and adolescents already identified in the literature) (Fink et al., 2015; Kelly, Wells, Pawson, LeClair, & Parsons, 2014; Mackesy-Amiti, Donenberg, & Ouellet, 2015).

The study is also important within the context of future research attempting to effectively scale projects for big data analyses. Primarily, the methodology allows for the machine learning processes to “pre-filter” millions of Tweets, better isolating content that is highly relevant to a research question (such as prescription opioid analgesic abuse) based upon a researcher's own desired set of inclusion/exclusion criteria drawn from behavioral risk factors identified in the literature or through traditional survey instruments. Following this process of machine driven filtering, a smaller subset of tweets/content can be identified for more in-depth analysis and confirmation by human coders to better assess the accuracy of identified themes to highly correlated content. Overall, the methodology represents an innovative approach that has

[1] The histogram representation for each tweet was obtained from *inference phase* of BTM, and the *cosine similarity* was calculated between each pair of tweets.

Table 7
Summary of results of evaluating the quality of the themes obtained after the final round of data pruning.

Drug name	#-tweets	Average FP-rate	Average cluster purity across themes	Purity of a set of random tweets
Percocet	1,231,641	55%	0.4348	0.2378
Oxycontin	741,272	28%	0.5678	0.2219
Oxycodone	380,838	14%	0.6729	0.1976

the advantages of reliably identifying themes of interest from large amounts of data collected at a point of time from the Twitter public API.

4.1. Limitations

One of our primary aims in this study was to increase the signal to noise ratio in the tweets that contain the three NMUPD keywords so that the filtered dataset contains only those messages which are relevant to NMUPD usage and behavior. Certain aspects of our methodology have inherent limitations in achieving this goal. Firstly, we faced scenarios where the text of the tweet would satisfy all the inclusion criteria (hence the theme detected from this tweet, and similar ones might be marked for inclusion in subsequent iterative rounds of filtering), but the intent of the tweet remained vague (e.g. due to its brevity or lack of sufficient description). Hence, while human coders reviewing a sample of tweets would likely mark such tweets as being not relevant, such tweets (and the themes they belonged to), inadvertently increase the false positive rate. Several tweets contained hyperlinks, the content of which might have helped to better contextualize the original tweets; especially those with vague intents. However, this study primarily considered only the text of the tweets, and did not conduct further analysis into the content of hyperlinks, which in past studies have been identified as containing pictures, videos and other media confirming drug abuse behavior (Katsuki et al., 2015). The second aim was to identify themes about NMUPD usage and behavior in the Twittersphere. This study is more of an exploratory study that attempts to determine the major themes prevalent on Twittersphere when it comes to opioid analgesic NMUPD. However, in order to gain a full understanding, crowdsourced or large scale human coding of data augmented by a cohort of NMUPD users to corroborate findings would be the optimal methodology for future consideration. Lastly, a non-trivial amount of content on Twitter contains special characters that do not fall under the alphanumeric character categorization. Hence, by ommitting content that is not alphanumeric, we might inadvertently be discarding some useful information.

4.2. Future directions

The data spans a period of 6 months (June to November 2015). However, for the purposes of this study, the timestamps of the tweet were not taken into account during the learning process.

Future studies could attempt to identify changes in NMUPD behavioral trends and attitudes over time through a longitudinal study design. We also observed that a significant number of themes that are attributable to the possible sale of controlled substances by illicit online pharmacies. Future studies should specifically examine this potential pathway for illicit prescription drug access and its negative impact on NMUPD behavior, dependence, and addiction.

Role of funding sources

TM and TK received funding for the data collection phase of this project from the Alliance for Safe Online Pharmacies (SA15-092) (ASOP), a 501(c)(4) social welfare organization engaged on the issue of illicit online pharmacies, for this research and greatly acknowledge this support. JK and GL acknowledge support from the National Science Foundation (grants CCF-0830535 and IIS-1054960). GL acknowledges support from the Alfred P. Sloan Foundation.

