DrugTracker: A Community-focused Drug Abuse Monitoring and Supporting System using Social Media and Geospatial Data (Demo Paper)

Han Hu, NhatHai Phan*, Xinyue Ye
{hh255,phan,xinyue.ye}@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Ruoming Jin, Kele Ding
{rjin1,kding}@kent.edu
Kent State University
Kent, Ohio, USA

Dejing Dou
dou@cs.uoregon.edu
University of Oregon
Eugene, Oregon, USA

Huy T. Vo
huy.vo@nyu.edu
City University of New York
New York City, New York, USA

ABSTRACT

In this paper, we present a community-focused drug abuse monitoring and supporting system, called **DrugTracker**, that utilizes social media and geospatial data in near real-time. Through the system, users can: **(1)** Detect drug abuse risk behaviors from social media platforms, e.g., Twitter; **(2)** Analyze drug abuse risk behaviors by querying consolidated and live datasets with keywords, spatial entities, and time constraints; and **(3)** Explore the query results and associated data through a web-based user interface in thematic choropleth, heatmap, and statistical charts. To protect the privacy of the Twitter users, whose data is collected, the system automatically hides the re-identification elements in tweets and aggregates the geo-tags into areas such as census tracts. For the demonstration purpose, our DrugTracker system is populated with a database that contains about 10 million tweets from the year 2017, that were annotated as drug abuse risk behavior positive by our deep learning model.

CCS CONCEPTS

• **Information Systems → Data Mining**; • **Collaborative and social computing → Social networks**.

KEYWORDS

drug abuse, deep learning, visualization, social media

ACM Reference Format:
Han Hu, NhatHai Phan*, Xinyue Ye, Ruoming Jin, Kele Ding, Dejing Dou, and Huy T. Vo. 2019. DrugTracker: A Community-focused Drug Abuse Monitoring and Supporting System using Social Media and Geospatial Data (Demo Paper). In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359076

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11…$15.00
https://doi.org/10.1145/3347146.3359076

1 INTRODUCTION

Even though drug abuse has reached epidemic proportions [17], there still lacks tools and means to prevent drug abuse epidemics effectively, especially for local communities and organizations, who are at the front and center of the fight. Several well-known resources have been developed for toxicovigilance monitoring include the National Poisoning Data System [12], the US Food and Drug Administration, the Drug Abuse Warning Network [1] and the MedWatch program [6]. These existing systems usually provide statistical data in the typical yearly fashion, which does not offer adequate information about drug abuse activities in a timely manner. This leads to difficulty in managing available resources and efforts (e.g., antidotes, recovery education, etc.), and to challenges in policy-making towards achieving best practices in prevention and recovery.

In addition, the prevalent usage of social network sites, mobile apps, forums, and the internet marketplace, has increasingly been recognized as a major factor in the spread of drug abuse epidemics [13]. Social media apps and the internet also make the purchase of illegal drugs more convenient; access to drugs can be just a few keystrokes away [16]. As both the exchange of information and the obtaining of drugs become easier and faster, the drug use trends become more volatile, diversified, and potentially lethal. Increasingly, one drug can result in damage, even loss of lives in a short time window (a few hours/days) [16]. It is an urgent demand to detect and monitor drug abuse activities on online social media.

Our Contributions. Towards addressing these problems, we develop a *community-focused drug abuse monitoring and supporting system*, called **DrugTracker**, using social media and geospatial data to provide local communities and organizations with the tools and capabilities to identify and understand the specific needs of drug misusers and/or abusers in near real-time. A such system will have crucial benefits to connect local communities and organizations with individuals and families, who are struggling with drug abuse, towards a better prevention and recovery outcome. In our system, well-trained deep learning models are integrated into the data collection process to detect tweets that contain drug abuse risk behaviors. Then end-users can operate the web-based interactive monitoring interface to browse the collected data in a spatial-temporal context in order to acquire insightful patterns

about drug abuse risk behaviors. *Our system is available on GitHub: https://github.com/hu7han73/DrugVis.*

2 RELATED WORK

Traditionally, the population of drug abusers and the drug usage trend are primarily estimated based on hospital emergency room (E/R) records and surveys. However, basing estimates on surveys and on E/R visits is increasingly becoming insufficient, because of those methods of estimating drug abusing populations and drug usage trends do not necessarily take into account the dynamics and possibly virtual nature of drug using communities, nor the fast pace with which drug abuse information cascades through virtual communities. In the past decade, researchers have utilized social media data and public forums, such as Twitter, Instagram, Reddit, and Bluelight, etc., to monitor drug abuse, with aim to potentially prevent drug abuse, and to reduce its harm [9, 18].

For instance, a number of studies have shown that social media, such as Twitter, can be excellent data sources for drug abuse and drug safety surveillance [4, 14]. The potential for exploring and reducing prescription drug abuse through social media is studied [15]. A more systematic study consists of a structured heterogeneous information network to model the users and posted tweets as well as their rich relationships constructed for automatic detection of opioid addictions [5].

From monitoring system perspectives, several well-known resources for toxicovigilance monitoring include the National Poisoning Data System [12], the US Food and Drug Administration, the Drug Abuse Warning Network [1] and the MedWatch program [6]. Myslin et al. [11] evaluated Twitter for information about cigarette smoking. Moreno et al. [10] mined Facebook regarding alcohol use. Bosley et al. [2] studied how Twitter users sought information about cardiac arrest. Recently, Coloma et al. [3] illustrated the potential of social media in drug safety surveillance. The NPDS provides data about calls to poison centers nationwide, and the information may be used to track the risks of prescription drug abuse. MedWatch focuses on providing information about adverse drug events. However, it does not monitor patterns of drug abuse in near-real time.

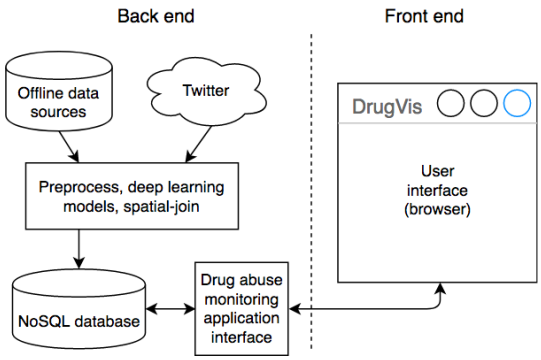


Figure 1: Basic architecture of DrugTracker system.

However, there is still a large gap between existing drug abuse surveillance methods and what is desired to help them connect and share the right information and resources with the population that needs them and with individuals in a timely manner. For instance, there is still a lack of approaches to utilize information networks that integrate online social media data with offline geospatial data for discovering and understanding drug abuse, due to a variety of challenges, such as data discrepancy, uncertainty, sparsity, noise, and bias. The spread of drug abuse behaviors, and the social structures and communities of drug abusers in online social media, are also largely unknown.

Table 1: Geo-tag types and frequency

Geo Type	Sub-type	Percentage
Place	Country	0.29%
	Admin	13.63%
	City	70.69%
	Neighborhood	0.32%
	POI	1.06%
Coordinates	-	14.01%

To meet the challenges that local communities and organizations face with regard to drug abuse, we develop a *community-focused online drug abuse monitoring and supporting system using social media and geospatial data.* Different from existing systems, our DrugTracker system aims to provide local communities and organizations with the tools to detect and analyze drug abuse risk behaviors in near real-time.

3 DRUGTRACKER SYSTEM

We select to implement DrugTracker as a web-based visualization system, since web-based systems are more flexible and requires virtually no setup process for end-users. Our system (Figure 1) includes two major parts: **(1)** The back-end, which runs on a server and provides data services, including data collection, data pre-processing, deep learning models for drug abuse risk behavior detection, and data management; and **(2)** The front-end, which runs on web browsers to provide interactive User Interfaces (UIs), for making queries and visualizing analysis results.

3.1 Back-end Services

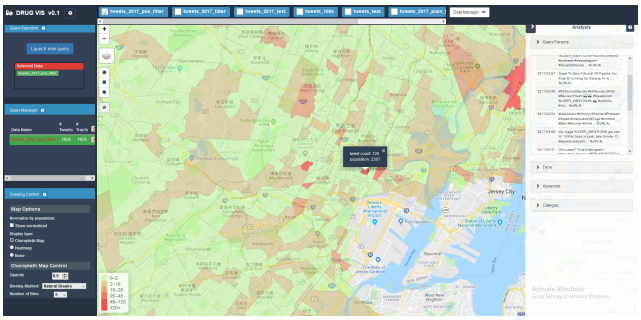
The back-end does the heavy lifting in the system, which runs a complete pipeline of collecting and processing data coming from social media and other sources (e.g., census data). There are three major modules in the back-end, including the data collection module, the data processing module, and the data management module.