Contributors

TM and TK conducted the data collection for the study. TM and JK designed the study and conducted the data analyses and wrote the initial draft of the manuscript. All authors contributed to the formulation, drafting, completion, and approval of the final manuscript.

Conflict of interest statement

TM is a non-compensated member of the ASOP academic advisory panel of ASOP and both TM and TK receiving funding for the data collection phase of this project from ASOP through a pilot research grant exploring prescription drug abuse risks online. Funder had no role or input in the study. There was no involvement of anyone other than the authors in the conception, design, collection, planning, conduct, analysis, interpretation, writing, and discussion to submit this work. Authors report no other financial relationships with any organizations that might have an interest in the submitted work.

References

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer (ISBN 978-0-387-31073-2).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Centers for Disease Control and Prevention (CDC) (2011). Vital signs: Overdoses of prescription opioid pain relievers—United States, 1999–2008. *MMWR. Morbidity and Mortality Weekly Report*, *60*(43), 1487–1492.

Centers for Disease Control and Prevention (CDC) (2013). Vital signs: Overdoses of prescription opioid pain relievers and other drugs among women—United States, 1999–2010. *MMWR. Morbidity and Mortality Weekly Report*, *62*(26), 537–542.

Chen, L., Hossain, K. S., Butler, P., & Ramakrishnan, N. (2016). Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. *Data Mining and Knowledge Discovery*, *30*(3), 681–710. http://dx.doi.org/10.1007/s10618-015-0434-x.

Fink, D. S., Hu, R., Cerdá, M., Keyes, K. M., Marshall, B. D. L., Galea, S., & Martins, S. S. (2015). Patterns of major depression and nonmedical use of prescription opioids in the United States. *Drug and Alcohol Dependence*, *153*, 258–264. http://dx.doi.org/10.1016/j.drugalcdep.2015.05.010.

Forman, R. F. (2003). Availability of opioids on the internet. *JAMA: The Journal of the American Medical Association*, *290*(7), 889. http://dx.doi.org/10.1001/jama.290.7.889.

Forman, R. F., & Block, L. G. (2006). The marketing of opioid medications without prescription over the Internet. *Journal of Public Policy & Marketing*, *25*, 133–146.

Forman, R. F., Woody, G. E., McLellan, T., & Lynch, K. G. (2006). The availability of web sites offering to sell opioid medications without prescriptions. *The American Journal of Psychiatry*, *163*(7), 1233–1238. http://dx.doi.org/10.1176/appi.ajp.163.7.1233.

Han, B., Compton, W. M., Jones, C. M., & Cai, R. (2015). Nonmedical prescription opioid use and use disorders among adults aged 18 through 64 years in the United States, 2003–2013. *JAMA: The Journal of the American Medical Association*, *314*(14), 1468–1478. http://dx.doi.org/10.1001/jama.2015.11859.

Hanson, C. L., Burton, S. H., Giraud-Carrier, C., West, J. H., Barnes, M. D., & Hansen, B. (2013a). Tweaking and tweeting: Exploring twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research*, *15*(4), e62. http://dx.doi.org/10.2196/jmir.2503.

Hanson, C. L., Cannon, B., Burton, S., & Giraud-Carrier, C. (2013b). An exploration of social circles and prescription drug abuse through twitter. *Journal of Medical Internet Research*, *15*(9), e189. http://dx.doi.org/10.2196/jmir.2741.

Harrison, L., & Hughes, A. (1997). Introduction—the validity of self-reported drug use: Improving the accuracy of survey estimates. *NIDA Research Monograph*, *167*, 1–16.

Katsuki, T., Mackey, T. K., & Cuomo, R. (2015). Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of twitter data. *Journal of Medical Internet Research*, *17*(12), e280. http://dx.doi.org/10.2196/jmir.5144.