For the data collection, we use tweets as a major source of geo-tagged social media data for its availability and abundance. We collect tweets through the publicly available Streaming API [19] using a typical keyword-based crawler well-integrated with trained deep learning models (i.e., CNN and LSTM models) designed to detect drug abuse risk behaviors in tweets [7, 8]. In our previous works [7, 8], we built our human labeled drug abuse risk behavior dataset, and demonstrated that a deep learning model, which was trained with both labeled data and large number of unlabeled data, can achieve state-of-art classification performance (86.63% of Accuracy, 89% of Recall, 86.83% of F1-value) on our dataset. The module is able to continuously collect newest tweets, to feed tweets to deep learning models, and to update the system with live data, so that the drug trend can be tracked and analyzed in near real-time.

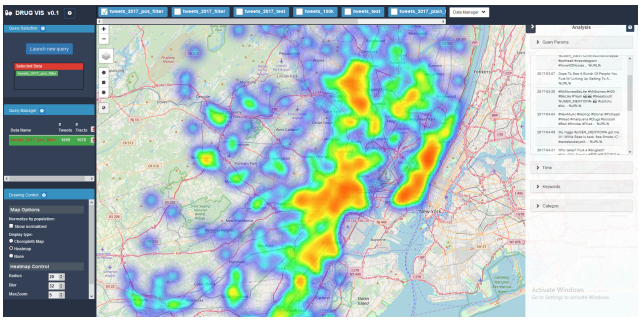
The use of geo-tags in tweets is not straight forward. Statistics of the 2017 dataset, as shown in Table 1, tell us that there are two major types of available geo-tags: '*place*' Object and '*coordinates*' Object. The '*coordinates*' Objects are GPS points that comes from Twitter users who have location service turned on, while each '*place*' Object refers to a named place entity that the tweet is associated with (but not necessarily originating from) [19]. However, the '*place*' Objects have several types but the resolution/granularity of each Object of same type may vary. In this demonstration, we opt to only use '*coordinates*' as geo-tags, as they provide the highest location precision with an adequate quantity of data points.

For offline data, such as geographical data and demographic data, we use publicly available data from trusted sources (e.g., Census Bureau). A work-flow is built to pre-join the tweets data with offline data as a pre-processing step for performance reason. In our system, we first collect Census Tract data, in the format of Shapefile, and population census data, in the form of spread-sheets. Then we joined (table-join and spatial-join) these data with tweets, so that each tweet record in our database is associated with offline data.

The pre-processed tweets data is stored and managed in a NoSQL database (MongoDB). For the fields in the original tweet objects, only those fields that are used by the front-end are stored. Several further pre-processing steps are preformed, including: **(1)** The time-stamp of each tweet, which is in UTC, is converted to local time using geo-location information; **(2)** Re-identification information in each tweet's texts, i.e., User Mention and URL, are removed; and **(3)** The sub-category of each tweet is extracted by identifying drug abuse-related keywords. Indexes are created for the fields of time-stamp, text, keywords, and geo-location to support fast text queries and spatial queries. Python and PHP scripts are served as the interface between the back-end and the front-end that execute queries and generate responses to the front-end.



(a) Choropleth



(b) Heatmap

Figure 2: Mapping Area: (a) Choropleth and (b) Heatmap.

3.2 Front-end Interactive Visualization

The front-end is built based on the open-sourced NeighborVis System [20]. The basic UI layout and some components are inherited, while new functions are added to incorporate the different functions offered in our system. The front-end is a dynamic web page constructed with HTML and Javascript. Javascript framework Leaflet is used for the core mapping functionality. Other frameworks, such as Heatmap-js and D3 are used for drawing heatmap and charts, respectively.

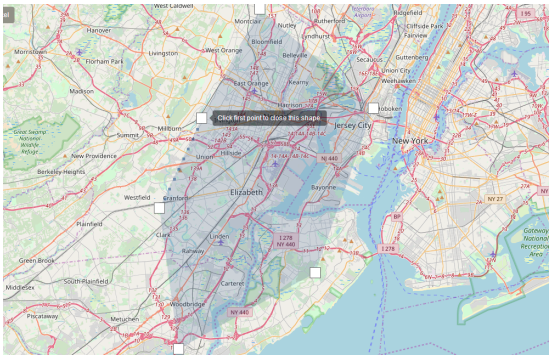
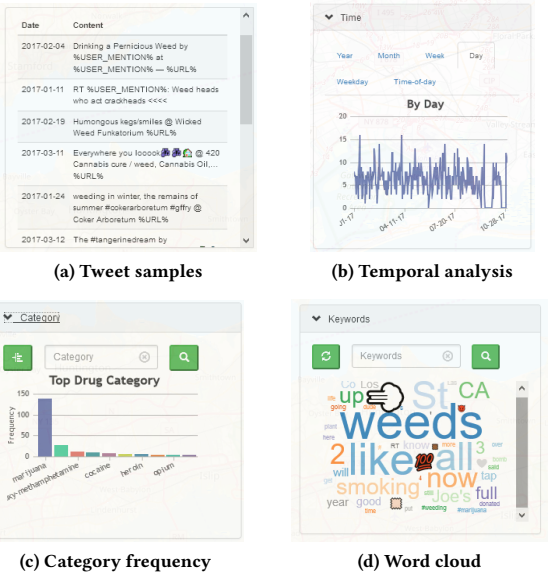


Figure 3: An example of polygon-shaped query area.

One noticeable change we made to the system is that the tweets displayed to the users in the front-end are always aggregated into some desired form instead of individually. This is done for two reasons: **(1)** To protect the sensitive information that the tweets and the classification results contain; and **(2)** To enable the displaying and analysis of tweets at population scale, as it is impractical to show millions of tweets on the map. The aggregation can be done on different types of administrative regions or entities, depending on the users' needs. For demonstration, we aggregate our data into Census Tract level. The UI of our system is demonstrated in Figures 2-4. The visualization layout is divided into three sections from left to right: **(1)** Query Management; **(2)** Mapping Area; and **(3)** Statistical Information.

Query Management. To begin, users should provide a query by clicking a button on the left panel to launch a query interface (Figure 2). A query has two mandatory constraints: temporal constraint and spatial constraint, and one optional constraint: content constraint. The temporal constraint limits the local posting-time of tweets and is specified by a start and an end date. The spatial constraint limits the location where the tweets are from and can be either a user defined area (optional shapes are circle, rectangle, and polygon Figure 3), or a list of states. The actual query area has a minimum granularity (Census Tract level in demonstration) to prevent the disclosure of individual tweet's location. The content constraint can be a list of keywords and phrases, e.g., "get high," "smoke blunt," etc. Users can also query multiple datasets at once to improve efficiency.

Once the query is submitted, the back-end will process the query and send aggregated results to front-end for displaying in the mapping area. Our system provides two basic mapping options: Choropleth and Heatmap (Figure 2). Choropleth displays the number of tweets that match the query within each area with a color mapping from green (low) to red (high). Jenks natural breaks is the default classification (binning) method. The choropleth can also display



(a) Tweet samples (b) Temporal analysis

(c) Category frequency (d) Word cloud

Figure 4: Analysis functions: (a) tweet samples, (b) Temporal analysis, (c) Category frequency, and (d) Word cloud.

the tweets data conjugated with selected offline data (e.g., normalize the number of tweets by the population in each Census Tract). Heatmap provides a way to view the data in a more analog way, by showing the density in a spectrum of colors from blue (low) to red (high). On the left-side panel, users can select which dataset and query to view, and can fine-tune the parameters of the generated maps (e.g., opacity, number of bins, and heatmap intensity).

There is a collapsible panel (Figure 4) on the right consisting of five sub-panels, each of which presents a set of information corresponding to the current query that aids the analysis, including: **(1)** Query information, shows the parameters that are used for this query; **(2)** Random samples of tweets in the query (Figure 4a); **(3)** Temporal analysis of different scopes, including by year, by month, by week, by day, by weekday, and by hour-of-day (Figure 4b); **(4)** Word cloud of most popular words in among the tweets in the query (Figure 4d); and **(5)** Bar chart of frequency of each category (e.g., type of drug mentioned in each tweet) (Figure 4c). If the user is interested in more detailed analysis, by clicking the items in these panels, e.g., a keyword in (Figure 4d), and a category in (Figure 4c), a sub-query with updated parameters will be launched and new results will be displayed. The user can easily switch between queries to compare them.

4 CONCLUSION AND FUTURE WORKS

In this paper, we developed a community-focused drug abuse monitoring and supporting system, called DrugTracker, using social media and geospatial data in near real-time. Our DrugTracker system provides vital source of information when combating the drug abuse epidemiology, and proposed a function rich visualization system that can help local communities and organizations being informed about drug trends, locating drug abuse hot-spots, and reaching online users who may in need for help.

Several future works can be done on the proposed system. Here we just name a few: (1) Integrates more varieties of offline geospatial data that fits the needs of different aggregation levels; (2) Integrates more advanced privacy preserving methods that enables more detailed analysis with lower risk of unwanted leak of privacy; and (3) Further enrich the system with social connections (e.g., following, user mention) to enable the association of social connection information with geospatial data that aids the analysis.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from the National Science Foundation (NSF) grants CNS-1650587, CNS-1747798, CNS-1624503, and CNS-1850094.