Kelly, B. C., Wells, B. E., Pawson, M., LeClair, A., & Parsons, J. T. (2014). Combinations of prescription drug misuse and illicit drugs among young adults. *Addictive Behaviors*, *39*(5), 941–944.

Longo, D. L., Compton, W. M., Jones, C. M., & Baldwin, G. T. (2016). Relationship between nonmedical prescription-opioid use and heroin use. *New England Journal of Medicine*, *374*(2), 154–163. http://dx.doi.org/10.1056/NEJMra1508490.

Mackesy-Amiti, M. E., Donenberg, G. R., & Ouellet, L. J. (2015). Prescription opioid misuse and mental health among young injection drug users. *The American Journal of Drug and Alcohol Abuse*, *41*(1), 100–106. http://dx.doi.org/10.3109/00952990.2014.940424.

Mackey, T. K., Liang, B. A., & Strathdee, S. A. (2013). Digital social media, youth, and nonmedical use of prescription drugs: The need for reform. *Journal of Medical Internet Research*, *15*(7), e143. http://dx.doi.org/10.2196/jmir.2464.

McCabe, S. E., West, B. T., Teter, C. J., & Boyd, C. J. (2012). Co-ingestion of prescription opioids and other drugs among high school seniors: Results from a national study. *Drug and Alcohol Dependence*, *126*(1–2), 65–70. http://dx.doi.org/10.1016/j.drugalcdep.2012.04.017.

McCabe, S. E., West, B. T., & Boyd, C. J. (2013). Leftover prescription opioids and nonmedical use among high school seniors: A multi-cohort national study. *Journal of Adolescent Health*, *52*(4), 480–485. http://dx.doi.org/10.1016/j.jadohealth.2012.08.007.

President Obama Is Taking More Steps to Address the Prescription Drug Abuse and Heroin Epidemic (2016, March 29i). *President Obama is taking more steps to address the prescription drug abuse and heroin epidemic*. Gov: Whitehouse Retrieved May 31, 2016, from https://www.whitehouse.gov/blog/2016/03/29/president-obama-taking-more-action-address-prescription-drug-abuse-epidemic-0

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health related topics on twitter. An exploration of tobacco-related tweets as a test topic.

Lecture Notes in Computer Science, *6589*, 18–25. http://dx.doi.org/10.1007/978-3-642-19656-0_4.

Raine, C., Webb, D. J., & Maxwell, S. R. J. (2009). The availability of prescription-only analgesics purchased from the Internet in the UK. *British Journal of Clinical Pharmacology*, *67*(2), 250–254. http://dx.doi.org/10.1111/j.1365-2125.2008.03343.x.

Rudd, R. A., Aleshire, N., Zibbell, J. E., & Gladden, R. M. (2016). Increases in drug and opioid overdose deaths—United States, 2000–2014. *MMWR. Morbidity and Mortality Weekly Report*, *64*(50–51), 1378–1382. http://dx.doi.org/10.15585/mmwr.mm6450a3.

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., et al. (2012). Digital epidemiology. *PLoS Computational Biology*, *8*(7), e1002616. http://dx.doi.org/10.1371/journal.pcbi.1002616.

Schepis, T. S., & McCabe, S. E. (2016). Trends in older adult nonmedical prescription drug use prevalence: Results from the 2002–2003 and 2012–2013 national survey on drug use and health. *Addictive Behaviors*, *60*, 219–222. http://dx.doi.org/10.1016/j.addbeh.2016.04.020.

Shutler, L., Nelson, L. S., Portelli, I., Blachford, C., & Perrone, J. (2015). Drug use in the twittersphere: A qualitative contextual analysis of tweets about prescription drugs. PubMed–NCBI. *Journal of Addictive Diseases*, *34*(4), 303–310. http://dx.doi.org/10.1080/10550887.2015.1074505.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *The 22nd international conference* (pp. 1445–1456). http://dx.doi.org/10.1145/2488388.2488514 (New York, New York, USA: ACM).