REFERENCES

[1] Substance Abuse and Mental Health Services Administration. 2017. *Drug Abuse Warning Network.* https://www.samhsa.gov/data/data-we-collect/dawn-drug-abuse-warning-network
[2] Justin C Bosley, Nina W Zhao, Shawndra Hill, Frances S Shofer, David A Asch, Lance B Becker, and Raina M Merchant. 2013. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. *Resuscitation* 84, 2 (2013), 206–212.
[3] Preciosa M Coloma, Benedikt Becker, Miriam CJM Sturkenboom, Erik M van Mulligen, and Jan A Kors. 2015. Evaluating social media networks in medicines safety surveillance: two case studies. *Drug safety* 38, 10 (2015), 921–930.
[4] Tao Ding, Warren K Bickel, and Shimei Pan. 2017. Social media-based substance use prediction. *arXiv preprint arXiv:1705.05633* (2017).
[5] Yujie Fan, Yiming Zhang, Yanfang Ye, Wanhong Zheng, et al. 2017. Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies. In *CIKM'17*. 1259–1267.
[6] The FDA. 2019. *MedWatch: The FDA Safety Information and Adverse Event Reporting Program.* https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program
[7] Han Hu, NhatHai Phan, James Geller, Stephen Iezzi, Huy Vo, Dejing Dou, and Soon Chun. 2019. An Ensemble Deep Learning Model for Drug Abuse Detection in Sparse Twitter-Sphere. In *MEDINFO'19*.
[8] Han Hu, NhatHai Phan, James Geller, Huy Vo, Bhole Manasi, Xueqi Huang, Sophie Di Lorio, Thang Dinh, and Soon Ae Chun. 2018. Deep Self-Taught Learning for Detecting Drug Abuse Risk Behavior in Tweets. In *CSoNet'18*. 330–342.
[9] Hsien-Wen Meng, Suraj Kath, Dapeng Li, and Quynh C. Nguyen. 2017. National substance use patterns on Twitter. *PLOS ONE* 12, 11 (11 2017), 1–15.
[10] Megan A Moreno, Dimitri A Christakis, Katie G Egan, Libby N Brockman, and Tara Becker. 2012. Associations between displayed alcohol references on Facebook and problem drinking among college students. *Archives of pediatrics & adolescent medicine* 166, 2 (2012), 157–163.
[11] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15, 8 (2013), e174.
[12] American Association of Poison Control. 2017. *National Poison Data System (NPDS).* https://aapcc.org/data-system
[13] Laura Orsolini, Duccio Papanti, John Corkery, and Fabrizio Schifano. 2017. An insight into the deep web; why it matters for addiction psychiatry? *Human Psychopharmacology: Clinical and Experimental* 32, 3 (2017), e2573.
[14] NhatHai Phan, Soon Ae Chun, Manasi Bhole, and James Geller. 2017. Enabling real-time drug abuse detection in tweets. In *ICDE'17*. 1510–1514.
[15] Kevin R. Scott, Lewis Nelson, Zachary Meisel, and Jeanmarie Perrone. 2015. Opportunities for exploring and reducing prescription drug abuse through social media. *Journal of addictive diseases* 34, 2-3 (2015), 178–184.
[16] Samantha Schmidt. 2018. *'It is taking people out': More than 70 people overdose on K2 in a single day in New Haven.* https://www.washingtonpost.com/news/morning-mix/wp/2018/08/16/it-is-taking-people-out-more-than-70-people-overdose-on-k2-in-a-single-day-in-new-haven/?noredirect=on&utm_term=.e81dd0a336d8
[17] National Institute of Drug Abuse. 2019. *Opioid Overdose Crisis.* https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis
[18] Douglas C. Throckmorton, Scott Gottlieb, and Janet Woodcock. 2018. The FDA and the Next Wave of Drug Abuse - Proactive Pharmacovigilance. *New England Journal of Medicine* 379, 3 (2018), 205–207.
[19] Twitter. 2019. *Twitter API Documentation.* https://developer.twitter.com/en/docs
[20] Ye Zhao, Andrew Curtis, Xinyue Ye, Jing Yang, Chao Ma, Shamal AL-Dohuki, Farah Kamw, and Suphanut Jamonnak. 2018. *NeighborVis.* http://vis.cs.kent.edu/NeighborVis/index.html